APPLICATIONS OF DISCOURSE STRUCTURE FOR SPOKEN DIALOGUE SYSTEMS

by

Mihai Rotaru

B.S., Universitatea de Vest, Romania, 1997

M.S., Universitatea de Vest, Romania, 1999

Submitted to the Graduate Faculty of

the Department of Computer Science in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2008

UNIVERSITY OF PITTSBURGH

SCHOOL OF ARTS & SCIENCES

This dissertation was presented

by

Mihai Rotaru

It was defended on

September 11, 2008

and approved by

Dr. Diane J. Litman, Professor, Department of Computer Science

Dr. Rebecca Hwa, Assistant Professor, Department of Computer Science

Dr. Janyce Wiebe, Professor, Department of Computer Science

Dr. Carolyn P. Rosé, Assistant Professor, School of Computer Science, Carnegie Mellon University

Dissertation Advisor: Dr. Diane J. Litman, Professor, Department of Computer Science

APPLICATIONS OF DISCOURSE STRUCTURE
FOR SPOKEN DIALOGUE SYSTEMS

Mihai Rotaru, PhD

University of Pittsburgh, 2008

Language exhibits structure beyond the sentence level (e.g. the syntactic structure of a sentence). In particular, dialogues, either human-human or human-computer, have an inherent structure called the **discourse structure**. Models of discourse structure attempt to explain why a sequence of random utterances combines to form a dialogue or no dialogue at all. Due to the relatively simple structure of the dialogues that occur in the information-access domains of typical spoken dialogue systems (e.g. travel planning), discourse structure has often seen limited application in such systems.

In this research, we investigate the utility of discourse structure for spoken dialogue systems in more complex domains, e.g. tutoring. This work was driven by two intuitions.

First, we believed that the "position in the dialogue" is a critical information source for two tasks: performance analysis and characterization of dialogue phenomena. We define this concept using **transitions** in the discourse structure. For performance analysis, these transitions are used to create a number of novel factors which we show to be predictive of system performance. One of these factors informs a promising modification of our system which is implemented and compared with the original version of the system through a user study. Results show that the modification leads to objective improvements. For characterization of dialogue phenomena, we find statistical dependencies between discourse structure transitions and two dialogue phenomena which allow us to speculate where and why these dialogue phenomena occur and to better understand system behavior.

Second, we believed that users will benefit from direct access to discourse structure information. We enable this through a graphical representation of discourse structure called the **Navigation Map**. We demonstrate the subjective and objective utility of the Navigation Map through two user studies.

Overall, our work demonstrates that discourse structure is an important information source for designers of spoken dialogue systems.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

While I do hope that the rest of this document will be an enjoyable and scientifically sound read, there is one detail missing from the pages that follow. It is something that any finished work lacks at a quick look, be it a scientific paper or dissertation, a movie, a building or anything that just hasn't popped into existence in a quirk of matter and space. You might have guessed already what I am talking about: it is the very process that has produced this work. In my case, it is the transformational journey from a mind that was pretty good at taking things apart and cutting deep into their fabric to a mind that (hopefully) knows how to put these things together and formulate perspectives and visions. I will take advantage of the minimal guidelines for a preface and briefly talk about this process.

There are two main reasons why I wanted to share the existence of this process with the reader. First, there is my fascination with knowing how things are constructed. Second, from my discussions with other fellow Ph.D. students, I know that this journey has been challenging for many. And I am not ashamed to say that this includes me. There are times when one feels lost or helpless. No wonder we all read Ph.D. Comics (www.phdcomics.com).

I will start by telling you a little bit about my past. When it comes to computers and computer science, things have always fitted me like a glove. At 12 years old I interacted for the first time with a computer and something clicked. I was so fascinated by it that I managed to learn Basic by myself from a book given by my math teacher, Ecaterina Pupaza. Driven by this propensity and with the help of my computer science teacher, Stefan Bojinca, high-school was a time of great reward from learning new programming languages and algorithms, and competing in many informatics contents. During my undergraduate studies, I learned even more about computer science and in particular, I was fascinated by the Artificial Intelligence classes I took with prof. Viorel Negru. I also got a good taste from professional software development from the 3 years I worked for a small startup I co-founded with Dan Bohus.

So everything was going nice and smoothly for me and then came the Ph.D. I mean, yes, I took with great pleasure many graduate level course but something else was not working as smoothly as I was accustomed. I was asked to do more than just understanding of a concept or a theory, writing some piece of code or analyzing some data. I was baffled! I could read a scientific paper and pick up on all potential issues but still that wasn't enough. I was supposed to "connect" papers. But how do you do that? I tried

diving even deeper into papers and got lost in more and more details. I read guides written by other people but I did not know how to put them in practice. On top of all this confusion there was also frustration. Frustration with the results I was getting from my research. Come on! I thought that if you have a good idea, then the results should follow quickly! How can I believe in my own research if it is not working?

There was pain! And as many other fellow students, procrastination was a sweet anesthetic. A reaction so common, that some very smart people in the past added to the Ph.D. program two requirements: the comprehensive examination and the Ph.D. proposal. Yes, I wanted to give up my Ph.D. several times before each of them.

Turns out I was lacking two critical skills: vision and perseverance. I will not attempt to define any of them, but particularly informative and inspirational for this task is the transcript of the "You and Your Research" talk by Richard Hamming. Unfortunately, I can not provide you with a recipe or a training manual for those skills. For me they developed slowly in time through work, struggle and enthusiasm. I believe it helps if you learn these skills at a younger age as many educational systems do these days. In the end, it will be your own transformational journey and you have to be ready to take it.

Was this a solitary journey? No! I owe the most to two people: my advisor, Diane Litman, and my wife, Diana Rotaru. I have learned so much from both and this journey wouldn't have been possible without you. I really appreciate your patience as in some cases I proved to be a stubborn pupil. As a funny side note, I can tell you that in about one thousand e-mails I send to my advisor during my studies, I made only once the freudian typo as Diane humorously put it.

There were many other companions along this journey who have kindly offered their wisdom, experience, support, empathy, entertainment, food, drinks and friendship. Here is an incomplete list: Alex Gruenstein, Amruta Purandare, Antoine Raux, Arthur Ward, Beatriz Maeireizo Tokeshi, Behrang Mohit, Dan Bohus, David Schlangen, Emil Talpes, Erin Gemmill, Greg Nicholas, Gunes Erkan, Hua Ai, Joel Tetrault, Kate Forbes-Riley, Ludmila Khvan, Michal Valko, Richard Pelikan, Satanjeev Banerjee, Scott Siliman, Shime Pan, Tatiana Ilina, Tomas Singliar, Verena Rieser and Zuzana Jurigova. I would also like to thank my committee which has helped me steer this work into the document you are about to read. My family also gets a big hug!

This dissertation is my first major attempt to vision and perseverance. I hope I will have the reader's consent after the reading. None of the struggles behind it transpires from within these pages, but you can be sure they were, and if you are going through similar things, don't give up, you will get there! And you will look back and you will be very pleased.

One question summarizes everything: was all these worth it? Definitively YES!

1 INTRODUCTION

Verbal communication is one of the most important and oldest forms of human communication. It is one of the main skills we learn as we grow up and it is central to our everyday life whether we are ordering food or engaging in a heated discussion about the last night's "Scrubs" episode. Unlike other skills (e.g. being able to use a computer), it is expected of most individuals in a society to have a minimal mastery of verbal communication. Thus, it is only natural for people to want to interact with non-human entities via speech. Wouldn't it be easier to say to the bedside lamp to turn off after we have just cozily tucked ourselves into bed or to ask our computer to count the number of correct student turns instead of writing a script in a programming language?

Spoken Dialogue Systems (**SDS**) is the field of Computer Science that studies computer systems that interact with users via speech. Advances in key technologies behind SDS (automated speech recognition, natural language understanding, dialogue management, language generation and synthesis) have allowed researchers to build systems in a variety of domains. Information access domains have especially received a lot of attention due to the relatively simple structure of the dialogues in these domains: e.g. air travel planning (Rudnicky et al., 1999), weather information (Zue et al., 2000), bus schedule information (Raux et al., 2005), train schedule information (Swerts et al., 2000), PowerPoint presentation command & control (Paek and Horvitz, 2000). Recently, a number of research groups have turned their attention to investigating SDS in more complex domains: e.g. tutoring & collaborative learning (Graesser et al., 2001; Kumar et al., 2007; Litman and Silliman, 2004; Pon-Barry et al., 2006), procedure assistants (Rayner et al., 2005), medication assistance (Allen et al., 2006), planning assistants (Allen et al., 2001), etc. The increased robustness of spoken dialogue systems has led to many commercial applications. There are already companies specializing in developing speech solutions for other businesses (e.g. Nuance, TellMe, SpeechCycle, etc.). An increasing number of businesses are automating tasks that were previously performed by human operators in call center applications (e.g. checking credit card balances, making travel reservations, checking baggage status, etc.).

One of the main research problems in the SDS field is how to design a robust, efficient and usable SDS. To address this problem, a number of research questions have been pursued. Of interest to this work are the following:

1. What are the factors that influence the success of a SDS? What is the best strategy to use to account for these factors?
2. What dialogue phenomena should we take into consideration when designing a system? What characterizes those phenomena and how to detect and handle them?
3. What information to communicate to the user? What is the best channel/modality to use for this information?

There has been a considerable amount of work in the past that addresses these questions. Those of interest to us will be briefly described here. *Performance analysis* is one approach to answering the first question group. This SDS task tries to find factors that are associated with system performance by modeling performance in terms of a set of interaction parameters. It was pioneered by the PARADISE framework developed by (Walker et al., 1997). Knowledge of these factors can then be used to focus the design effort on specific aspects. The second question group is motivated by the observation that certain dialogue phenomena and the way they are handled by a SDS have a big impact on the robustness and usability of that SDS (e.g. speech recognition problems). One approach towards detecting and handling dialogue phenomena is to first try and *characterize the phenomena*: why do these phenomena occur and where. The third group of questions is related to the capabilities and limitations of how people process audio and visual information. We discuss in greater details these approaches and other related work in Sections 4.6, 5.6 and 6.5.

However, a particular characteristic of dialogues has received little attention when addressing these questions: *the fact that every dialogue has an underlying structure*. The formal name for this property is the **discourse structure**. This property explains how a well formed dialogue, or more generally a discourse, differs from a random set of sentences/utterances. For example, a dialogue between two persons tends to follow a certain structure: there might be greetings in the beginning, and then various topics are discussed. For each topic the dialogue might go into specific subtopics. Finally, the dialogue might conclude with a farewell. We will give a more concrete example of discourse structure shortly.

In this work, we investigate applications of discourse structure for SDS: we explore whether the discourse structure information has any value for answers to the three questions described above. More specifically, we look at two types of applications: on the system side and on the user side. On the system side, we investigate the utility of discourse structure for *performance analysis* (question 1 above) and for *characterizing dialogue phenomena* (question 2 above). On the user side, we investigate the utility of discourse structure for *graphical output* of a SDS (question 3 above).

The catalyst for our work was the observation that dialogues managed by SDS in complex domains have a *richer discourse structure*. This is primarily due to an increased task complexity for these domains. Indeed, in typical information access SDS, the task is relatively simple: get the information from the user and return the query results with minimal complexity added by confirmation dialogues. For example, a SDS in the air-travel domain (Walker et al., 2002) has to obtain from its users information like departure city, arrival city, date and time and based on these constraints to query the database and return the flights that satisfy the criteria. If multiple flight legs are required the process is repeated.

In contrast, for SDS in complex domains the situation is different. Take for example tutoring. A tutoring SDS has to discuss concepts, laws and relationships and to engage in complex subdialogues to correct user misconceptions. We illustrate this complexity with an example from the ITSPOKE speech-based tutoring system (Litman and Silliman, 2004), the testbed of this work. ITSPOKE is a speech-enabled version of the WHY-Atlas (VanLehn et al., 2007; VanLehn et al., 2002) text-based conceptual physics tutoring dialogue system (more details on the system in Section 2.1). Figure 1 shows one of the ITSPOKE tutoring plans for a physics problem (see upper part of the figure). Several tutoring topics are covered to address the problem. The system will analyze two time frames with the student: before the keys are released and after the keys are released. While the discussion for the "before release" topic is short, the discussion for the "after release" topic is quite lengthy. The system needs to discuss for both the man and the keys the forces acting on them, the net force and then, based on that information, infer the acceleration's direction and value. Next, the relationship acceleration-velocity is being discussed to compare the man's velocity and the key's velocity which in turn will be used for comparing the displacements and draw conclusions. The complexity of this tutoring plan will result in dialogues with an average of 21 system-user exchanges (i.e. the system asks questions and users answer). In contrast, for many information access SDS the average number of system-user exchanges typically ranges between 7 and 14 (e.g. (Litman and Pan, 2002; Raux et al., 2006)). Please note that this tutoring plan does not show the subdialogues initiated by the system for incorrect user answers.

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.*

Tutoring plan:

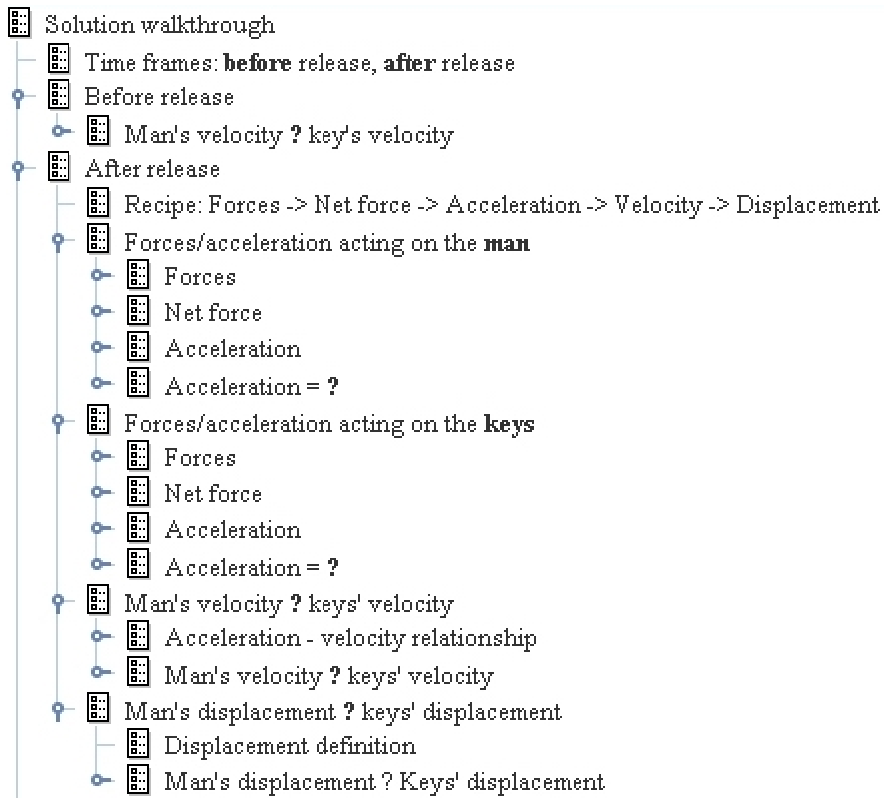


Figure 1. A sample ITSPOKE tutoring plan

(problem discussed: upper part, tutoring plan: lower part)

It is easy to see how this complex tutoring plan will translate into a dialogue with a rich discourse structure when we apply the Grosz & Sidner theory of discourse (Grosz and Sidner, 1986). According to this theory, each discourse (monologue or dialogue) 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 in discourse segments each with an associated discourse segment purpose/intention. For task-oriented dialogues, the underlying task structure acts as the skeleton for the discourse structure. Going back to the example in Figure 1, each line (tutoring topic) in the tutoring plan is equivalent to a discourse segment intention. As a result, a dialogue that follows this tutoring plan will have a discourse segment hierarchy similar to the hierarchical organization of the tutoring topics. This hierarchy is relatively complex reaching three levels of nesting. Moreover, some discourse segments are made of a substantial number of sub-segments (e.g. the “After release” segment is made of 5 sub-segments; the “Forces/acceleration acting on the man/keys” segment is made of 4 sub-segments). Please note, that we have not even included here the subdialogues initiated by the system for incorrect answers. These subdialogues will further increase the level of nesting and the complexity of the overall dialogue.

1.1 PROBLEM STATEMENT

We investigate the utility of discourse structure for spoken dialogue systems (SDS) in complex domains. To validate this utility, we pursue two types of applications: on the system side and on the user side. This classification is based on the entity directly exposed to the discourse structure information: the system and/or the system designer for applications on the system side or the user for applications on the user side.

On the system side, we investigate if the discourse structure information is useful for two important SDS tasks: *performance analysis* (Section 4) and *characterization of discourse phenomena* (Section 5). For performance analysis, we investigate if knowledge of discourse structure leads to discovery of factors that have a direct impact on performance. For characterization of dialogue phenomena, we investigate if discourse structure can be used to formulate meaningful hypotheses about where and why discourse phenomena occur.

On the user side, we investigate if the discourse structure information is useful for the graphical output of a SDS (Section 6). More specifically, we study whether users benefit from having direct graphical access to the discourse structure while interacting with a SDS.

We use tutoring as the complex domain and the ITSPOKE spoken dialogue system as the testbed for our experiments.

1.2 GENERAL APPROACH & RESULTS

The applications of discourse structure on the system side are driven by our intuition that the “position in the dialogue” is a critical information source for the two applications. To define the concept of “position in the dialogue”, we exploit the discourse structure information by looking at transitions in the discourse structure hierarchy (Section 3.3).

For performance analysis (Section 4), we use discourse structure transitions in combination with other dialogue phenomena to derive a number of factors (i.e. interaction parameters). We show how these factors are linked to performance in an empirical study that looks at the correlation between them and system performance (learning). One of these factors informs a promising modification of our system. We implement this modification and we compare it with the original version of the system by running a user study. The differences between the two versions are studied on a number of objective and subjective metrics. Our results indicate that factors which use discourse structure are correlated to system performance. Some of these factors have intuitive explanations and inform promising modifications of our system. Analysis of the data from the user study shows that the implemented modification leads to objective improvements for our system (e.g. performance improvements for certain users but not at the population level and fewer system turns).

For characterization of dialogue phenomena (Section 5), we look at statistical dependencies between discourse structure transitions and dialogue phenomena. We mine dependencies in a corpus of dialogues. Analyses of the dependencies tell us where dialogue phenomena occur more/less than expected and allow us to formulate hypotheses behind this increase/decrease. We use this approach for two dialogue phenomena: user affect and speech recognition problems. Our results show dependencies between discourse structure transitions and both phenomena. Several transitions in the discourse structure are associated with an increase or decrease of uncertainty and have specific interaction patterns with speech recognition problems. These dependencies offer insights into system behavior and provide additional motivations for our graphical representation of discourse structure.

Our user-side application of discourse structure (Section 6) is motivated by our intuition that users benefit from direct access to discourse structure information. We enable this direct access through a graphical representation of discourse structure called the Navigation Map. We investigate the value of the Navigation Map from user’s perspective (subjective utility) and in terms of performance (objective utility). We run a user study for each investigation. When exposing users to both a Navigation Map version of our system and a baseline version with no graphical support during the conversation, we find that users prefer the Navigation Map version and rate it better on a number of dimensions (e.g. ability to identify and follow tutoring plan, integrate instruction, concentration, etc.). When compared to a baseline that provides graphical support during conversation in the form of a dialogue transcript, we find that the Navigation Map version leads to improvements on a number of objective metrics (e.g. increased system performance for certain users but decreased performance for others and shorter dialogues).

A summary of our analyses, results and other work items is available below in Table 1. For each item we show the section where more details are available, a brief description, the statistical confidence in the outcome and whether it is a positive or negative outcome.

Table 1. Summary of results

Result	**Confidence**	**+/-**
(4) APPLICATIONS FOR PERFORMANCE ANALYSIS		
(4.4) Predictiveness		
(4.4.1) Several discourse-structure based parameters are predictive of system performance (learning)		
• transition-correctness bigrams	**significant**	+
• transition-certainty bigrams	**significant**	+
• transition-transition bigrams	**significant**	+
(4.4.2) PopUp–Incorrect bigram predictiveness generalize to other corpora	**trend/significant**	+
(4.5) Informativeness		
(4.5.2) Implemented a new PopUp–Incorrect strategy		+
(4.5.3) Run a between-subjects user study		+
(4.5.4) Comparison with control condition (`PI` vs `R`)		
• (4.5.4.2) System performance: very small improvement (0.07 NLG effect size)	non-significant	-
• (4.5.4.3) Aptitude-treatment interaction: pretest split (PRE Split)		
• PRE Split x Condition: no interaction with learning	non-significant	
• + 0.15 NLG effect size for low pretesters	non-significant	
• + 0.00 NLG effect size for high pretesters	non-significant	
• (4.5.4.4) PopUp–Incorrect predictiveness: significantly correlated with system performance for `R` users, but not for `PI` users	**significant**	+
• (4.5.4.4) Splitting based on the number of PopUp–Incorrect events (PI Split)		
• PI Split x Condition: trend interaction with learning	*trend*	
RELNLG subset (2.3.3.2)		
• PI Split x Condition: significant interaction with learning	**significant**	
• + 0.96 NLG effect for high PopUp–Incorrect users	**significant**	+
• - 0.50 NLG effect for low PopUp–Incorrect users	non-significant	+/-
• `PI` users: low and high PopUp–Incorrect users have similar learning		+
• (4.5.4.5) Subjective metrics: no differences	non-significant	+
• (4.5.4.6) Dialogue time: very small reduction in dialogue time	non-significant	+
• (4.5.4.7) Number of system turns: reduction in number of system turns	*trend*	+
(5) APPLICATIONS FOR CHARACTERIZATION OF DIALOGUE PHENOMENA		
(5.4) Characterization of user affect (uncertainty)		
Statistical dependencies between discourse structure transitions and uncertainty	**significant**	+

Statistical dependencies between discourse structure transitions and uncertainty even after we discount for user correctness		
• Dependencies exist for correct user turns	**significant**	**+**
• Dependencies exist for incorrect user turns	**significant**	**+**
(5.5) Characterization of Speech Recognition problems		
Statistical dependencies between discourse structure transitions and ASR misrecognitions	**significant**	**+**
Statistical dependencies between discourse structure transitions and Rejections	**significant**	**+**
Statistical dependencies between discourse structure transitions and Semantic Misrecognitions	*trend*	**+**
(6) APPLICATIONS FOR GUI: THE NAVIGATION MAP		
(3.4) Implementation of the Navigation Map for ITSPOKE		**+**
(6.3) Subjective utility		
(6.3.1) Run a within-subjects user study		**+**
(6.3.2) Comparison with a control condition with no graphical support (NM vs noNM)		
(6.3.2.1) Presence of the NM has a positive effect on how users rate several questions in the following categories:		
• Overall impression (easy to learn and concentrate, expectation, reuse)	**trend/significant**	**+**
• During the conversation (identify and follow the tutoring plan, integrate the instruction)	**trend/significant**	**+**
• After the conversation (understand tutor's main point)	*trend*	**+**
(6.4) Objective utility		
(6.4.1) Run a between-subjects user study		**+**
(6.4.2) Comparison with a control condition with graphical support (*NM* vs. *R*)		
• (6.4.2.1) System performance: very small decrease (- 0.06 NLG effect size)	non-significant	**-**
• (6.4.2.2) Aptitude-treatment interaction: pretest split (PRE Split)		
• PI Split x Condition: interaction with learning	*trend*	
• + 0.46 NLG effect size for high pretesters	non-significant	**+**
• - 0.66 NLG effect size for low pretesters	non-significant	**-**
RELNLG subset (2.3.3.2)		
• PI Split x Condition: interaction with learning	**significant**	
• + 0.85 NLG effect size for high pretesters	**significant**	**+**
• - 0.66 NLG effect size for low pretesters	non-significant	**-**
• (6.4.2.5) Dialogue time: reduction in dialogue time	**significant**	**+**
• (6.4.2.6) Number of system turns: reduction in the number of system turns	non-significant	**+**
• Subjective metrics:		
• (6.4.2.3) Positive effect on average rating in the Conversation category (Conv)	*trend*	**+**
• (6.4.2.4) Correlation between average user rating on every category (Overall, Conv, Essay and ALL) and learning for *NM* users but not for *R* users	**significant**	**+/-**

1.3 CONTRIBUTIONS OF THIS WORK

Due to the relatively simple nature of the underlying domains in previous SDS, discourse structure has seen limited applications in these systems. The main contribution of this work is to *establish discourse structure as a useful information source for SDS in complex domains*. In particular, we show that the discourse structure information is valuable for two empirical approaches to dialogue design: performance analysis and characterization of dialogue phenomena. In addition, we show that discourse structure information is also valuable for users of a SDS through a graphical representation, which we call the Navigation Map.

Our work also contributes in other aspects. Through our discourse structure transitions (3.3), we offer a way of defining the notion of “position in the dialogue”, a definition which is domain-independent and automatically computable. Defining this concept was crucial for our performance analysis and characterization of dialogue phenomena investigations. Through our transition–phenomena bigrams (4.3), we provide a way to look at dialogue phenomena in their dialogue context. Our results show that looking at dialogue phenomena (correctness and user affect) in their dialogue context offers more insights about system performance compared to looking at dialogue phenomena out of context (4.4). Through our transition–transition bigrams (4.3), we offer a way of quantifying the structure of a dialogue. Our results (4.4) show that the structure of the dialogue, as measured by parameters derived from these bigrams, can discriminate between “good” and “bad” dialogues with our system (where good/bad is defined in terms of learning). Statistical dependencies between discourse structure transitions and dialogue phenomena suggest that discourse structure transition as an informative feature for predictive models of dialogue phenomena (5.4 and 5.5). These dependencies also show how discourse structure transitions can be used to better understand the behavior of a system. It is interesting to note that the two applications make different conceptual use of discourse structure transitions. For performance analysis, we use transitions in collaboration with dialogue phenomena to model system performance. For characterization of dialogue phenomena, we use transitions to model dialogue phenomena by looking at statistical dependencies between them.

For performance analysis, we provide one of the few examples of a complete application of this approach to empirical dialogue design. By examining existing dialogues, we discover *new* factors related to performance of a SDS. These factors relate to issues in the design of the SDS that were not anticipated beforehand (e.g. the PopUp–Incorrect behavior - 4.5.1). We do not stop at identifying these issues, but we also show that by addressing them we obtain performance improvements (at least for certain users – 4.5).

We also show that the information presented on the graphical output of a SDS (if available) can have a considerable impact on system performance. We show that by replacing the transcript of the dialogue with a graphical representation of discourse structure we obtain a number of subjective and objective improvements. It is interesting to note that this improvement comes without making any changes to actual spoken dialogue between the system and users. We provide one of the first graphical representations of discourse structure: the Navigation Map (3.4). This representation can be easily extended to change the way users interact with the SDS. For example, the Navigation Map can be easily extended with a new functionality that will allow users to control the dialogue flow even in a system directed conversation (e.g. skipping topics).

This work contributes to the Computational Linguistics field through our novel representation and application of discourse structure. Since our investigation is performed on a SDS in the tutoring domain, our findings also open new avenues for research with the potential for advancing the state-of-the-art in speech-based intelligent tutoring systems.

1.4 A GUIDE FOR READERS

The rest of this document is organized as follows.

In **Section 2 – Preliminaries**, we present the ITSPOKE system, the testbed of this research. We describe several ITSPOKE experiments, the collected corpora, a number of metrics and annotations on these corpora, and the statistical tools that will be used in our analyses.

In **Section 3– Discourse Structure**, we introduce the concept of discourse structure in more detail and identify promising elements for our purposes. We introduce the concept of discourse structure transition which will be used heavily in Sections 4 and 5. We present an automatic annotation of the discourse structure hierarchy and discourse structure transitions. We also describe our manual annotation of discourse structure and our automatic graphical representation of discourse structure (the Navigation Map).

In **Section 4 – Applications for Performance Analysis**, we demonstrate the utility of discourse structure for performance analysis in SDS, our first system-side application. We describe how discourse structure transitions can be used to derive a number of factors which are later shown to be correlated with system performance through an empirical study. One of these factors informs a promising modification of our system which is implemented and compared with the original version of the system through a user study. Results show that the modification leads to objective improvements.

In **Section 5 – Applications for Characterization of Dialogue Phenomena**, we demonstrate the utility of discourse structure for the characterization of dialogue phenomena, our second system-side application. We look at statistical dependencies between discourse structure transitions and two dialogue phenomena: user affect and speech recognition problems. Results show that several transitions are associated with an increase or decrease of uncertainty (our choice of user affect) and have specific interaction patterns with speech recognition problems.

In **Section 6 – Applications for GUI: The Navigation Map**, we motivate the intuition behind the Navigation Map and demonstrate its utility. We validate the subjective and objective utility of the Navigation through two users studies.

We conclude in Section 7 with a brief summary of our main findings and contributions and discuss a number of interesting directions for the future.

Readers interested in specific applications of discourse structure are recommended to first familiarize themselves with our system (2.1) and with the concept of discourse structure (3.1). Next, readers are recommended to read about discourse structure transitions (3.3) if interested in system side applications and then move on to the application of interest (Sections 4 or 5). Readers interested in user-side applications can read about the Navigation Map (3.4) and then skip to Section 6. Readers unfamiliar with correlations and ANOVA/ANCOVA tests are recommended to review Section 2.4 before reading the result sections.

2 PRELIMINARIES

2.1 THE ITSPOKE SYSTEM

ITSPOKE (**I**ntelligent **T**utoring **Spoke**n Dialogue System) (Litman and Silliman, 2004) is a speech-enabled version of the text-based Why2-Atlas conceptual physics tutoring system (VanLehn et al., 2007; VanLehn et al., 2002). Students discuss with ITSPOKE a set of five qualitative physics problems (Appendix A.2). Dialogues between students and ITSPOKE are mediated through a web interface supplemented with a headphone-microphone unit. An example screenshot of the ITSPOKE graphical interface is shown in Figure 2. For each problem, the interaction has the following format. First, ITSPOKE asks the student to read the problem text (top-right box in Figure 2) and to type an essay answering the problem (bottom right box in Figure 2). Next, based on the analysis of the student essay, ITSPOKE engages the student in spoken dialogue (using head-mounted microphone input and speech output) to correct misconceptions and elicit more complete explanations. A transcript of the dialogue is available in the dialogue history box (left box in Figure 2). At the end of the dialogue, the student is asked to revise the essay and, based on the analysis of the essay revision, the system decides whether to do another round of tutoring/essay revision or to move on to the next problem.

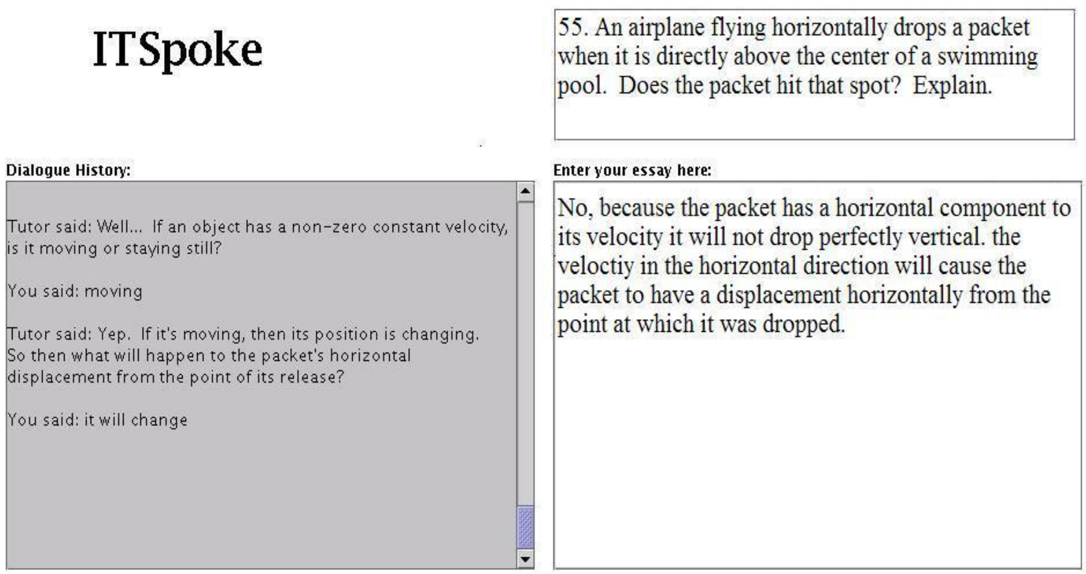


Figure 2. ITSPOKE interface during a dialogue with a student

The Why2-Atlas back-end is responsible for the essay analysis, for selecting the appropriate instruction and for interpreting the student answers (Jordan et al., 2001; Rosé et al., 2003). ITSPOKE handles speech input and speech output and manages some speech related problems (i.e. timeouts and rejections). During the dialogue, student speech is digitized from microphone input and sent to the Sphinx II recognizer (Huang et al., 1993). The default acoustic model included in the Sphinx II distribution is used; the stochastic language models were trained with student utterances from evaluations of Why2-Atlas and from previous pilot studies of ITSPOKE. Sphinx II's best "transcription" (recognition output) is then sent to the Why2-Atlas back-end for syntactic, semantic and dialogue analysis. Finally, the text response produced by Why2-Atlas is sent to the Cepstral[1] text-to-speech system and played to the student over headphones or speakers. A version of ITSPOKE that uses human prerecorded prompts is also available (Forbes-Riley et al., 2006).

A complete dialogue between ITSPOKE and a user is available in Appendix A.4.4.

[1] www.cepstral.com

2.2 USER EXPERIMENTS & CORPORA

The ITSPOKE system has been used in several user experiments. In this work we will use data from 4 studies: **F03** (2.2.1), **S05** (2.2.2), **NMPrelim** (2.2.3) and **Main** (2.2.4). Out of them, only the NMPrelim and Main experiments were designed and performed by the author[2]. We first describe the steps and procedures that are common to all experiments. Then we discuss additional details for each experiment and describe the collected corpus. Details about experiment design choices are available in other sections which will be referenced in the text. Corpora statistics are available in Table 2.

The experiment procedure for these studies contains at least the following steps in order:

1. **Overview** – In this step, users were explained the procedure used in this experiment and a quick overview of each step was given.
2. **Reading** – In this step, users were asked to read a short material with background information about the tutored domain. The information was presented on the screen in form of a webpage. The material is available in Appendix A.1.
3. **Pretest** – In this step, users took a test designed to measure their initial physics knowledge. In all experiments except S05 (2.2.2), the pretest contained the same 26 multiple choice questions and 4 essay questions (see Appendix A.3). The test was administered through a web interface: users were presented with one problem per page and had to click a button to move to the next question. Users could also go back and correct previous questions.
4. **Instruction** – In this step, users worked through a set of 5 problems with ITSPOKE (except for NMPrelim – see Section 2.2.3). The same problems in the same order were discussed in all experiments. In some experiments (NMPrelim and Main), the essay analysis component was disabled and only a single "walkthrough" instruction per problem was covered with users (marked in the last column of Table 2)
5. **Posttest** – In this step, users took another test isomorphic to the pretest designed to measure their physics knowledge post-instruction (see Appendix A.3). The posttest and the pretest had the same number of questions and were administered through the same web interface. User learning can be measure through various ways of looking at the difference between the pretest and the posttest scores (see Sections 2.3.1.1 and 2.3.1.2).

---

[2] Whenever we attribute something to the author, note that other people may have contributed, primarily the thesis advisor, Diane J. Litman. For a detailed list of acknowledgements, please see the Preface.

In all experiments, the majority of users were undergraduate students at University of Pittsburgh, recruited from on-campus advertisements. They were required to be native speakers of American English, to have not taken college physics, to have not taken part in other physics user studies, and were paid for their involvement.

Table 2. Corpora statistics

Experiment	Condition	# of users	# of problems	User turns	System turns	Walkthrough only
F03	F03	20	5	2334	2964	N
S05	S05SYN	33	5	4636	5762	N
	S05PR	31	5	4428	5483	N
NMPrelim	F	13	2	461	530	Y
	S	15	2	495	566	Y
Main	*R*	25	5	1946	2252	Y
	PI	27	5	1963	2302	Y
	NM	27	5	1995	2324	Y

2.2.1 F03 Experiment

The **F03** experiment took place in Fall 2003 and was part of a larger study that compared speech-based versus typed-based tutoring (Litman et al., 2004). The experiment followed the standard procedure. The resulting corpus contains 20 users (3 male and 17 female).

Besides the annotations common to all corpora, the F03 corpus was also annotated for human correctness (2.3.2.1) and certainty (2.3.2.2).

2.2.2 S05 Experiment

The S05 experiment took place in Spring 2005 and was designed to compare synthesized versus prerecorded speech output (Forbes-Riley et al., 2006). The experiment followed the standard procedure plus an extra survey step at the end (for more details about this survey see (Forbes-Riley et al., 2006)). There were 2 conditions: **S05SYN** (synthesized-speech) and **S05PR** (prerecorded-speech). Both conditions used the same version of ITSPOKE (slightly improved over the one used in F03) but used a different speech output component. Users were randomly assigned to conditions.

There were 33 users in the S05SYN corpus (15 male and 18 female) and 31 users in S05PR corpus (13 male and 18 female).

2.2.3 NMPrelim Experiment

The NMPrelim experiment took place in October-November 2006 and was designed to test the subjective utility of the Navigation Map (**NM**), a graphical representation of the dialogue structure (described in Section 3.4). The experiment had 2 conditions: **F** and **S**. More details about the design of this study are available in Section 6.3.1.

For both conditions, we used the prerecorded version of ITSPOKE. In addition, the essay analysis component was disabled. As a result, for all users, the system went through the same "walkthrough" dialogue for each problem which assumed no information in the user essay. Note that the actual dialogue depends on the correctness of user answers. After the dialogue, users were asked to revise their essay and then the system moved on to the next problem (i.e. there is only one cycle of dialogue – essay rewriting).

The experiment procedure includes the basic ITSPOKE procedure plus several additional/modified steps (italicized in the list): 1) overview, 2) reading, 3) pretest, 4) *interaction with the system*, 5) posttest, 6) *NM survey*, and 7) *open question interview*. In the interaction step, users worked through 2 problems only and completed a system questionnaire after *each* problem (2.2.3.1 - note that this questionnaire and the NM survey are different). In the NM survey step, users completed a short survey about their NM usage (2.2.3.2). At the end of the experiment, there was an open question interview in which we asked users about their experience with the system, what they liked or disliked and some specific questions for about the NM.

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 were no significant differences between the two conditions in terms of pretest scores ($p<0.63$); in both conditions users learned (significant difference between pretest and posttest, $p<0.01$).

2.2.3.1 System Questionnaire

The system questionnaire was designed to measure the perceived (subjective) utility of the system for users. It was given through a web-based interface shown in Figure 3. It contains 16 questions that probed user's perception of the system on various dimensions. For each questions, a 5-point Likert scale was used (1 – Strongly Disagree, 2 – Disagree, 3 – Somewhat Agree, 4 – Agree, 5 – Strongly Agree). The questions are grouped in 3 categories: **Overall** (Q1-Q6), **Conv**ersation (Q7-Q13), **Post** (Q14-Q16). Questions in the Overall category 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 in terms of ease of use, enjoyment, reuse etc. Questions in the Conversation category probe user's perception during the dialogue with the tutor. These questions are related to dimensions we hypothesize the NM will influence (6.3): concentration, identification of the tutoring structure, integration of the information etc. Questions in the Post category look at the post-conversation perceptions. They probe the relevance of the conversation and its relationship with the essay update.

When comparing ratings, a higher rating is better for most questions with the exception of Q7 and Q11. These questions measure “negative” factors (high level of concentration and task disorientation) and for them a lower rating is better. They were introduced primarily as a deterrent for negligence in rating.

The questionnaire was administered after each problem in this study. The same questionnaire was also used in the Main study (2.2.4) however it was only given once at the end of the instruction.

	Overall:	**Strongly Agree**	**Agree**	**Somewhat agree**	**Disagree**	**Strongly Disagree**
1	The tutor increased my understanding of the subject	○	○	○	○	○
2	It was easy to learn from the tutor	○	○	○	○	○
3	The tutor helped me to concentrate	○	○	○	○	○
4	The tutor worked the way I expected it to	○	○	○	○	○
5	I enjoyed working with the tutor	○	○	○	○	○
6	Based on my experience using the tutor to learn physics, I would like to use such a tutor regularly	○	○	○	○	○
	During the conversation with the tutor:	**Strongly Agree**	**Agree**	**Somewhat agree**	**Disagree**	**Strongly Disagree**
7	... a high level of concentration is required to follow the tutor	○	○	○	○	○
8	... the tutor had a clear and structured agenda behind its explanations	○	○	○	○	○
9	... it was easy to figure out where the tutor's instruction was leading me	○	○	○	○	○
10	... when the tutor asked me a question I knew why it was asking me that question	○	○	○	○	○
11	... it was easy to loose track of where I was in the interaction with the tutor	○	○	○	○	○
12	... I knew whether my answer to the tutor's question was correct or incorrect	○	○	○	○	○
13	... whenever I answered incorrectly, it was easy to know the correct answer after the tutor corrected me	○	○	○	○	○
	At the end of the conversation with the tutor:	**Strongly Agree**	**Agree**	**Somewhat agree**	**Disagree**	**Strongly Disagree**
14	... it was easy to understand the tutor's main point	○	○	○	○	○
15	... I knew what was wrong or missing from my essay	○	○	○	○	○
16	... I knew how to modify my essay	○	○	○	○	○

Figure 3. System questionnaire web interface

2.2.3.2 NM Survey

While the system questionnaires (2.2.3.1) probed users’ perception of the system, in the second to last step in this experiment, users had to fill a NM survey which explicitly asked how the NM helped them, if at all. It was given through a web-based interface which is shown in part in Figure 4. It contains 9 questions. For 8 out the 9 questions, a 5-point Likert scale was used (1 – Strongly Disagree, 2 – Disagree, 3 – Somewhat Agree, 4 – Agree, 5 – Strongly Agree). The last question allowed users to specify other areas where the NM was useful for them. In addition, there were 3 open-ended questions where users can type an answer. The open-ended questions asked users what they liked the most and the least about the NM and what they would change in the NM.

	Overall	Strongly Agree	Agree	Somewhat agree	Disagree	Strongly Disagree
1	The presence of the Navigation Map helped me					
	The Navigation Map helped me...	**Strongly Agree**	**Agree**	**Somewhat agree**	**Disagree**	**Strongly Disagree**
2	... in following the conversation with the tutor					
3	... by summarizing the tutor's questions					
4	... by summarizing the correct answer to tutor's questions					
5	... while writing the essay					
6	... in understanding how to approach other problems					
7	... to learn easier					
8	... concentrate on the tutoring process					
9	... in other areas, please specify					

Figure 4. The NM survey

2.2.4 Main Experiment

The Main experiment took place between October 2007 and February 2008 and was designed to test the objective utility of two modifications of the original ITSPOKE system inspired by the discourse structure. The experiment had 3 conditions: ***R***, ***NM*** and ***PI***. *R* is the regular ITSPOKE system described in Section 2.1. *PI* is the regular ITSPOKE system with a modification of the PopUp–Incorrect strategy (described in Section 4.5.2). *NM* is the regular ITSPOKE system with the Navigation Map instead of the dialogue history (described in Section 3.4). *R* was the control condition in this experiment and *NM* and *PI* were the two experimental conditions. The experiment was designed for two pairwise comparisons: *NM* vs. *R* and *PI* vs. *R* (i.e. *R* was shared). More details about the design of this study are available in Section 4.5.3 and 6.3.1.

For all conditions, we used the prerecorded version of ITSPOKE. In addition, the essay analysis component was disabled. As a result, for all users, the system went through the same “walkthrough” dialogue for each problem which assumed no information in the user essay. Note that the actual dialogue depends on the correctness of user answers. After the dialogue, users were asked to revise their essay and then the system moved on to the next problem (i.e. there is only one cycle of dialogue – essay rewriting).

The experiment procedure includes the basic ITSPOKE procedure plus several extra steps (italicized in the list): 1) overview, 2) *memory test*, 3) reading, 4) pretest, 5) interaction with the system, 6) *system questionnaire*, 7) posttest, 8) *open question interview*. A detailed account of the discussion between the experimenter and users at each step is available in Appendix A.4.1. The memory test step was introduced to test an exploratory hypothesis and is described in Appendix A.4.3. In the sixth step, users rated the system using the system questionnaire from the NMPrelim experiment (see Section 2.2.3.1). Users in the *NM* condition had an additional question (qNM) discussed in Section 2.2.4.3. At the end of the experiment, there was an open question interview in which we asked users about their experience with the system, what they liked or disliked and some specific questions for each condition. More details are available in Appendix A.4.1.

User assignment to condition was done through a pseudo-random procedure. Details and validation of this procedure are available in Section 2.2.4.2.

We had 85 users in this experiment however only data from 79 users forms the corpus we will use in our analyses. Removal of the six users is discussed in Section 2.2.4.1. The resulting corpus is well balanced between conditions as Table 3 shows. There were about 50% more female users than male users in total, however the conditions have similar ratio of male/female users.

Table 3. Distribution of users per condition and gender (Main experiment)

	Condition			
Gender	*R*	*NM*	*PI*	Total
male	10	11	11	32
female	15	16	16	47
Total	25	27	27	79

2.2.4.1 Eliminated users

Out of the 85 users which took part in the Main experiment, we eliminated 6 users from our analyses for various reasons.

We eliminated one user in the *R* condition because the user skipped the fifth problem due to a system bug.

We eliminated 4 users (2 in *PI* and 2 in *NM*) because they exhibited negligence in the posttest step. Removal of problematic users is reported in other studies. For example, (Chi and VanLehn, 2008) remove a user "for deliberately wasting time". In our case, the criterion for elimination was users that have at least 2 posttest questions answered in 3 seconds or less[3]. The criterion was designed to capture a "click-through" behavior: users that click randomly an answer and move on to the next question. The four users eliminated were: two *NM* users (one with 16 such questions, the other with 7) and two *PI* users (one with 18 such questions, the other with 4). Note that our detection of "click-through behavior" is not foolproof: during the experiment, there were several days when due to network delays, it was impossible to advance to the next question in less than 4-10 seconds.

Since in our analyses we will compare the three conditions in terms of learning, it was very important to have reliable measures for the pretest and posttest metrics. We do not know the reasons behind the behavior of these four users. It is possible that it was induced by the interaction with the system. In these cases, eliminating the users will create a sample bias as we are eliminating users that do not like the system. However, there may be many other personal reasons behind this behavior (e.g. tiredness due to the long experiment session, being late for other appointments, disinterest, etc.). Since we do not have a clear indication of the actual reason, we elected to eliminate these users[4].

Finally, another *R* user was eliminated for being an outlier on several metrics (see outlier definition in Section 2.4.1). The logs for this user indicate a large number of speech recognition errors (timeouts, rejections and semantic errors, see Section 2.3.1.5). Recorded audio files indicate that this is due to user speaking very softly or keeping the microphone too far. This led to an inflated number of rejection tutor turns (e.g. "Could you please repeat that") or unwarranted remediation turns or dialogues (due to semantic errors). As a result this user is an outlier on several dimensions: dialogue time (TotalTime=4561 seconds, z-score=4.5, see Section 2.3.3.2) and number of system turns (TutTotal=140 turns, z-score=4.8, see Section 2.3.3.4). Figure 5 shows the distribution of the total dialogue time for the population in this experiment. The outlier can be easily seen on the right side of the graph. Although the user learned (pretest score is 9 out of 26, posttest score is 17 out of 26 - 2.3.1.1), his/her experience with the system is atypical of other users in this experiment. To reduce the impact it has on the *R* population statistics[5], we eliminated this user.

---

[3] For a reported average reading speed of 250 words per minute (Legge et al., 1985), reading the shortest posttest question and its available answers (47 words) would take at least 10 seconds. To account for very fast readers, we chose a threshold of 3 seconds.

[4] Since the posttest scores are unreliable for these 4 users regardless of the reasons behind their behavior, there was no need to rerun the learning-related analyses that will be presented in Sections 4.5 and 6.4 with these users included.

[5] Keeping the user increases the significance of the differences between *R* and the other two conditions in many of our analyses.

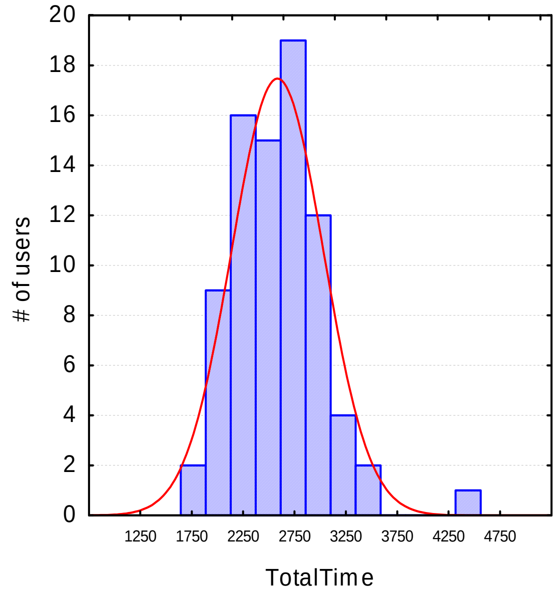


Figure 5. Distribution of TotalTime

2.2.4.2 Validation of the assignment procedure

User assignment to condition was done through a pseudo-random procedure. After the pretest step, the experimenter assigned the user to one of the conditions based on user's pretest score and gender. Using a histogram of the current pretest scores, the experimenter ensured a similar distribution of pretest scores in each condition for new users: the condition with the least number of users with a similar pretest score was selected. Ties were broken randomly in general but in certain cases the decision was influenced by two other factors: balance of gender and balance of number of users per condition.

We used this procedure because we noticed that in the S05 experiment there were some differences between the pretest distributions even though users were assigned randomly to conditions (e.g. there was a difference of 1.0 between the means - Table 5). Given the analyses we were planning to make, we wanted to have distributions as similar as possible.

To verify if the procedure produced similar subpopulations, we compare the `R`, `NM` and `PI` populations in terms of their pretest score (PRE score - Section 2.3.1.1). Table 4 shows the average and standard deviation for this metric in each condition. We find that the averages and standard deviations are very similar for the three conditions. Indeed, a one-way ANOVA with PRE score as dependent variable and condition as categorical factor (`R` vs. `NM` vs. `PI`) finds no significant differences between the three conditions ($F(2,76)=0.003$, $p<0.99$). The distribution of PRE scores for each condition is available in Appendix A.4.2 (Figure 34).

Table 4. Average and standard deviation for the PRE score

Condition	PRE
R	12.6 (4.2)
NM	12.6 (4.1)
PI	12.5 (4.4)

Based on these findings, we conclude that the pseudo-random procedure used to assign users to conditions was successful as it did not produce different populations in terms of pretest scores.

2.2.4.3 qNM system questionnaire question

Analyses done on the NMPrelim corpus (6.3.2.1) find that a small proportion of users disliked the NM. In the open question interview, these users indicated that they had trouble listening to the system and following the NM at the same time or that the NM was moving too fast. Since for these users, their ratings for the NM version of the system were lower, we wanted to identify them in the Main experiment.

Thus, for *NM* users, we added an extra question to the system questionnaire: **qNM** ("During tutoring it was hard for me to follow the information on the screen while listening to the tutor speech"). Unfortunately, the answers for qNM are not a reliable measure of user's having problems with the NM. During the open question interviews we discovered that the qNM text was ambiguous for users[6]. Several users which agreed or strongly agreed with this question indicated that they had no problems following the NM but that they wanted graphical representations of the problem or even animations. A better formulation of this question would have been "During tutoring it was hard for me to follow the Navigation Map while listening to the tutor speech".

2.3 ANNOTATIONS & METRICS

In this section we describe the annotations and the metrics used in this study. We begin with annotations and metrics that are common to all corpora and then we describe the ones specific to each corpus (or at least used only for that corpus).

---

[6] For the Main experiment we did not run a pilot user study to test our experiment design (the system modifications were minimal or tested in previous experiments; the system questionnaire without qNM was used in a previous study). It is likely that the pilot study would have discovered this design issue.

Figure 6 shows some of the annotations on a small ITSPOKE dialogue excerpt from the F03 corpus. For each user turn, we show the output of the automated speech recognition (ASR) component and the human transcript of what the student said (2.3.1.3). The annotations for each user turn are listed in parentheses and will be discussed below.

ITSPOKE: Good. Apply the same force to an object of mass M3, which is three times the mass of M1. What is the magnitude M3's acceleration in terms of A1 [acceleration of the object M1]?

STD$_1$: third of a one (***CERT=Uncertain, HSEM=C***) [*TSEM=correct*]
ASR$_1$: they're it but a one (***AsrMis, SemMis***) [*ASEM=incorrect*]

ITSPOKE: The acceleration of M3 [...] would be one-third the acceleration of M1. [...] In general for the same force what remains constant?

STD$_2$: weight (***CERT=Neutral, HSEM=I***) [*TSEM=incorrect*]
ASR$_2$: great (***AsrMis***) [*ASEM=incorrect*]

ITSPOKE: Well... The product of the mass of a body and its acceleration remains constant for a given force. What is the relationship between a force acting on the box, the mass of the box, and its acceleration?

STD$_3$: newton's second law (***CERT=Uncertain, HSEM=PC***)
ASR$_3$: the cancel the fire (***Rej***)

ITSPOKE: Could you please repeat that?

…

Figure 6. Annotated ITSPOKE dialogue excerpt

(human transcript, speech recognition problems, correctness, certainty)

2.3.1 Common annotations & metrics

2.3.1.1 Test scores

Both tests (pretest and posttest) contain 26 multiple-choice questions and 4 essay questions. We calculate for each test a score (**PRE** score for pretest, **POST** score for posttest) defined as the number of correctly[7] answered multiple-choice questions. The 4 essay questions are not used as manual annotation of the correctness of the answer is required (see (VanLehn et al., 2007) for alternatives to scoring essay questions).

Table 5 shows the average and the standard deviation of the PRE/POST scores for our corpora. It is interesting to note that most PRE averages are similar with the exception of S05PR which has a higher

[7] Note that since the test interface allowed users to go back and review their answers, the answer with the latest time stamp was used in calculating the PRE/POST scores.

average. Although POST scores can not be directly compared across experiments, they are in a similar range with the exception of the NMPrelim experiment (a direct effect of users working only through 2 of the 5 problems). Interestingly, POST scores are not affected in the Main experiment by the limitation of the instruction to walkthrough dialogues and disabling of the essay interpretation component.

Table 5. PRE/POST scores for all corpora

Experiment	Condition	PRE	POST
F03	F03	12.5 (4.4)	17.9 (4.6)
S05	S05SYN	12.5 (4.4)	18.2 (4.2)
	S05PR	13.5 (3.3)	18.6 (3.5)
NMPrelim	F	12.8 (4.8)	15.2 (4.6)
	S	12.0 (3.6)	14.6 (3.6)
Main	*R*	12.6 (4.2)	19.1 (3.1)
	PI	12.5 (4.4)	19.4 (2.8)
	NM	12.6 (4.1)	18.8 (3.9)

2.3.1.2 Learning metrics

The primary performance metrics for tutoring systems is learning due to interaction with the system. Other metrics are also important but secondary to learning (e.g. user satisfaction – Section 2.2.3.1, dialogue efficiency – see Sections 2.3.3.3 and 2.3.3.4). There are several ways of measuring learning. All use the PRE/POST scores but differ in terms of the perspective they offer.

The simplest way to measure learning is the **POST** score. Ideally, we would like to see all users achieving the perfect test score. This metric disregards the fact that users have different levels of knowledge when they come in. As a result, improvements are not measured: i.e. it treats the same users with the same POST score regardless of how high or low their PRE score was. For example, it treats the same users that reached a POST of 20 even if they started with a PRE of 6 or 19.

To account for the PRE score, other metrics look in various ways at the distance between the PRE and the POST score. Learning gain is defined as the arithmetic difference between the POST score and the PRE score. However, this metric assigns similar learning to users that improve from a PRE of 6 to a POST of 7 and users that improve from a PRE of 19 to a POST of 20. In addition, it suffers from a "ceiling effect": users with a higher PRE score can improve less than users with a lower PRE score (e.g. users with a PRE of 20 have a maximum learning gain of 6 while users with a PRE of 6 have a maximum learning gain of 20). For these reasons, we will not use this metric in our analyses.

The **Normalized Learning Gain** (**NLG**) fixes the learning gain issues by normalizing the distance between PRE and POST to the distance between PRE and the maximum POST. In our case NLG = (POST-PRE)/(26-PRE). When PRE/POST scores are measured as percentage (e.g. in our case, PRE divided by 26), NLG is defined as (posttest-pretest)/(1-pretest). In effect, this metric measures the percentage improvement relative to the perfect improvement: an NLG of 0.0 means no improvement, an NLG of 0.5 means we are half-way there while an NLG of 1.0 means maximum improvement.

However, even NLG has an important drawback: stability issues for higher PRE scores. More specifically, the higher the PRE score, the more sensitive is the NLG score to small variations in the POST score. For example, if a student starts with a PRE of 22 and achieves a POST score of 25, the NLG will be 0.75. However, if the student would miss one of the 25 correctly answered questions (e.g. lost concentration for that problem or was misled by the problem text), the NLG will become 0.50. Thus, one small user mistake transforms the user from a high learner to a medium learner (average NLG is around 0.5 in most corpora). The same is not true of users with a lower PRE score. One way to address this issue is to eliminate from analyses users with higher PRE score (see Section 2.3.3.2).

When comparing a control condition and an experimental condition in terms of NLG, a common metric that is reported in pedagogical studies is the **effect size** (e.g. (Bloom, 1984)). Effect size is defined as (average NLG experimental – average NLG control)/(standard deviation NLG control) and measures the improvement offered by the experimental condition. An effect size of 1.0 is approximate of one letter grade improvement. An effect size of 2.0 for adult tutoring in replacement of classroom instruction (Bloom, 1984) has been the main catalyst behind work on computer-based tutors.

As another alternative to measuring learning, we can apply the ANCOVA test and use the **adjusted posttest scores** (posttest scores that account for the pretest score). Details will be discussed in Section 2.4.4.

To test if certain phenomena are associated with learning, **correlations** and **partial correlations** are typically used. Details will be discussed in Section 2.4.2.

2.3.1.3 User turn transcripts

During the interaction with ITSPOKE, there is a spoken dialogue between the system and the user: the system asks questions and the user answers back. Two transcripts of the user turn are available: system and manual transcript. The system transcript is obtained by running the user speech through the Automated Speech Recognition (**ASR**) component – Sphinx II. The top recognition hypothesis is used. This transcript is interpreted by the Why2-Atlas backend in terms of semantics to determine the appropriate system response.

Because the system transcript in not perfect, after each experiment, a human annotator transcribed all user turns. A web interface was used in which the annotator could listen to the student speech and transcribe the content. Non-linguistic events (e.g. laughs, coughs, sighs, background noise, etc) were ignored when transcribing.

Figure 6 shows an example of the system transcript (ASR lines) and the human transcript (STD lines).

2.3.1.4 Correctness

ITSPOKE uses the Why2-Atlas backend to drive the conversation. The Why2-Atlas backend has a semantic interpretation component which identifies concepts in the user input. Based on what concepts are present and the authored tutoring information, Why2-Atlas will decide if additional statements or questions are needed before moving on to the next question. A deterministic procedure that uses the output of the semantic interpretation component and the authored tutoring information was developed to assign 3 labels of correctness to any user input: *Correct*, *Partially Correct* and *Incorrect*. The system can ask the student to provide multiple pieces of information in her answer (e.g. the question “Try to name the forces acting on the packet. Please, specify their directions.” asks for both the names of the forces and their direction). If the student answer is correct and contains all pieces of information, it was labeled as Correct (e.g. “gravity, down”). The Partially Correct label was used for turns where part of the answer was correct but the rest was either incorrect (e.g. “gravity, *up*”) or omitted some information from the ideal correct answer (e.g. “gravity”). Turns that were completely incorrect (e.g. “no forces”) were labeled as Incorrect.

Depending on the user turn transcript (2.3.1.3) that is fed to the semantic interpretation component, two versions of correctness can be automatically computed: ASR correctness (**ASEM**) and transcript correctness (**TSEM**). An additional correctness label “Unable to Answer” was created to automatically mark turns where the user used either variations of “I don’t know” or simply did not say anything.

Figure 6 shows the two versions of correctness (ASEM and TSEM labels). In cases where due to speech recognition problems the system and human transcripts differ enough, the ASEM and TSEM label can be different (e.g. Figure 6, STD_1).

2.3.1.5 Speech recognition problems (SRP)

Three types of SRP have been annotated in the corpus: Rejections, ASR Misrecognitions and Semantic Misrecognitions. **Rejections** occur when ITSPOKE is not confident enough in the recognition hypothesis thus discarding the current recognition and asking the student to repeat (e.g. Figure 6, STD_3). When ITSPOKE recognizes something different than what the student actually said (i.e. human transcript is different from the system transcript - 2.3.1.3) but was confident in its recognition hypothesis, we call this an **ASR Misrecognition** (e.g. Figure 6, $STD_{1,2}$).

Semantic accuracy is more relevant for dialogue evaluation, as it does not penalize for word errors that are unimportant to overall utterance interpretation. For ITSPOKE, the semantic interpretation is defined in terms of correctness (2.3.1.4). We define **Semantic Misrecognition** as cases where ITSPOKE was confident in its recognition hypothesis and the correctness interpretation of the system transcript (ASEM) is different from the correctness interpretation of the manual transcript (TSEM) (e.g. Figure 6, STD_1).

2.3.2 F03 experiment

The F03 corpus has been used in several previous studies which have enriched the corpus with a number of manual and automatic annotations. Two of them will be used in this work: human correctness and certainty. Other annotations not used here include: generic emotion annotation (Litman and Forbes-Riley, 2004), user acoustic/prosodic profiles (Forbes-Riley, 2005), clustering of tutor goals (Ai and Litman, 2007), concept repetition (Tetreault and Litman, 2006), identification of domain-specific concepts (Purandare and Litman, 2008), etc.

2.3.2.1 Human correctness

To eliminate the noise introduced by the automated speech recognition component and the semantic interpretation component, a human annotation of the correctness (**HSEM**) was performed on the F03 corpus. The annotator used the human transcripts and his physics knowledge to label each student turn with one of the 4 labels of correctness described in Section 2.3.1.4. A comparison of HSEM and TSEM results in an agreement of 90% with a Kappa of 0.79 (Carletta, 1996).

2.3.2.2 Certainty annotation

While in most computer tutors student correctness is used to drive the conversation, other factors might be of importance. Student certainty is hypothesized to play an important role in the learning and tutoring process. Researchers hypothesize that student uncertainty creates an opportunity for constructive learning to occur (VanLehn et al., 2003) and studies have shown a positive correlation between uncertainty and learning (Craig et al., 2004). (Forbes-Riley and Litman, 2005) show that student certainty interacts with a human tutor's dialogue decision process (i.e. the choice of feedback).

A human annotator has annotated certainty in all F03 user turns. The annotation manual asks the annotator to label based on the perceived uncertainty or confusion about the material being learned expressed by the user. The annotator used the audio recordings of each user turn. Four labels were used:

certain, *uncertain* (e.g. Figure 6, STD_1), *mixed* and *neutral*. In a small number of turns, both certainty and uncertainty were expressed and these turns were labeled as mixed (e.g. the student was certain about a concept, but uncertain about another concept needed to answer the tutor's question). To test the reliability of the certainty annotation, a second annotator was commissioned to annotate the corpus for the presence or absence of uncertainty (a binary version of the initial certainty annotation). We performed a *binary* comparison of the two annotations. The 4-way certainty annotation was converted to a binary annotation: uncertain vs. others (certain, mixed and neutral). The comparison yields an agreement of 90% with a Kappa of 0.68.

2.3.3 Main Experiment

For the Main corpus we first show how we partitioned the population into a high and a low split based on metrics of interest (2.3.3.1). Then, we identify a particular subset of interest in this corpus (2.3.3.2). We also computed a series of other metrics (2.3.3.3 and 2.3.3.4). Note that these metrics, the subset and the splitting procedure are not specific to this corpus and can be reproduced in any other ITSPOKE corpora. We only needed them for the analyses that compare the three conditions from the Main experiment: *R*, *NM* and *PI*.

2.3.3.1 High-low splits

It is a common analysis practice in tutoring research to investigate the relationship between user aptitudes and his/her performance while working with the system (e.g. (McNamara and Kintsch, 1996; VanLehn et al., 2007; VanLehn et al., 2005; Ward and Litman, 2006)). Several studies have shown that the treatment condition can produce effects only on specific subsets of the populations and that, in some cases, the treatment has opposite effects depending on the subset. Typically, the subsets are generated by splitting the user population based on the mean or median of the aptitude metric: a high subset and a low subset. We used a mean split in this work.

One of the aptitudes measured in this study is the initial physics knowledge (the PRE score – see 2.3.1.1). **PRE Split** was generated by splitting users using the mean PRE score in the Main Experiment (Mean PRE = 12.5). Table 6 shows the number of users in each condition for each low/high split. Details for the other aptitude metric, the working memory span, are available in Appendix A.4.3.

Table 6. Number of users in each PRE split subset

	PRE Split	
Condition	Low	High
R	15	10
PI	16	11
NM	16	11

2.3.3.2 RELNLG Subset

In Section 2.3.1.2 we mentioned that the NLG score is less stable for users with a higher PRE. Thus, in our learning-related analyses we will also investigate what happens if we eliminate these users. But what is a good cutoff threshold? We decided to use the average POST score in the *R* condition (19.1 in Table 5) as the cutoff threshold. From a learning perspective, users with an initial knowledge (PRE score) higher than the average post-instruction knowledge (POST score) are less interesting: they know already more than what the population will achieve on average. Since in this experiment the instruction was not tailored to user essays (see Section 2.2.4), it is very likely that the system discussed with these users a lot of things they already knew. As another argument, even if these users do not work with the system, it is likely they will achieve a higher POST score than average since the pretest and posttest are isomorphic. Other studies remove users with a perfect PRE score (e.g. (Chi and VanLehn, 2008)).

The remaining subset (i.e. users with a PRE score of 19 or less) is called the **RELNLG subset** (reliable NLG). The subset contains only 22 of the 25 *R* users, 25 of the 27 *NM* users and 25 of the 27 *PI* users, for a total of 72 users. As we will see, all of our learning-related results become clearer on this subset.

Note that the PRE Split was not recomputed for the RELNLG subset (i.e. we did not split based on the mean PRE on the RELNLG subset). As a result, when looking at PRE Split, all the low pretesters subsets remain the same; the removed students come out of the high pretesters subsets.

2.3.3.3 Dialogue time

We define the dialogue time as the amount of time spent by users conversing with the system in each problem. Thus, dialogue time is defined as the duration between the start time of the first tutor turn in the dialogue for a given problem and the end time of the last tutor turn in that dialogue. Note that we could have used the total time spent on each problem however, besides the dialogue time, this duration also includes the time user spent reading the problem and typing the initial essay and its revision. In addition, users were told they can take breaks during the essay part. Thus the total time spend on each problem is a less reliable measure.

We compute the dialogue time for each problem (**P1Time-P5Time**) and a total dialogue time as the sum of the five problem dialogue times (**TotalTime**).

The dialogue time is influenced by two factors: the recognition performance and user correctness. Since in the Main experiment the dialogue was not tailored to the user initial essay, all users went through the same instruction for each problem. Deviations from this plan are only due to speech errors (i.e. rejections and timeouts) and/or incorrect answers (e.g. an incorrect answer will engage a remediation subdialogue for certain questions). Thus the dialogue time is a good measure of overall correctness of the user and how well they are recognized by the system. The latter is in turn a result of the ASR performance and user's ability to adapt his answers to what the system expects them to say (especially for correct answers).

The time users spend with a system is an important metric as researchers strive for learning efficiency when designing tutoring systems: deliver increased learning as fast as possible. When no improvements in learning over a baseline are observed, it is a positive result to achieve the same amount of learning in a shorter time spam (e.g. (VanLehn et al., 2007)).

2.3.3.4 Number of system turns

We define number of system turns as the total number of turns the system has uttered in the dialogue for a problem. Note that this number will also include system turns that deal with speech errors (i.e. repetition of the last turn for timeouts and "Could you please repeat that" turns for rejections). Also, a tutor turn can include one or more goals (e.g. for some incorrect users answers, the next system turn will include a correction of the incorrect answer and the next question).

We compute the metric for each problem (**P1Tut-P5Tut**) and as a sum over all problems (**TutTotal**).

The difference between the dialogue time metric (2.3.3.3) and this metric is that the number of system turns metric ignores the duration of the system turn and the duration of the user turns[8], where duration is dependent on the number of words in these turns and the speaking rate. In addition, the number of turns stays the same regardless of the system correctness of the user answer for questions that do not require remediation subdialogues. In general, an increase in number of turns is due to extra remediation dialogue. In these cases, the metric incorporates the size of the remediation dialogue in system turns. Rejections due to speech problems can also increase the number of turns through additional

[8] The duration of the user turns represents a very small proportion of the total dialogue time in ITSPOKE (about 6%). In the Main experiment, the average total dialogue time is 2550 seconds out of which 147 seconds is the average total user turn duration.

system turns that handle these rejections (e.g. "Could you please repeat that?" system turns). Another phenomenon that increase the number of system turns is timeouts (i.e. when users do not answer the question in the allotted time). In such cases, the system simply repeats the question. Note that the number of user turns is similar to the number of system turns since the interaction follows a question-answer format where the system asks questions and the user has to answer.

2.4 STATISTICAL TOOLS

All the analyses in this work were performed using the Statistica software package[9]. We provide a quick introduction to some of the statistical tools we used focusing on how to interpret our presentation of the results.

2.4.1 Outliers

Outliers are by definition atypical, infrequent observations which do not appear to follow the characteristic distribution of the rest of the data. Occurrence of outliers may reflect genuine properties of the underlying phenomenon, or be due to measurement errors or other anomalies which should not be modeled. One of the users we removed from the Main experiment corpus (2.2.4.1) is a good example of an anomaly which we do not want to model. The user has a large number of speech errors due to speaking very softly and/or keeping the microphone too far. Since ITSPOKE had a hard time getting the actual answer from this user, the user was forced to go through more instruction than warranted. This resulted in a considerably larger instruction time compared to other users as can be observed visually in Figure 5.

Outliers can have a profound effect on certain statistics. For example, in linear regression, a single outlier is capable of considerably changing the slope of the regression line and, consequently, the value of the correlation. Thus, detecting and removing outliers is common practice in data analysis.

Several methods can be used to label and detect outliers. In our analyses we used the box-plot rule method as implemented in the Statistica software package. According to this method, outliers are points for which the distance from the closest of the 75th or 25th percentile is larger than 1.5 times the distance between the 25th and 75th percentile. Figure 7 shows this method visually for the outlier discussed above (recall Figure 5). The 25th-75th percentile range is shown as a rectangle with the median in the

[9] http://www.statsoft.com/

middle. The plot whiskers are the non-outlier range according to the definition above (e.g. the top whisker reaches 75th percentile + 1.5 * (75th percentile – 25th percentile)). We observe that the user we discussed is the only point outside the outlier range.

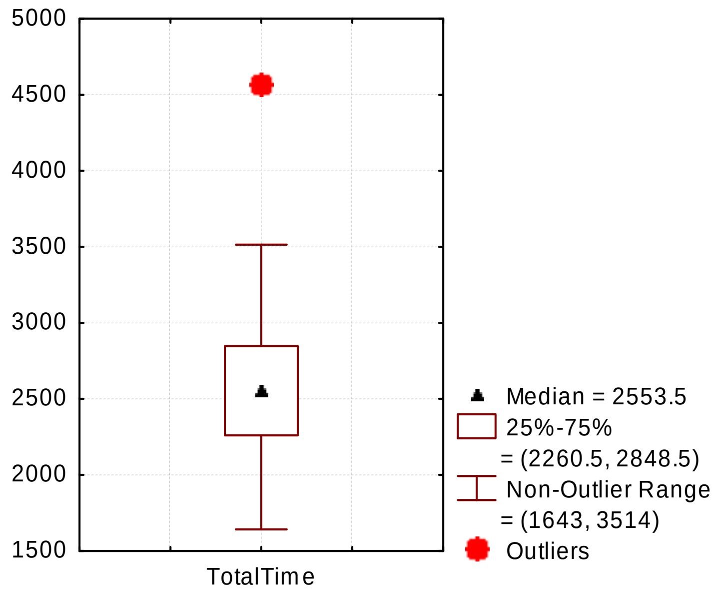


Figure 7. Box plot method for detecting outliers (TotalTime)

Whenever we find outliers in our analyses (2.2.4.1 and 6.4.2.3) using the box-plot method, we will show in this document the distribution and the z-score. The z-score is another method for labeling outliers. It has the advantage of being intuitive and easy to interpret. The z-score is defined as the distance between a point and the mean in terms of standard deviations. A commonly used heuristic for labeling outliers is a z-score of 3.0 or higher[10] (Iglewicz and Hoaglin, 1993).

A quick tutorial on outliers written by Agata Fallon and Christine Spada is available online[11]. For more details see (Iglewicz and Hoaglin, 1993).

---

[10] The z-score heuristic is less robust than the box-plot rule method because the outlier is included in the computation of the mean and standard deviation for the z-score. In contrast, the box-plot method does not have this problem because it uses ranks instead of means (i.e. the 25% and 75% percentile are not affected by how far the outlier is from the mean).

[11] http://www.cee.vt.edu/ewr/environmental/teach/smprimer/outlier/outlier.html

2.4.2 Correlations and partial correlations

There are many situations in which we would like to study the relation between two variables: are they related in any way or does one variable predict the other? For example, we might be interested to know if the time users spend with our system tells us anything about the system performance (i.e. learning - 2.3.1.2) or if user's correctness is predictive of his/her learning.

One of the most commonly used methods is to look at the linear relation between two variables using the Pearson's **correlation**. A correlation coefficient is computed (**R**) that measures the strength of the linear relationship between two variables using a linear regression. The correlation coefficient ranges between -1 and 1, where a value of -1 or 1 means a perfect linear relation between the two variables (e.g. Y=2*X+1) and 0 means no significant linear relation. The sign of the correlation coefficient tells us about the type of relation: positive correlation ($R>0$) or negative correlation ($R<0$). Intuitively, a positive/negative correlation means that an increase for one variable is associated with a proportional increase/decrease for the other variable. Figure 8 shows the correlation between the pretest score and the posttest score (PRE, POST - 2.3.1.1) in the Main experiment (2.2.4) and their linear relation (the regression line). Each point in the graph represents one or more users with the corresponding PRE and POST scores (larger points represent 2 or more users). There is a significant positive correlation between the two variables ($R=0.63$, $p<0.001$): users with a higher PRE score tend to have a higher POST score.

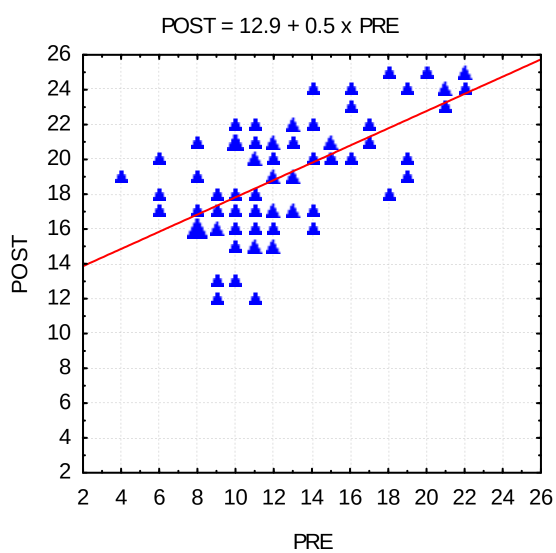


Figure 8. PRE-POST correlation in the Main experiment

To investigate correlations between a variable and learning a methodology commonly used in tutoring research (e.g. (Chi et al., 2001)) is to look at Pearson's **partial correlation** between the variable and the posttest score that accounts for the pretest score. Accounting for the pretest score in the posttest score (i.e. the residuals from a linear regression) is a way of measuring learning because we are discounting for levels of knowledge users have when they come in (as measured by the pretest score). A partial correlation coefficient (R) is reported for partial correlations. Its interpretation is similar to a normal correlation coefficient between the variable of interest and learning as the second variable.

Note that correlations do not imply causality. There might be other confounding variables that are responsible for the observed linear relation between two variables.

2.4.3 ANOVA

ANOVA (**AN**alysis **O**f **VA**riance) is a statistical model that analyzes the relative contributions of explained and unexplained sources of variance in a continuous response variable. The explained sources are called factors. ANOVA is used for categorical factors. When continuous factors are needed, an extension of ANOVA called ANCOVA (2.4.4) is used.

ANOVA tests for significant differences between means using the F-statistic. The means come from partitioning the response variable based on the factors. The continuous response variable is also called the dependent variable; the factors are also called independent variables or predictors. For example, we would like to know if users exposed to two versions of a system achieve different posttest scores (POST score - 2.3.1.1). To test this hypothesis, we run an ANOVA with POST as dependent variable and the condition as a categorical factor. ANOVA will tell us if the factor has a significant effect on the dependent variable. It effectively compares the mean POST of the subpopulation that used one version with the mean POST of the subpopulation that used the other version of the system.

After the effect of a factor or a combination of factors has been observed, posthoc tests are usually performed to understand how the effect works on the dependent variable. These tests compare the means in pairs. In our analyses we used the Fisher LSD test from Statistica which is equivalent to the t-test. Other more conservative posthoc tests can be used (e.g. tests that account for the number of comparisons like the Bonferroni test) but reaching significance on these tests is harder given our sample size and the effects we observe. For more details about ANOVA and its applications see (Doncaster and Davey, 2007).

We will briefly describe here how to interpret our presentation of the ANOVA results. We apply ANOVA in 3 ways in our analyses: one-way ANOVA, factorial ANOVA and repeated-measures ANOVA.

One-way ANOVA is used to measure the effect of a single categorical factor on a dependent variable. In our analyses, we will use only factors with two values. For binary factors, one-way ANOVA is equivalent to the t-test. Thus a significant effect of the factor translates into a significant difference between the two groups that come from partitioning based on the factor values. For example, in Section 6.4.2.5, we investigate the effect of the condition (i.e. the version of the system) on the total dialogue time. For that, we run a one-way ANOVA with TotalTime as dependent variable and Condition as the categorical factor (*NM* vs. *R* - 2.2.4). We find a significant effect of the Condition on TotalTime ($F(1,50)=8.31$, $p<0.006$). Since the average TotalTime for *NM* users is 2395 and for *R* users is 2671 (Table 35), we conclude that the total dialogue time for *NM* users is significantly shorter than for *R* users.

Factorial ANOVA is used to measure the effect of two or more factors. In our analyses, we will only use two binary factors (a 2×2 design): a two-way ANOVA. For two-way ANOVAs, we are interested if the combination of the two factors has any effect on the dependent variable. The effect of individual factors can be measured using a one-way ANOVA. For example, we saw above that the Condition has a significant effect on TotalTime. We might be interested to know if this effect is influenced by whether users have a lower or a higher pretest. For that, we run a factorial ANOVA with TotalTime as the dependent variable and two factors: the Condition (*NM* vs. *R*) and the PRE Split (Low vs. High - low vs. high pretesters in 2.3.3.1)). We find a trend effect of the combination Condition × PRE Split on TotalTime. The easiest way to understand this effect is through a graphical representation. Figure 9 shows the average TotalTime for each of the 4 groups created by the two factors (low *NM*, high *NM*, low *R* and high *R*). The 95% confidence intervals for each mean are also shown. Two confidence intervals that do not intersect or intersect minimally are an indication of a statistical significant difference between the two means according to our post-hoc test (Fisher LSD). We observe from the graph that the effect of the Condition on TotalTime depends on the PRE Split. It is for high pretest users where the Condition has the greatest effect (significant difference between high *NM* and high *R*). There is a small effect for low pretest users but far from significance. Visually, when the combination has an effect the two lines in the graph should have very different slopes.

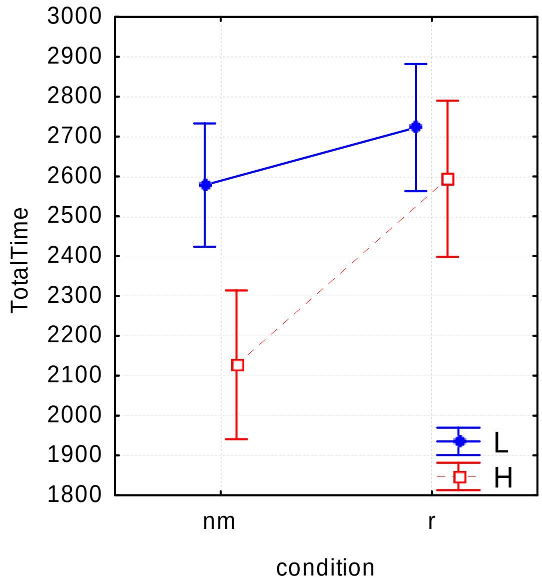


Figure 9. Interpretation of a sample 2x2 factorial ANOVA

Whenever we find a trend/significant effect of a combination of factors in a factorial ANOVA we will present them graphically as in Figure 9 and include in the text the significance of the posthoc tests.

Repeated measures ANOVA is used when we want to administer the same test to the same subjects repeatedly over a period of time. For example, in one of our experiments (NMPrelim - 2.2.3) subjects were asked to rate two versions of the system using the same questionnaire. This corresponds to an ANOVA with one within-subjects factor: the version of the system. A significant effect in this example means that the version of the system has a significant effect on the rating (see 6.3.2.1). In some analyses, we include in addition a between-subjects factor. For example, in all our experiments all subjects are administered a physics test twice: the pretest and the posttest (note that the tests are isomorphic/comparable - 2.3.1.1). However, each user worked with a specific version of the system. This corresponds to an ANOVA with one within-subjects factor (the Test Phase) and one between-subjects factor (the Condition). A significant effect of the Test Phase factor means that there are significant differences between the pretest and posttest (i.e. if posttest > pretest then we observe learning). A significant effect of the combination Test Phase × Condition means that learning depends on the condition and it can be visualized graphically similarly to Figure 9 (though we never find these effects in our data).

2.4.4 ANCOVA

ANCOVA (**AN**alysis of **COVA**riance) is an extension of ANOVA that allows the use of continuous factors. It combines ANOVA with regression and tests the effect of other factors on the dependent variable after removing the variance accounted for by the continuous factor. ANCOVA is applied whenever in an experiment we can not control for a continuous factor that influences our dependent variable: e.g. the level of intelligence/knowledge of the subject as measured by a specific test. In our experiments, the pretest score (2.3.1.1) is an example of a continuous factor that we can not control for.

Many tutoring studies (e.g. (VanLehn et al., 2007)) use ANCOVA to study the effect of factors on learning. In our work we also use ANCOVA to study learning. We describe below how the test is applied and how to interpret the results. As discussed in Section 2.3.1.2, using the posttest score to measure learning does not account for the fact that users have different levels of knowledge when they come in (as measured by the pretest score). Indeed, in all our corpora as in many other studies, the pretest score is positively correlated with the posttest score. To better measure learning, we can account for the pretest score in the posttest score by running an ANCOVA with posttest score as the dependent variable and pretest score as the continuous factor. Any number of factors for which we want to study the effect on learning (e.g. the experimental condition, aptitude) can be added to this instantiation of ANCOVA as categorical factors.

There are two ways of interpreting the ANCOVA results: via adjusted posttest scores or via regression lines. In our analyses we will present both as they offer slightly different perspectives on the effects.

ANCOVA works by factoring out the pretest score from the posttest scores. The resulting scores are called the **adjusted posttest scores**. First, a linear regression between pretest and posttest is performed in the population and residuals are computed. Residuals are the differences between the posttest score predicted by the linear regression and the observed posttest score for each user. The adjusted posttest score is computed by adding the posttest average and the residual and can be seen as another measurement of learning (2.3.1.2). Next, an ANOVA (2.4.3) is run to study the effect of factors on these adjusted posttest score. All interpretations of ANOVA that we discussed in Section 2.4.3 apply.

Another way of interpreting ANCOVA is through regression lines. ANCOVA tests if factors affect significantly the relationship (i.e. correlation/regression) PRE-POST. The population is partitioned based on the factor(s) and a regression line is computed for each group. A significant ANCOVA effect means that the regression lines are different: either significantly different slopes, or if no significant difference in slope, then significant intercepts. This effects are best viewed graphically by plotting the PRE and POST score and the corresponding regression lines. Figure 10 shows a sample one-way

ANCOVA with the PI Split as the independent variable (Low vs. High – low vs. high pretesters in 2.3.3.1). The test is discussed in more detail in 4.5.4.4. For this test, we find a trend effect of the PI Split (F(1,47)=3.71, p<0.06), which means that there is a trend difference between the two regression lines. When comparing two pretest-posttest regression lines, a regression line with a higher intercept and a smaller positive slope is more desirable since the smaller slope is an indication of the fact that the pretest score has a smaller effect on the posttest score (i.e. most people will reach a similar posttest score); a higher intercept means that the posttest scores achieved are higher. By looking at the scatter plot and the regression lines we can visualize better the performance across a range of pretest scores.

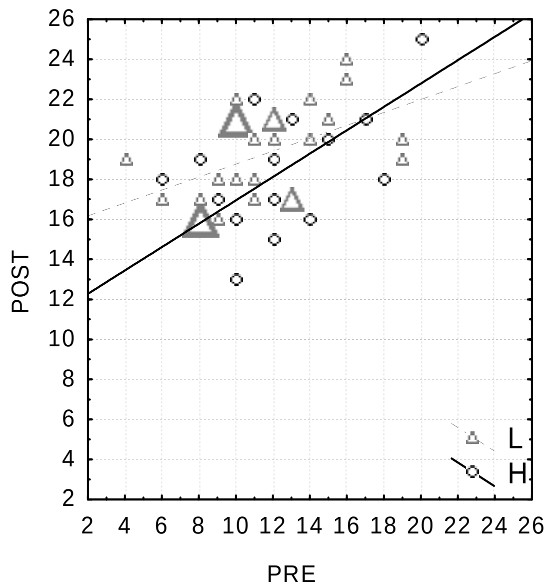


Figure 10. Sample scatter plot and regression lines for a one-way ANCOVA

3 DISCOURSE STRUCTURE

3.1 THEORY

There is more to language than word and sentence level phenomena (e.g. phonology, morphology, syntax, semantics, etc). Indeed, "language does not consist of isolated, unrelated sentences, but instead of collocated, related groups of sentences" (Jurafsky and Martin, 2000). Discourse is a generic terms used to refer to such a group of sentences. Based on whether the communication of information is unidirectional or bi(multi)directional, there are two types of discourse: monologues and dialogues. This document is an example of a monologue. The conversation between ITSPOKE and users is an example of a dialogue.

The same way words are organized in a sentence according to a structure (e.g. syntactic, semantic), sentences are organized in a discourse according to a structure called the **discourse structure**. Models of discourse structure attempt to explain why a sequence of random sentences combines to form a discourse or no discourse at all. For example, at the higher level, this document follows the structure outlined in the table of contents.

In this work we will use the Grosz & Sidner theory of discourse structure (Grosz and Sidner, 1986). This theory models discourse structure using three interacting components: the linguistic structure, the intentional structure and the attentional state. The linguistic structure contains the utterances in the discourse grouped in **discourse segments**. The intentional structure identifies the discourse-relevant purpose of each discourse segment and the relationships between these segments. The attentional state is a dynamic model of objects, properties and relationships that are salient at each point in the discourse. The theory can be used to explain discourse phenomena like cue phrases (Passonneau and Litman, 1997) and referring expressions (Grosz et al., 1995).

In Figure 11 we provide a quick introduction to the Grosz & Sidner theory of discourse structure and discuss the elements that we will use in this work. The figure shows an excerpt from a conversation with ITSPOKE (left) and the annotation of the relevant discourse structure information (text highlights on the left and the right part of the figure). The conversation discusses the problem from Figure 1 and follows the problem plan from the same figure. According to the theory, each discourse has an associated **discourse purpose**. Intuitively, the discourse purpose is the intention/purpose behind engaging in the particular discourse rather than taking other actions. In our example, the discourse purpose is the system's intention to walk the user through the solution to the problem. It is shown on the right in Figure 11 marked as "Solution walkthrough". To achieve the discourse purpose, a number of "smaller" intentions/purposes have to be achieved. In our example, the system needs to establish two time frames (purpose: "Two time frames: before release, after release") and then discuss what happens in each time frame (purpose: "Before release" and "After release") in order to achieve its discourse purpose. These purposes/intentions naturally aggregate the sentences in **discourse segments**, with each discourse segment having an associated **discourse segment purpose/intention**. We can see in Figure 11 how the first two sentences in $Tutor_1$ aggregate in a discourse segment that has the purpose to establish the two time frames ("Two time frames: before release, after release").

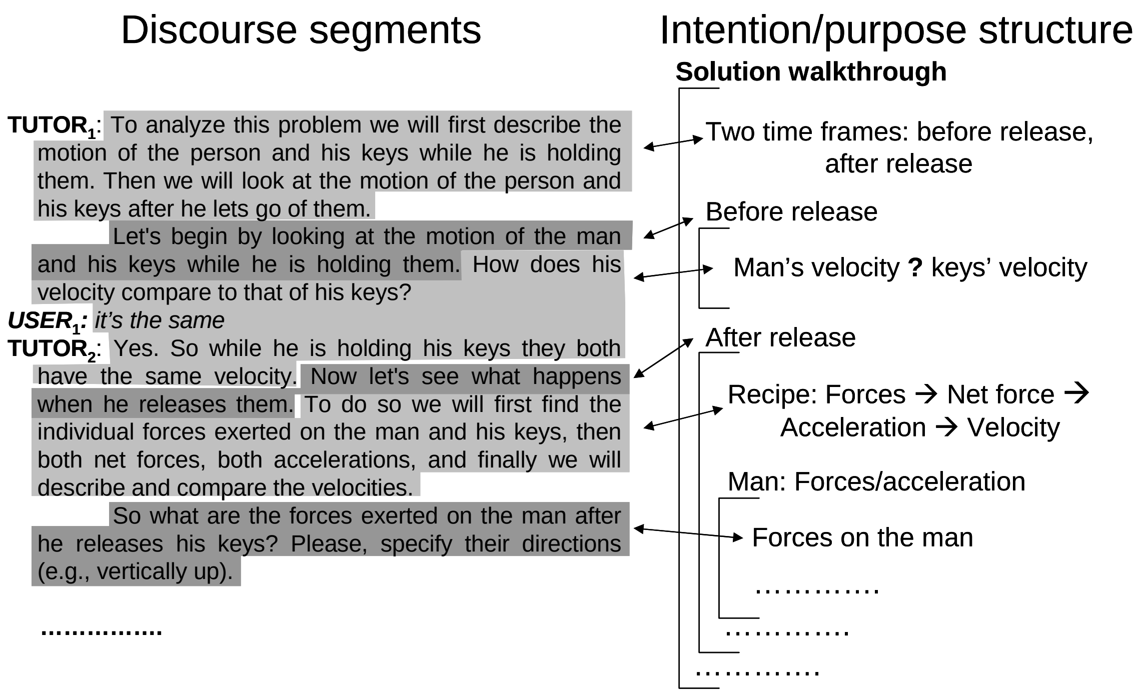


Figure 11. Elements from the Grosz & Sidner theory of discourse structure

There are two structural relations between discourse segments which specify how their purposes relate to each other and to the overall solution: a satisfaction-precedence relation and a dominance relation. The satisfaction-precedence relation enforces a partial ordering of the discourse segments. In Figure 11 we represent this relation graphically via the top to bottom ordering of discourse segment purposes. For example, we can not talk about what happens in each time frame before introducing the two time frames (i.e. "Two time frames: …" has to be satisfied before "Before release" and "After release"). The dominance relation identifies discourse segment purposes that contribute to another discourse segment purpose. In Figure 11 we represent this relation graphically through indentation and bracketing. For example, to achieve the purpose of discussing what happens in the after release time frame ("After release"), the system presents the recipe of discussion ("Recipe: Forces …") and then discusses individual concepts (e.g. "Forces on the man"). As a result of these relations, discourse segments can be arranged in a hierarchical tree-like structure which we refer to as **discourse segment hierarchy**. For other elements of the theory not discussed here, we refer the reader to the paper (Grosz and Sidner, 1986).

The implications of the Grosz & Sidner theory of discourse structure have been investigated for a variety of research problems. In monologue settings, researchers have used discourse structure for understanding specific lexical and prosodic phenomena (Hirschberg and Nakatani, 1996), natural language generation (Hovy, 1993), essay scoring (Higgins et al., 2004), etc. For dialogues, discourse structure was used for modeling prosodic cues (Levow, 2004), predictive/generative models of postural shifts (Cassell et al., 2001), generation/interpretation of anaphoric expressions (Allen et al., 2001), etc. In terms of SDS architectures, Grosz & Sidner theory has laid the foundation for expressive dialogue managers (Bohus and Rudnicky, 2003; Horvitz and Paek, 1999; Rich et al., 2001). Many of these dialogue managers borrow concepts from this theory: hierarchical task representation, a stack (agenda) to maintain entities in focus, etc.

We investigate the utility of discourse structure for a variety of SDS tasks. In the next section we will identify which elements from the Grosz & Sidner theory we use and why.

We would like to mention that there are other competing theories of discourse structure. Rhetorical Structure Theory (**RST**) (Mann and Thompson, 1987) is primarily used for monologues. Similar to Grosz & Sidner, this theory looks at spans of text and the (rhetorical) relationships that hold between them (e.g. contrast, elaboration, result, support, etc). Differences between the two theories (e.g. RST combines discourse and domain/world information) are discussed in §7.4 of (Grosz and Sidner, 1986). The main reason behind our choice for the Grosz & Sidner theory is the fact that they use only two structural relationships between discourse segments (dominance and satisfaction-precedence). As a result, manual and automatic annotation of discourse structure is less complex and raises less issues compared to an RST annotation of dialogue (Stent, 2000). For dialogue settings, dialogue acts (e.g. (Core and Allen, 1997)) have seen a lot of applications as features for predictive models of linguistic/paralinguistic phenomena (e.g. (Stolcke et al., 2000)). However, dialogue acts look only at lower-level communicative relationships (e.g. question-answer pairing) and for the applications we had in mind we were interested in higher level relationships.

3.2 PROMISING ELEMENTS

In order to identifying promising elements from the Grosz & Sidner theory of discourse, we had to answer the following question: why would discourse structure have any use for SDS? We had two answers for this question.

The first answer is based on our intuition that the "*position in the dialogue*" is of importance. For example, we believed that when it comes to speech recognition problems or user affect, the position in the dialogue should tell us something about the likelihood of encountering those phenomena (characterization of dialogue phenomena application – Section 5). As another example, if we want to predict how much an user has learned by looking at his/her dialogue with our system, we believed that user's answers have more weight depending on their position in the dialogue (performance analysis application - Section 4).

But how to define "position in the dialogue"? This is where discourse structure comes into use. Before we explain how discourse structure was used, we need to spell out the requirements for our definition. We wanted our definition to be *domain-independent* and *automatically computed*. The first property ensures that our definition can be applied to other domains without any changes. The second property stems from the nature of the applications we envisioned (performance analysis and characterization of dialogue phenomena): all these applications have runtime implications thus availability of that particular discourse structure information at runtime will enable us to apply in practice the findings from the offline analysis.

We propose using discourse structure transitions to define the "position in the dialogue". This information source exploits the discourse segment hierarchy by identifying certain types of transitions that occur in this hierarchy as the dialogue advances. It also satisfies the two requirements. "Position in the dialogue" will be defined based on the current discourse structure transition. Discourse structure transitions require an automatic annotation of the discourse segment hierarchy and an automatic way of computing the transitions which will be discussed in Section 3.3. In later sections we will show how this definition of the "position in the dialogue" and the above intuition can be used for two important SDS tasks: performance analysis (Section 4) and characterization of dialogue phenomena (Section 5).

As another way of using discourse structure in SDS, we hypothesized that users will benefit from direct access to the discourse structure information. We believed that users will benefit from having explicit access to the intention/purpose behind the current discourse segment and the relationships between the current discourse segment and previous or future discourse segments. We enable this access through a graphical representation of discourse structure, the Navigation Map, which we describe in Section 3.4 along with a manual annotation of discourse segment purpose/intention and of the discourse segment hierarchy. Support for this hypothesis will be discussed further in Section 6.1; the utility of the Navigation Map will be investigated in Sections 6.3 and 6.4.

Previous work in the SDS tasks we investigate makes little use of the discourse structure information. In the Related Work section for each of the applications we investigate (4.6, 5.6 and 6.5), we will compare our use of discourse structure with previous work.

3.3 AUTOMATIC ANNOTATION & TRANSITIONS

As mentioned in Section 3.2, we believe that the "position in the dialogue" is an important information source for several SDS tasks. We will use the notion of discourse structure transition[12] to define the "position in the dialogue". We exploit the discourse segment hierarchy by identifying certain types of transitions that occur in this hierarchy as the dialogue advances. For transitions to satisfy the automatically computed property, we need an automatic annotation of the discourse segment hierarchy. We first describe our automatic approximation of the discourse structure hierarchy and then we describe how we compute the discourse structure transitions.

The discourse segment hierarchy requires the identification of all discourse segments in the dialogue and their nesting structure. Note that other elements from the Grosz & Sidner theory of discourse (e.g. discourse segment intention/purpose, the attention stack) are not needed. We argue that the discourse structure hierarchy or at least an approximation of it can be automatically obtained in dialogue systems with dialogue managers inspired by the Grosz & Sidner theory (Bohus and Rudnicky, 2003; Rich and Sidner, 1998).

We exemplify our automatic annotation of the discourse structure hierarchy in the ITSPOKE system. This approach takes advantage of the fact that the tutored information was structured in the spirit of the Grosz & Sidner theory. A dialogue with ITSPOKE follows a question-answer format (i.e. system initiative): ITSPOKE asks a question, the student provides the 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 that resembles the discourse segment hierarchy (see Figure 12). Tutor questions are grouped in segments which correspond roughly to the discourse segments. Similarly to the discourse

---

[12] Publications: The work presented in this section was first published in (Rotaru and Litman, 2006). The discourse structure transition information has since been used in other studies (Ai et al., 2006; Forbes-Riley et al., 2007; Forbes-Riley et al., 2008; Forbes-Riley et al., 2007; Rotaru and Litman, 2006).

segment purpose, each question segment has an associated tutoring goal or purpose. For example, in ITSPOKE there are question segments discussing about forces acting on the objects, others discussing about objects' acceleration, etc.

In Figure 12 we illustrate ITSPOKE's behavior and our hierarchy annotation. Note that the conversation is for a different problem and tutoring plan than the ones used in Figure 1/Figure 11. First, based on the analysis of the student essay, ITSPOKE selects a question segment to correct misconceptions or to elicit more complete explanations. This question segment will correspond to the top level discourse segment (e.g. DS1 in Figure 12). Next, ITSPOKE asks the student each question in DS1. If the student answer is correct, the system moves on to the next question (e.g. $Tutor_1 \rightarrow Tutor_2$). If the student answer is incorrect, there are two alternatives. For simple questions, the system will simply give out the correct answer and move on to the next question (e.g. $Tutor_3 \rightarrow Tutor_4$). For complex questions (e.g. applying physics laws), ITSPOKE will engage in a *remediation subdialogue* that attempts to remediate the student's lack of knowledge or skills. The remediation subdialogue is specified in another question segment and corresponds to a new discourse segment (e.g. DS2 in Figure 12). The new discourse segment is dominated by the current discourse segment (e.g. DS2 dominated by DS1). $Tutor_2$ system turn is a typical example; if the student answers it incorrectly, ITSPOKE will enter discourse segment DS2 and go through its questions ($Tutor_3$ and $Tutor_4$). Once all the questions in DS2 have been answered, a heuristic determines whether ITSPOKE should ask the original question again ($Tutor_2$) or simply move on to the next question ($Tutor_5$).

ESSAY SUBMISSION & ANALYSIS

DS 1

TUTOR$_1$: Consider Newton's laws applied to two objects that move together. What three quantities does Newton's Second Law describe the relationship between?

Student answer$_1$: correct (e.g. force, mass, accel.)

TUTOR$_2$: If two bodies are connected so that they move together and you know the acceleration of the first body, what is the acceleration of the second body?

Student answer$_2$: incorrect (e.g. zero)

DS 2

TUTOR$_3$: If the two bodies always move together and one body speeds up, what happens to the other?

Student answer$_3$: incorrect (e.g. lags behind)

TUTOR$_4$: The second body will speed up too. If the first body accelerates at a particular rate, will the second body accelerate at an equal or different rate?

Student answer$_4$: correct (e.g. equal)

TUTOR$_5$: If a force acts on one body such that it moves, what happens to the second body?

Student answer$_5$: incorrect but rejected (e.g. stays)

TUTOR$_6$: Could you please repeat that?

…

Figure 12. Automatic discourse segment hierarchy annotation in ITSPOKE

Note that this annotation of the discourse structure is an approximation. Indeed, if we apply this procedure for the automatic annotation of discourse segment hierarchy to the conversation from Figure 11, the two system questions presented in the excerpt form a single discourse segment (i.e. they are part of the same question segment). However, Figure 11 shows that a manual annotation of discourse structure yields multiple discourse segments. Nonetheless, this automatic approximation provides the backbone for our manual annotation of the discourse structure we describe in Section 3.4. A similar discourse structure approximation was used in (Levow, 2004). Their system performs multiple tasks (e.g. e-mail, calendar) and in their annotation, the dialogue segment for each task defines a discourse segment.

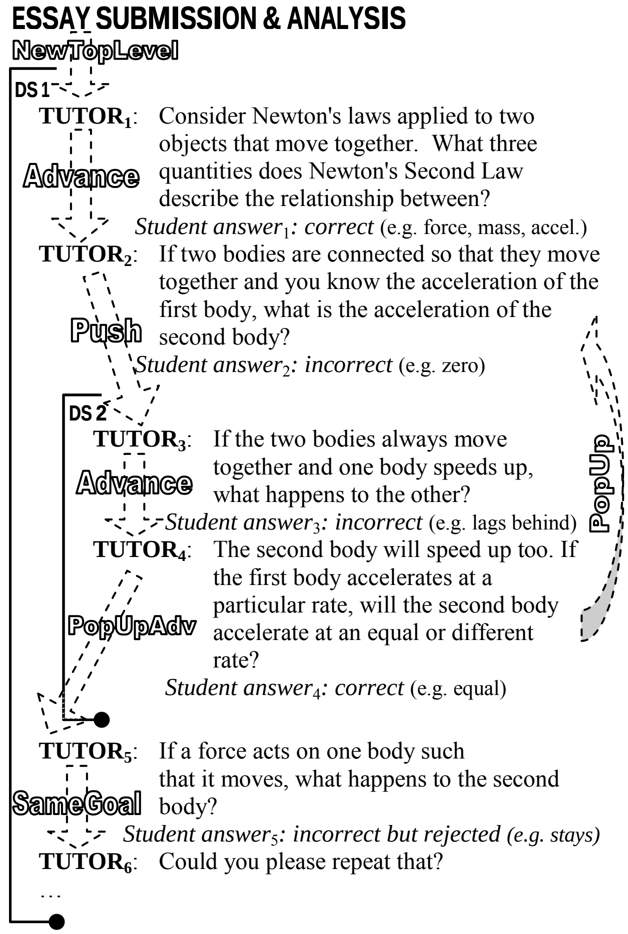


Figure 13. Transition annotation

(Each transition labels the turn at the end of the arrow)

Discourse structure transitions are defined for each system turn and capture the position in the discourse segment hierarchy of the current system turn relative to the previous system turn. We define six labels. In Figure 13 we show the transition annotation of the dialogue excerpt from Figure 12. **NewTopLevel** label is used for the first question after an essay submission (e.g. $Tutor_1$). If the previous question is at the same level with the current question we label the current question as **Advance** (e.g. $Tutor_{2,4}$). The first question in a remediation subdialogue is labeled as **Push** (e.g. $Tutor_3$). After a remediation subdialogue is completed, ITSPOKE will pop up and it will either ask the original question again or move on to the next question. In the first case, we label the system turn as **PopUp**. Please note that $Tutor_2$ will not be labeled with PopUp because, in such cases, an extra system turn will be created between $Tutor_4$ and $Tutor_5$ with the same content as $Tutor_2$. This extra turn includes variations of “Ok, back to the original question” to mark the discourse segment boundary transition. If the system moves on to the next question after finishing the remediation subdialogue, we label the system turn as **PopUpAdv** (e.g. $Tutor_5$). In case of rejections (2.3.1.5), the system question is repeated using variations of "Could you please repeat that?". Also, if the user does not answer the system question, ITSPOKE repeats the question again and listens for an answer from the user. We label such cases as **SameGoal** (e.g. $Tutor_6$).

We define the notion of "position in the dialogue" based on the transition label for the current system turn (e.g. current position in the dialogue is a PopUp transition). Indeed, the discourse structure transition information looks at the *horizontal relative position* of each system turn in the discourse segment hierarchy. The Push, PopUp and PopUpAdv transitions capture cases where we cross the discourse segment boundaries (going down one level for Push or going up one level for PopUp and PopUpAdv). The Advance transition captures cases in which we remain in the same discourse segment. The SameGoal transition represents cases where no extra information is produced due to a speech recognition rejection or due to user not answering.

Note that the discourse structure transition information satisfies the two properties we were looking for: automatically computed and domain-independence. Transitions are automatically computed as the discourse segment hierarchy is automatically extracted and the computation of the transition information from this hierarchy is also automatic. Discourse structure transitions are also domain independent: this information is directly computed from the discourse structure hierarchy and does not depend on the underlying domain.

There are other ways of exploiting the discourse segment hierarchy to define the notion of "position in the dialogue". In (Rotaru and Litman, 2006), we look at the depth information: the vertical position of the current system turn in the discourse segment hierarchy. Depth information also satisfies the two properties. While discourse structure depth was informative for the characterization of speech recognition problems (see (Rotaru and Litman, 2006)), it was not particularly useful for the performance analysis application (4.4). Consequently, we will not use discourse structure depth in this work.

Discourse structure has been implicitly used to define the notion of "position in the dialogue" in previous work. Details will be given in the "Related Work" sections of our system-side applications (4.6 and 5.6). Some of the definitions include: dialogue acts (Batliner et al., 2003; Walker et al., 2001), dialogue sequencing information (Gabsdil and Lemon, 2004), current system state (Bohus, 2007), type of subdialogue (conversation-related versus task-related or specific subtasks as in (Walker et al., 2001)). However all these definitions ignore or flatten the discourse segment hierarchy and/or are domain-dependent.

3.4 MANUAL ANNOTATION & THE NAVIGATION MAP

As mentioned in Section 3.2, we hypothesized that users will benefit from direct access to the discourse segment hierarchy and the discourse segment purpose information. We enable this access through our graphical representation of discourse structure which we call 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. 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.

The NM requires access to 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. After our first experiment with the NM (NMPrelim - 2.2.3), we developed an annotation manual which was used to annotate all 5 ITSPOKE problems. The annotation manual is available in Appendix C.

We describe our annotation and the NM design choices we made. Figure 14 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). ITSPOKE discusses the same problem/tutoring plan from Figure 1 and it the same dialogue excerpt used in Figure 11. Note that Figure 14 is not a screenshot of the actual system interface; the NM is the only part from the actual system interface (see Figure 21 for an actual screenshot). Figure 15 shows the NM after turn $Tutor_1$.

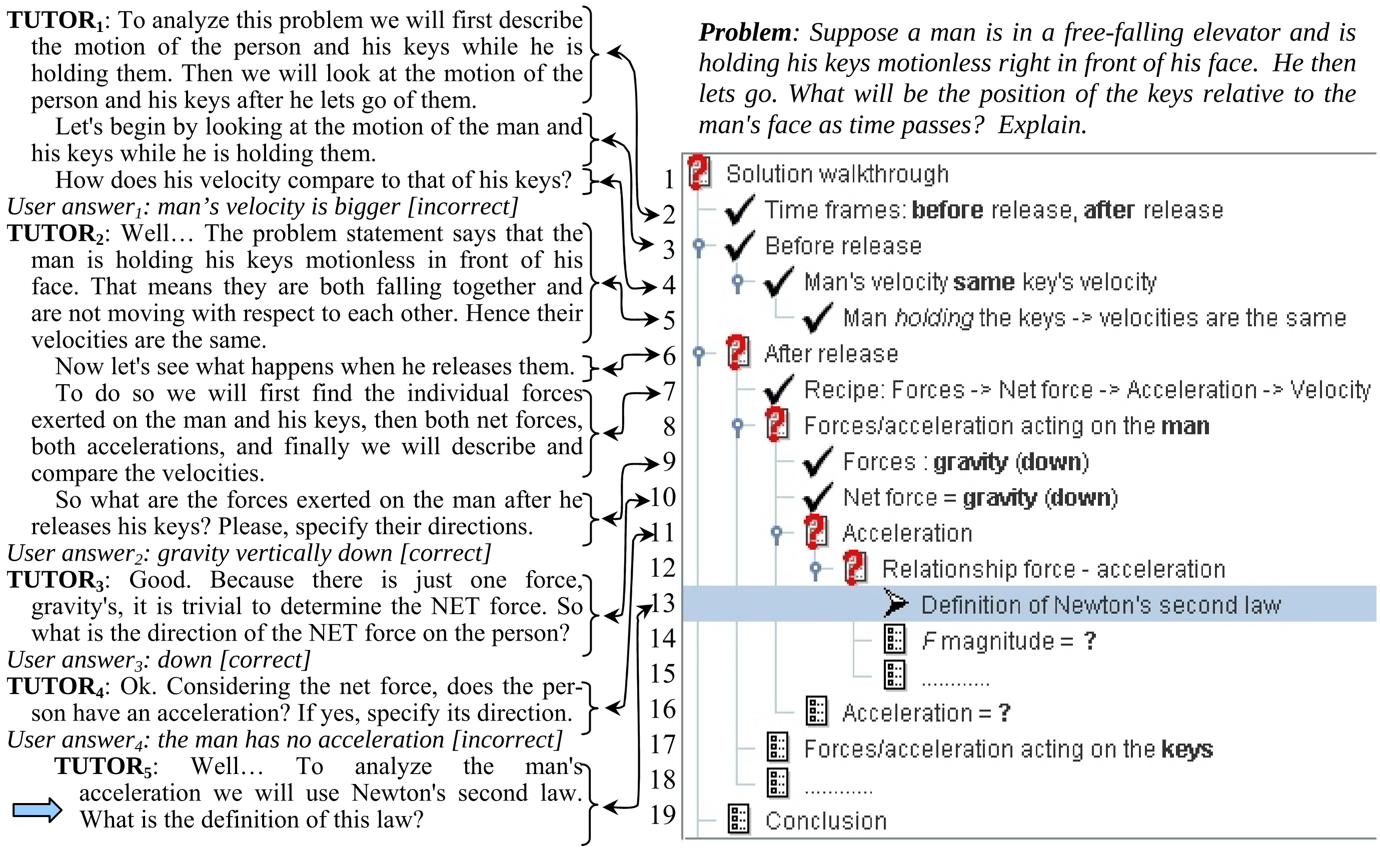


Figure 14. Transcript – NM correspondence

(Transcript of a sample ITSPOKE conversation (left). The NM as the user sees it after turn Tutor5 (right))

We manually annotated each system question/explanation for its intention(s)/purpose(s) (FOCUS intention annotation in Appendix C.4). Please note that some system turns have multiple intentions/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 14, NM$_2$). Next, the system indicates that it will discuss about the first time frame (Figure 14, NM$_3$) and then it asks the actual question (Figure 14, NM$_4$).

In addition to our manual annotation of the discourse segment purpose, we manually organized all discourse segments from a question segment in a hierarchical structure that reflects the discourse structure. We opted for a manual annotation of the discourse segment hierarchy instead of the automatic one described in Section 3.3 because we hypothesized that the latter is too coarse. For example, for the tutoring plan from Figure 1, all the leaf nodes in that plan will be in the same discourse segment according to the automatic annotation. Nonetheless, the automatic annotation is used as a skeleton for the manual annotation.

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[13] 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.

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

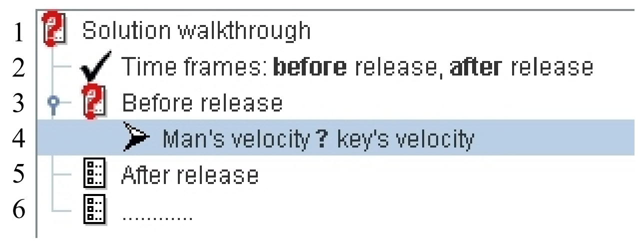


Figure 15. NM state after turn $Tutor_1$

Correct answers. In Figure 15 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 14). Thus, the NM was enhanced to display the correct answer after the system is done with each question (i.e. the user answer is correct or the incorrect answer was remediated). 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 (EXIT intention annotation in Appendix C.4).

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 the instruction is heading, thus facilitating their

[13] Since a system prompt can contain more than one discourse segment, we had to manually identify the discourse segment boundaries in the prerecorded system prompts. This allowed us to synchronize the system speech with the NM: whenever the system starts speaking about a new discourse segment, the NM displays and highlights it. In our first NM experiment (NMPrelim), the discourse segment boundaries in the prerecorded system prompts were approximated based on the number of words in the discourse segment.

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 14, NM_{14}, NM_{16}, NM_{17} and NM_{19}; Figure 15, NM_5); additional discourse segments at the same level are signaled through a dotted line. Since in some cases the next discourse segment can hint/describe the solution to the current question, each discourse segment has an additional purpose annotation that is displayed when the segment is part of the visible horizon (LOOKAHEAD intention annotation in Appendix C.4).

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 14, NM Line 3 is collapsed in the actual system and Lines 4 and 5 are hidden (shown in Figure 14 to illustrate our discourse structure annotation.). The user can expand nodes as desired using the mouse.

Information highlight. Bold and italics font are used to highlight important information (what and when to highlight was manually annotated). For example, in Figure 14, NM_2 highlights the two time frames as they are key steps in approaching this problem. Correct answers are also highlighted.

Discussion

It is important to understand if the NM design choices we made add information that is not part of the discourse structure information. The correct answer feature can be argued to be related to the discourse structure information. This is because the system intention behind each question is not only to get an answer from the user but also to *ground* the correct answer with the user (i.e. either through positive feedback for correct user answer (e.g. $Tutor_2$ in Figure 11) or additional explanations for incorrect answers (e.g. $Tutor_2$ in Figure 14)).

However, it is arguable whether the visual information included in the information highlight feature relates back to the discourse structure information. It is not very clear whether the information we highlight (e.g. important concepts, laws and correct answers) is part of discourse structure. One way to positively argue for information highlight is to bring into the picture the attentional state component of discourse structure: a dynamic model of objects, properties and relationships that are salient at each point in the discourse (recall 3.1). Many of the concepts, laws and correct answers we highlighted can be considered to be part of the attentional state. However, at the time of the NM annotation, we were not aware of this possible connection with the attentional state. Thus, deciding which piece of information to highlight was done in an ad-hoc fashion: whatever information the annotator believed it would be important to draw user's attention to. Since in our first NM experiment (NMPrelim – subjective utility) we were still exploring the utility of the NM, the information highlight feature was enabled in that study. However, given these issues with our annotation for information highlight and the fact that research done in parallel with this work shows that concept highlighting has a significant impact on performance

(Jackson and Graesser, 2007), in our second NM experiment (Main – objective utility) the information highlight feature was disabled. Even with information highlight disabled, the NM has a positive impact on objective and subjective metrics.

Although the auto-collapse feature was motivated by an efficient use of the available graphical real-estate, the feature can be related back to the attentional state. This is because the attentional state tends to function as a stack and older objects and properties are removed from the attentional state once the corresponding discourse segment has been completed. The auto-collapse feature does the same thing: it hides information that was already discussed and that is not in focus anymore.

The limited horizon feature presents (part of) the intention of future discourse segments which will become part of the discourse structure information.

3.5 ISSUES RELATED TO OUR USE OF DISCOURSE STRUCTURE

In this section we discuss some of the issues that arise from our use of discourse structure. First, it is important to highlight again that we make use of discourse structure in two ways: through discourse structure transitions (3.3) and through our graphical representation, the Navigation Map (3.4). Transitions exploit only the discourse segment hierarchy which is approximated using the procedure described in Section 3.3. The Navigation Map makes direct use of the discourse segment hierarchy and the discourse segment intentions, both manually annotated (the manual annotation of the hierarchy uses the automatic one as a skeleton). We will discuss some of the issues that arise from each use of discourse structure and conclude with some general remarks related to using discourse structure information in other spoken dialogue systems and/or domains.

Issues related to discourse structure transitions

While our transitions are domain-independent, note that the interpretation of each discourse structure transition is *dependent* on the discourse segment hierarchy annotation and the system. In our case, a Push transition signals entering a remediation dialogue. A PopUp or a PopUpAdv transition signals exiting a remediation dialogue, and asking the original question or continuing with the next question respectively. An Advance transition signals moving on to the next question in the question segment (even through an incorrect user answer for questions that do not require a remediation subdialogue). A NewTopLevel transition signals the beginning of a new dialogue. For other annotations of discourse structure hierarchy the interpretation might be different. Take for example our manual annotation from Section 3.4. As we showed in Figure 14 (a manual annotation of the example from

Figure 11) the first tutor turn has 3 discourse segments, and within this turn we have a NewTopLevel (NM_2 in Figure 14), an Advance (NM_3) and a Push transition (NM_4). Thus, a Push transition signals more than entering a remediation subdialogue in this annotation. In addition, a system turn can now have more than one transition associated with it. It is not obvious which one of these transition to use (e.g. only use the last one) or how to combine them (e.g. use all as a label: NewTopLevel–Advance–Push) for the two applications we will explore in Sections 4 and 5. Studying issues related to granularity and interpretation for other annotations and systems/domains is an interesting direction for the future.

We would like to highlight the connection between user correctness and the discourse structure transitions as we defined them in our system. However, the two concepts should not be confused. This connection exists in the first place because user correctness determines the course of the conversation in our system and, as a result, the shape of the discourse structure. In other systems, other factors might be of importance. For example, in simple information-access dialogue systems, speech recognition problems are of interest as they can reshape the conversation through confirmation or repair subdialogues. Second, user correctness is not equivalent with transitions. For example, an incorrect answer can be followed by a Push transition (i.e. for questions that are remediated through a remediation dialogue – e.g. $Tutor_2 \rightarrow Tutor_3$ in Figure 13), by an Advance transition (i.e. for questions that do not require remediation subdialogues – e.g. $Tutor_3 \rightarrow Tutor_4$ in Figure 13) or even by a PopUp or PopUpAdv transition (i.e. when the question does not require a remediation subdialogue and is the last question in a remediation subdialogue – e.g. $Tutor_4$ in Figure 13 if the Student $answer_4$ would have been incorrect). The SameGoal transition is unrelated to user correctness.

Similarly, remediation subdialogues and discourse structure transitions are tightly connected due to our definition of the discourse segment hierarchy. However, in other definitions of discourse segment hierarchy, transitions will transcend remediation subdialogues even for our system. For example, as discussed above, if we use the manual annotation from Section 3.4, we can have Push transitions inside a tutor turn (e.g. the first tutor turn in Figure 14 has 3 transitions inside: a NewTopLevel, an Advance and a Push). The interpretation and usage of discourse transitions for this annotation are left for future work.

Note that our second use of discourse structure, the Navigation Map, does not exhibit this confound with correctness and remediation dialogues.

Issues related to the Navigation Map (NM)

Since we are the first to look at the benefits of a graphical representation of discourse structure (see 6.5), we opted for a manual annotation to avoid dealing with the potential issues of an automatic annotation. For example, we hypothesized that our automatic annotation of the discourse segment hierarchy (3.3) will be too coarse for our purposes. Indeed, for the tutoring plan from Figure 1, all topics (i.e. leaf nodes) will form a single discourse segment according to our automatic annotation. However, we believed that it would be more helpful for users to have a richer hierarchy like the one from Figure 1. Nonetheless, we use the automatic annotation as a skeleton for the manual annotation. In addition, we did not have an automatic way of annotating the discourse segment purpose information, though previous work on summarization can be used as a starting point (e.g. (Mani, 2001)). Note that the NM is constructed automatically at runtime whether we use a manual or automatic annotation of the necessary information.

Because our focus is on applying the discourse structure, the reliability of our manual annotation is of secondary importance for our exploratory work. Nonetheless, we believe that our manual annotation is relatively robust as our system follows a carefully designed tutoring plan and previous studies have shown that naïve users can reliably segment discourse (e.g. (Passonneau and Litman, 1997)). Moreover, because ITSPOKE uses system initiative (i.e. the system drives the conversation by asking questions), the discourse structure annotation is simplified as we do not need to recognize the user plan as many user-initiative systems have to do (e.g. (Allen et al., 2000; Blaylock and Allen, 2006)). Our positive results (6.3 and 6.4) motivate additional studies to measure the reliability of the discourse structure annotation and to investigate the right choice of granularity. As part of our work, we developed an annotation manual (Appendix C), however it was only used by a single annotator (the author). This manual can be used a starting point to investigate reliability and granularity issues for the discourse structure annotation and for implementing the NM in a new system/domain.

Using discourse structure in dialogue systems

Two important properties of ITSPOKE facilitate our use of discourse structure: system initiative and hierarchical task representation. ITSPOKE uses system initiative: the system drives the conversation by asking questions and users respond. Thus, we always know which question will be selected next and how it relates to the current question: if the user is incorrect and the current question requires a remediation subdialogue, then the next question is the first question in the remediation subdialogue; otherwise we simply move on to the next question in the tutoring plan. This type of initiative allowed us to implement the "Limited horizon" feature in the Navigation Map. As discussed in Section 3.3, the task is represented in ITSPOKE using a hierarchical structure that resembles the discourse segment hierarchy (see Figure 12). Tutor questions are grouped in segments which correspond roughly to the discourse segments. Similarly to the discourse segment purpose, each question segment has an associated tutoring goal or purpose. For example, in ITSPOKE there are question segments discussing about forces acting on the objects, others discussing about objects' acceleration, etc. This hierarchical representation has eased our automatic annotation of discourse structure hierarchy and has served as a skeleton for our manual annotation.

Although many systems that use dialogue managers inspired by the Grosz & Sidner theory of discourse (e.g. RavenClaw (Bohus and Rudnicky, 2003), COLLAGEN (Rich et al., 2001)) behave similarly to ITSPOKE, using discourse structure in other systems raises additional complications. In particular, in systems with user initiative, the user holds the conversation plan and it is the system job to (partially) recognize or infer user's plan (e.g. (Allen et al., 2000; Blaylock and Allen, 2006)). In the author's unpublished work with Shimei Pan, we looked at such a system: a multimodal system that allows users to browse a large real-estate database (Zhou et al., 2005). Users can browse various entities like houses, cities, school districts, hotels, etc. Our choice of discourse segments for this system was user queries. To build the NM, we needed the discourse segment hierarchy: i.e. how the queries/discourse segments relate to other queries/discourse segments. To construct this hierarchy we had to infer at runtime the user's exploration plan which is a non-trivial task even for that domain.

The internal task representation can also complicate the use of discourse structure. For example, it is not obvious how to infer the discourse structure hierarchy for dialogue systems that use probabilistic dialogue managers (e.g. (Paek and Horvitz, 2000)) or certain plan-based dialogue managers (e.g. the information state approach to dialogue management (Larsson and Traum, 2000)). The main problem for these systems is that there is no tree-based representation that can implicitly guide an annotation of discourse segment hierarchy. In addition, these systems typically have a large number of system states (e.g. the information state representation as in (Larsson and Traum, 2000)) and figuring out all possible transitions between these states might be complicated (e.g. the process is governed by dialogue moves, update rules and update strategy in (Larsson and Traum, 2000)).

These task representation issues stem from some of the problems with the Grosz & Sidner theory of discourse. In particular, previous work has challenged the tree-like structure of the intentional structure assumed in the theory (recall 3.1). For example, (Rosé et al., 1995) argue that this assumption does not hold for dialogues with multiple concurrent threads (e.g. negotiation dialogues where multiple alternatives are kept active simultaneously and discussed in an interleaving fashion). They extend the theory by allowing multiple tree-like structures to form concurrently, one for each thread. In fact, even in the unpublished work mentioned above, we had to create a NM for each type of entity to address similar issues. Understanding the implications of these domains on our discourse structure transitions and the NM is an interesting direction for future work.

3.6 SUMMARY AND FUTURE WORK

In this work, we use the Grosz & Sidner theory of discourse structure (Grosz and Sidner, 1986). We had two intuitions behind our use of the discourse structure information. First, we believed that the "position in the dialogue" is an important information source and we show how we can define this concept using the notion of discourse structure transitions. These transitions capture events that happen in the discourse segment hierarchy: crossing of discourse segment boundaries (Push, PopUp, PopUpAdv), remaining in the same discourse segment (Advance) or lack of new information (SameGoal). We offer an automatic way of annotating the discourse segment hierarchy and consequently of automatically annotating discourse structure transitions.

Other researchers have acknowledged the importance of phenomena similar to our transitions. For example, (Hirschberg and Nakatani, 1996) looks at initial, medial and final utterances in a discourse segment which related to our Push, Advance and PopUp/PopUpAdv transitions. We propose a nomenclature for these phenomena via our discourse structure transitions and we contribute by using transitions to define the concept of "position in the dialogue". More details on how the "position in the dialogue" was defined in previous work are discussed in the Related Work sections of our system-side applications (4.6 and 5.6). While our definition of discourse structure transitions is domain independent and automatically computable (at least in dialogue managers inspired by the Grosz & Sidner theory (Bohus and Rudnicky, 2003; Rich and Sidner, 1998)), it would be interesting to apply it to other annotations of the discourse segment hierarchy and/or other systems/domains and to study issues related to the interpretation of discourse structure transitions. For example, one can investigate issues that arise from using a manual annotation of the discourse segment hierarchy or if the 6 transitions we defined are enough (e.g. maybe in some domains we should distinguish between two forms of Advance: Advance with a correct answer or Advance with an incorrect answer).

Our second intuition was that users will benefit from direct access to discourse structure. We will show in Section 6 that our graphical representation of discourse structure based on a manual annotation (the Navigation Map) leads to a number of subjective and objective improvements. These positive results motivate additional studies to measure the reliability of the discourse structure annotation and to investigate the right choice of granularity and its impact on the effectiveness of the Navigation Map.

We also highlight some of the issues that might arise from similar use of discourse structure in other systems/domains (Section 3.5): the inherent connection between discourse structure transitions and user correctness/remediation subdialogues in our system, issues related to the interpretation of discourse structure transitions, reliability and granularity issues for the NM, and the effects that the initiative type and the task representation have on obtaining the discourse structure information.

4 APPLICATIONS FOR PERFORMANCE ANALYSIS

4.1 INTRODUCTION

The success of a SDS depends on a large number of factors and how the system addresses/handles them. Some of these factors are known beforehand while others are unknown and can be identified only later. A number of these factors are very intuitive. For example, problems with automated speech recognition can derail a dialogue from the normal course: e.g. non-understandings, misunderstandings, end-pointing, etc. (e.g. (Bohus, 2007; Raux and Eskenazi, 2008; Swerts et al., 2000)). The strategy the system employs to handle or avoid these situations is also important and researchers have experimented with many such strategies as there is no clear winner in all contexts (e.g. (Bohus, 2007; Singh et al., 2002)).

A principled approach to identifying important factors and ways to handle them comes from performance analysis. In performance analysis, the behavior of SDS is analyzed from the perspective of a performance metric. First, the SDS behavior is quantified in form of *interaction parameters*. These parameters measure various events that occur in the dialogues between users and the system (e.g. speech recognition performance, number of turns, number of help requests, etc.). Next, two properties of interaction parameters are investigated: **predictiveness** and **informativeness**. Note that the two properties and their names are not clearly established in previous SDS research although the investigations behind these properties are a common practice in the field.

Predictiveness of an interaction parameter looks at the connection between the parameter and system performance. Typically, predictive models are built to associate the interaction parameters with the performance metric. These models can tell us which factors are linked to performance and the polarity of their impact (i.e. positive vs. negative). In addition, these models allow researchers to assess the performance of future system improvements without running additional costly user experiments. In the intelligent tutoring system research, univariate linear regression models are typically used (i.e. correlations with learning - 2.4.2). The PARADISE framework developed for SDS by (Walker et al., 1997) uses multivariate linear regression which allow for multiple interaction parameters to be present in the model at the same time.

Once the predictiveness of an interaction parameter is established, it is important to look at its informativeness. Informally, informativeness looks at how much the parameter can help us improve the system. We already know that the parameter is predictive of performance. But this does not tell us if there is a causal link between the two. In fact, the main drive is not to prove a causal link but to show that the interaction parameter will inform a modification of the system and that this modification will improve the system.

Previous SDS research (e.g. (Walker et al., 2000)) has looked at the predictiveness of interaction parameters that measure the dialogue efficiency (e.g. number of system/user turns, task duration, etc.) and the dialogue quality (e.g. recognition accuracy, number of rejections, number of help requests, etc.). An extensive set of parameters can be found in (Möller, 2005). However, most of these parameters do not take into account the discourse structure information or make a limited domain-dependent use of it (e.g. the DATE annotation scheme discussed in Section 4.6).

Here we investigate the predictiveness of new types of interaction parameters that use discourse structure information. Unlike most of the previous work, we take a step further and also look at the informativeness of these parameters.

4.2 PROBLEM STATEMENT

Statement: We investigate the utility of using the discourse structure information for performance analysis by looking at the predictiveness and informativeness of interaction parameters that use this information.

Hypothesis: Interaction parameters that use the discourse structure information are useful for performance analysis: they are predictive of system performance (predictiveness) and inform modifications of the system that result in system improvements (informativeness).

Intuition: Dialogue phenomena related to performance (e.g. correctness, user affect) are not uniformly important across the dialogue but have more weight based on their "*position in the dialogue*" (Conditioning intuition). "Good" and "bad" dialogues have *different structures* (Discrimination intuition). Discourse structure transitions can be used to define "position in the dialogue" and "different structures".

Approach: We compute interaction parameters that make use of the discourse structure transition information (4.3). First, we look at the predictiveness of these parameters by investigating correlations with system performance (4.4.1). We pick the two most promising parameters in terms of their ability to inform a valuable modification of the system and we test if their predictiveness generalizes to other corpora to further validate our choice (4.4.2). Next, we implement the modification suggested by one of these parameters (the PopUp–Incorrect bigram) and investigate if it leads to performance improvements to prove its informativeness (4.5).

Results: We find that the interaction parameters that use discourse structure (i.e. transition–correctness, transition–affect and transition–transition bigrams) are predictive of system performance. This is in contrast with the corresponding parameters that do not use discourse structure information (i.e. correctness and affect) which have little predictive power. Among discourse structure-based parameters, the PopUp–Incorrect transition–correctness bigram has an intuitive explanation and informs a sensible modification of our system. Its predictive power generalizes to other corpora and the system modification results in objective improvements (e.g. performance improvements for certain users but not at the population level and fewer system turns).

A summary of our analyses, results and other work items is available below in Table 7. For each item we show the section where more details are available, a brief description, the statistical confidence in the outcome and whether it is a positive or negative outcome.

Table 7. Summary of results: performance analysis

Result	Confidence	+/-
(4.4) Predictiveness		
(4.4.1) Several discourse-structure based parameters are predictive of system performance (learning)		
• transition-correctness bigrams	**significant**	+
• transition-certainty bigrams	**significant**	+
• transition-transition bigrams	**significant**	+
(4.4.2) PopUp–Incorrect bigram predictiveness generalize to other corpora	**trend/significant**	+
(4.5) Informativeness		
(4.5.2) Implemented a new PopUp–Incorrect strategy		+
(4.5.3) Run a between-subjects user study		+
(4.5.4) Comparison with control condition (*PI* vs *R*)		
• (4.5.4.2) System performance: very small improvement (0.07 NLG effect size)	non-significant	-
• (4.5.4.3) Aptitude-treatment interaction: pretest split (PRE Split)		
• PRE Split x Condition: no interaction with learning	non-significant	
• + 0.15 NLG effect size for low pretesters	non-significant	
• + 0.00 NLG effect size for high pretesters	non-significant	
• (4.5.4.4) PopUp–Incorrect predictiveness: significantly correlated with system performance for *R* users, but not for *PI* users	**significant**	+
• (4.5.4.4) Splitting based on the number of PopUp–Incorrect events (PI Split)		
• PI Split x Condition: trend interaction with learning	*trend*	
RELNLG subset (2.3.3.2)		
• PI Split x Condition: significant interaction with learning	**significant**	
• + 0.96 NLG effect for high PopUp–Incorrect users	**significant**	+
• - 0.50 NLG effect for low PopUp–Incorrect users	non-significant	+/-
• *PI* users: low and high PopUp–Incorrect users have similar learning		+
• (4.5.4.5) Subjective metrics: no differences	non-significant	+
• (4.5.4.6) Dialogue time: very small reduction in dialogue time	non-significant	+
• (4.5.4.7) Number of system turns: reduction in number of system turns	*trend*	+

4.3 DISCOURSE STRUCTURE-BASED INTERACTION PARAMETERS

We use the discourse structure transition information to derive interaction parameters that look at two classes of events: *transition–phenomena* events and *transition–transition* events. For each class of events we had an intuition related to their predictiveness which motivated our decision to look at the event class. We discuss below these intuitions and the interaction parameters we compute.

First, we use discourse structure transitions as contextual information for dialogue phenomena related to performance. This is based on our intuition that dialogue phenomena related to performance are not uniformly important but have more weight based on their position in the dialogue. We refer to this as the **Conditioning intuition**. For example, it was our intuition that although correctness is related to system performance (i.e. user learning), it is more important for users to be correct at specific places in the dialogue rather than overall in the dialogue. More specifically, if we look for example at how many time the user was incorrect overall and how many times the user was incorrect after a PopUp transition, our intuition was that the latter is more predictive of learning than the former. We will look at two phenomena related to performance in our system/domain: user correctness (2.3.1.4) and user certainty (2.3.2.2). In tutoring, user correctness measures directly their knowledge of the topic discussed in the question. User affect (e.g. certainty) and especially adaptation to user affect has been recently pursued as a modality to improve existing SDS and intelligent tutoring systems (Ang et al., 2002; Forbes-Riley et al., 2008; Pon-Barry et al., 2006).

Second, we look at trajectories in the discourse structure hierarchy. This is based on our intuition that the structure of a dialogue can tell us something about how much a user learns (i.e. system performance). In other words, "good" and "bad" dialogues have different discourse structures where "good" dialogues are from users that learn more and "bad" dialogues are from users that learn less. We refer to this as the **Discrimination intuition**. To compare two dialogues in terms of the discourse structure we look at what type of transition trajectories are present in each. Due to data sparsity, we only look at trajectories at length two (i.e. two consecutive transitions). Thus our intuition can be translated, for example, as follows: we find more two consecutive Push transitions (i.e. a Push–Push trajectory) in "good" dialogues compared to "bad" dialogues. In other words, interaction parameters that measure Push–Push trajectories should be predictive of performance.

We will refer to the two classes of events as **bigrams** because each looks at two consecutive observations: the transition of the system turn and the phenomena (e.g. correctness or certainty) present in the subsequent user turn for transition–phenomena events and the transition of two consecutive system turns for transition–transition events. Thus, in our analysis we will look at 3 types of bigrams: *transition–correctness, transition–certainty and transition–transition bigrams*. There are 6 possible transition values

(3.3), 4 possible correctness values (2.3.1.4) and 4 possible certainty values (2.3.2.2). As a result there are 24 transition–correctness bigrams, 24 transition–certainty bigrams and 36 transition–transition bigrams. Going back to Figure 13, the three incorrect answers will be distributed to three bigrams: Advance–Incorrect ($Tutor_2$–$Student_2$), Push–Incorrect ($Tutor_3$–$Student_3$) and PopUpAdv–Incorrect ($Tutor_5$–$Student_5$). In terms of transition–transition bigrams, the $Tutor_4$–$Tutor_5$ pair will be counted as an Advance–PopUpAdv bigram.

To derive interaction parameters, for each bigram we compute a *total* parameter (identified as bigram followed by "#"), a *percentage* parameter (identified as bigram followed by "%") and a *relative percentage* parameter (identified as bigram followed by "rel%"). The parameters are computed for each user using the dialogues from all 5 problems. The total parameter counts how many times we see the bigram for a user. The percentage parameter normalizes the bigram occurrence count to the dialogue length. The relative percentage parameter normalizes the bigram occurrence count to the number of time the transition occurs. For example, the PopUp–Incorrect rel % parameter represents the percentage of Incorrect user turns after a PopUp transition.

The Conditioning intuition we presented above suggests that our discourse structure based parameters (i.e. bigram parameters) should also have *relative* predictiveness: transition–phenomena events should be "more" predictive of performance compared with phenomena events. To test this, we also looked at two **unigram** events: correctness and certainty. For each unigram (e.g. Incorrect, Uncertain) we compute a total parameter (e.g. number of Incorrect answers) and percentage parameter (e.g. percentage of Uncertain answers).

4.4 PREDICTIVENESS

We first look at the predictiveness of the discourse structure-based interaction parameters and validate our intuitions. For each predictive parameter we discuss its potential informativeness. We select one of these parameters and we look if it generalizes to other corpora as a preliminary step towards understanding its informativeness.

4.4.1 Results

To test the absolute and relative predictiveness of the discourse structure-based parameters[14], we perform an empirical analysis on the F03 corpus (2.2.1). We use learning as our system performance metric. We focus primarily on correlations between our interaction parameters and learning as this modeling methodology is commonly used in the tutoring research (e.g. (Chi et al., 2001)). Because in our data the pretest score is significantly correlated with the posttest score (R=0.462, p<0.04), one way to look to look at correlations with learning is to study *partial* Pearson's correlations between our parameters and the posttest score that account for the pretest score (2.4.2).

Out of the three forms of correctness available in F03 (human, transcript and system correctness – see 2.3.1.4 and 2.3.2.1) we use human correctness. We opted for this correctness annotation to eliminate the noise introduced by the automatic speech recognition component and the semantic interpretation component.

We report only parameters with significant/trend correlations with learning[15]. For each correlation we report the unigram/bigram the parameter is derived from and the average and standard deviation for that bigram (i.e. the average/standard deviation of the total parameter). For brevity, for each unigram/bigram, we report only the best Pearson's Correlation Coefficient (R) associated with parameters derived from the unigram/bigram and the statistical significance of that coefficient R (p). As a result, in our discussions we will refer only to the unigram/bigram instead of specifying the actual unigram/bigram parameter(s)[16]. Correlation information for each parameter is available in Appendix B.1.1, Table 39.

Results: Unigrams

To measure the relative predictiveness of the discourse structure-based interaction parameters, we first look at the predictiveness of the unigram parameters. We find only one significant/trend correlation

---

[14] Publications: The work presented in this section was published in (Rotaru and Litman, 2006) and extended in (Forbes-Riley et al., 2008).

[15] Some unigrams/bigram events occur too little in the corpus (e.g. most of the transition–Partially Correct bigrams). As a result, the count for these unigram/bigrams is 0 for most students and can lead to spurious correlations. We observed that a threshold of 1.4 bigram occurrences on average eliminates these situations. All significant/trend correlations below this threshold are ignored.

[16] For example, for the PopUpAdv–Correct bigram reported in Table 9, we find that the count parameter (PopUpAdv–Correct #) and the percentage parameter (PopUpAdv–Correct %) are both correlated with learning (R=0.43, p<0.06 and R=0.52, p<0.02, respectively). We only report the best R (0.52 in this case) and its significance.

(Table 8): neutral turns (in terms of certainty) are negatively correlated with learning. We hypothesize that this correlation captures student involvement in the tutoring process: more involved students will try harder thus expressing more certainty or uncertainty. In contrast, less involved students will have fewer certain/uncertain/mixed turns and, in consequence, more neutral turns.

Surprisingly, none of the correctness unigrams is trend/significantly correlated with learning. The only unigram that comes close to a trend correlation is the Correct unigram ($R=0.37$, $p<0.12$).

Table 8. All trend and significant unigram correlations

Unigram	Avg (StdDev)	Best R	p
Neutral	43.6 (17.0)	**-0.47**	**0.04**

To summarize, we find that dialogue phenomena which we hypothesized to be linked with learning offer limited insights regarding learning: only one unigram is predictive of learning (at least for our system and in this corpus).

Results: Transition–Correctness bigrams

This type of bigram informs us whether accounting for the discourse structure transition when looking at student correctness has any predictive value (the Conditioning intuition from Section 4.3). We find several interesting trend/significant correlations (Table 9).

User behavior, in terms of correctness, after a PopUp or a PopUpAdv transition provides insights about user's learning patterns. In both situations, the student has just finished a remediation subdialogue and the system is popping up either by reasking the original question again (PopUp) or by moving on to the next question (PopUpAdv). We find that after PopUp, correct student answers are positively correlated with learning. In contrast, incorrect student answers are negatively correlated with learning. We hypothesize that this correlation indicates whether the student took advantage of the additional learning opportunities offered by the remediation subdialogue. By answering correctly the original system question (PopUp–Correct), the student demonstrates that he/she has absorbed the information from the remediation dialogue. This bigram is an indication of a successful learning event. In contrast, answering the original system question incorrectly (PopUp–Incorrect) is an indication of a missed learning opportunity; the more such events happen the less the student learns.

Table 9. All trend and significant transition–correctness bigram correlations

Bigram	Avg (StdDev)	Best R	p
PopUp–Correct	7.0 (3.3)	**0.45**	**0.05**
PopUp–Incorrect	2.0 (1.8)	**-0.46**	**0.05**
PopUpAdv–Correct	2.5 (2.0)	**0.52**	**0.02**
NewTopLevel–Incorrect	2.4 (1.8)	**0.56**	**0.01**
Advance–Correct	40.6 (9.8)	**0.45**	**0.05**

Similarly, being able to correctly answer the tutor question after popping up from a remediation subdialogue and moving on to the next question (PopUpAdv–Correct) is positively correlated with learning. Since in many cases, these system questions will make use of the knowledge taught in the remediation subdialogues, we hypothesize that this correlation also captures successful learning events.

Another set of interesting correlations is produced by the NewTopLevel–Incorrect bigram. We find that ITSPOKE's starting of a new essay revision dialogue that results in an incorrect student answer is positively correlated with learning. The content of the essay revision dialogue is determined based on the analysis of the student essay by the WHY2 backend. We hypothesize that an incorrect answer to the first tutor question is indicative of the system's picking of a topic that is problematic for the student. Thus, we see more learning in students for which more knowledge gaps are discovered and addressed by ITSPOKE.

Finally, we find that correct answers after an Advance transition are positively correlated with learning (Advance–Correct bigram). We hypothesize that this correlation captures the relationship between a student that advances without major problems and a higher learning gain.

To summarize, we find that transition–correctness bigrams produce many predictive interaction parameters that have intuitive interpretations and hypotheses behind them (e.g. successful/failed learning opportunities). Since the correctness unigrams do not produce predictive parameters, our results confirm the relative predictiveness of the transition–correctness bigrams and validate the Conditioning intuition.

Results: Transition–certainty bigrams

Next we look at the combination between the transition in the dialogue structure and student certainty (the Conditioning intuition from Section 4.3). Results are presented in Table 10. These correlations offer additional insight into the negative correlation between the Neutral unigram and student learning (recall Table 8). We find that out of all neutral student answers, those that follow Advance transitions are negatively correlated with learning. Similar to the Neutral unigram correlation, we hypothesize that the Advance–Neutral correlation captures student's lack of involvement in the tutoring process. This might be also due to ITSPOKE engaging in teaching concepts that the student is already familiar with.

Table 10. All trend and significant transition–certainty bigram correlations

Bigram	Avg (StdDev)	Best R	p
Advance–Neutral	23.7 (8.3)	**-0.73**	**0.00**
SameGoal–Neutral	3.1 (4.0)	**0.46**	**0.05**

In contrast, staying neutral in terms of certainty after a system rejection is positively correlated with learning (SameGoal–Neutral). These correlations show that based on their position in the discourse structure, neutral student answers will be correlated either negatively or positively with learning.

To summarize, we find that transition–certainty bigrams produces several predictive interaction parameters which offer additional insights into the predictive Neutral unigram. These results provide additional support for the Conditioning intuition.

Results: Transition–transition bigrams

Finally, we are looking at the transition–transition bigram correlations (Table 11). These bigrams help us find trajectories of length two in the discourse structure that are associated with student learning (the Discrimination intuition from Section 4.3).

The Advance–Advance bigram captures situations in which the student is covering tutoring material without major knowledge gaps. This is because an Advance transition happens when the student either answers correctly or his incorrect answer can be corrected without going into a remediation subdialogue. Similar to the Advance–Correct correlation (recall Table 9), we hypothesize that this correlation links higher learning gains to students that cover a lot of material without many knowledge gaps.

Table 11. All trend and significant transition–transition bigram correlations

Bigram	Avg (StdDev)	Best R	p
Advance–Advance	34.8 (9.1)	**0.47**	**0.04**
Push–Push	2.2 (1.7)	**0.52**	**0.02**
SameGoal–Push	1.4 (1.8)	**0.49**	**0.03**

The Push–Push bigrams capture another interesting behavior. In these cases, the student incorrectly answers a question, entering a remediation subdialogue; she also incorrectly answers the first question in the remediation dialogue entering an even deeper remediation subdialogue. We hypothesize that these situations are indicative of big student knowledge gaps. In our corpus, we find that the more such big knowledge gaps are discovered and addressed by the system the higher the learning gain. Please note that the Push–Push bigram is more specific than the Push–Incorrect bigram because the latter also includes cases where the incorrect student answer is corrected through an explanation (i.e. resulting in an Advance transition).

The SameGoal–Push bigram captures another type of behavior after system rejections that is positively correlated with learning (recall the SameGoal–Neutral bigram, Table 10). In our previous work (Rotaru and Litman, 2006), we performed an analysis of the rejected student turns and studied how rejections interact with student state. The results of our analysis suggested a new strategy for handling rejections in the tutoring domain: instead of rejecting student answers, a tutoring SDS should make use of the available information. Since the recognition hypothesis for a rejected student turn would be interpreted most likely as an incorrect answer thus activating a remediation subdialogue, the positive correlation between SameGoal–Push and learning suggests that the new strategy will not impact learning.

To summarize, we find that transition–transition bigrams produces many predictive interaction parameters that have intuitive interpretations and hypotheses behind them (e.g. discovery of deep user knowledge gaps). These results provide support for the Discrimination intuition.

Summary

Our results indicate that the discourse structure transition information produces many predictive parameters for performance analysis. We also provide support for the Conditioning and Discrimination intuitions. We find that while dialogue phenomena unigram parameters produce only one trend/significant correlation, transition–phenomena bigrams result in a large number of trend/significant correlations (recall Table 9 and Table 10). In addition, the transition–transition bigram parameters are also predictive (recall Table 11).

4.4.2 Generalization to other corpora

As a first step towards proving the informativeness of discourse structure-based interaction parameters, we select two predictive bigrams (PopUp–Incorrect and NewTopLevel–Incorrect) based on their potential informativeness and investigate if their predictiveness generalizes to other corpora. If their predictiveness is not an isolated finding specific to a population of users/version of the system, then this strengthens the connection between the parameter and performance. In turn, a strong connection suggests that experiments to prove informativeness are likelier to have positive results.

Potential for informativeness

The most promising bigram in terms of its ability to produce a valuable modification of the system is the PopUp–Incorrect bigram (recall Table 9). Our interpretation for this correlation coupled with the PopUp–Correct bigram correlation is that they capture failed and successful learning opportunities. We hypothesized that these correlations indicate whether the student took advantage of the additional learning opportunities offered by the remediation subdialogue whose completion is signaled by the PopUp transition. By answering correctly the original system question (PopUp–Correct), the user demonstrates that he/she has absorbed the information from the remediation dialogue. This bigram is an indication of a successful learning event. In contrast, answering the original system question incorrectly (PopUp–Incorrect) is an indication of a missed learning opportunity. Because successful learning opportunities are positively correlated with learning while failed learning opportunities are negatively correlated with learning, one way to modify the system is to reduce the number of failed learning opportunities by transforming them into successful learning opportunities. That is, whenever the system detects a failed learning opportunity (i.e. an incorrect answer after a PopUp transition), instead of giving away the correct answer as the system does now, we can modify the system to give additional instruction. For example, a more detailed explanation could make explicit the connection between the question and the points discussed in the remediation dialogue and how the latter combine to produce the correct answer. It should also be tailored to students that did not understand the original remediation dialogue (e.g. give more explanations and intuitions, spell out the connections between the issues discussed, etc).

Another interesting modification can be derived from the NewTopLevel–Incorrect bigram (recall Table 9). We hypothesized that an incorrect answer to the first tutor question after an essay revision is indicative of the system selecting a topic that is problematic for the student. Thus, we see more learning in students for which more knowledge gaps are discovered and addressed by ITSPOKE. Thus, one way of modifying the system is to change the system behavior after essay analysis. Instead of activating the tutoring topic based on the analysis of the student essay (Rosé et al., 2003), we can have the system try all possible tutoring topics one by one. A system prompt can introduce this revision process (e.g. "There are a few topics I would like to discuss with you before we are done with this problem"). If the student answers correctly the first question in the question segment associated with a tutoring topic, the system will abandon that topic and move to the next topic (e.g. system prompt "You seem to be comfortable with this topic so I will not continue. Let's move on to the next one"). An ordering of the tutoring topics is required for this modification.

We only selected these 2 bigrams because the other bigrams that were predictive of system performance (i.e. Advance–Correct, Advance–Neutral, SameGoal–Neutral, Advance–Advance, Push–Push and SameGoal–Push) did not inform obvious modifications of our system or these modifications raised additional challenges. For example, it is not very clear how to modify the system in light of the positive correlation between Push–Push and performance. One modification would be to redesign the instruction to feature many opportunities for Push–Push but it is not clear how to do this. As another example, for transition–uncertainty bigrams, we would need an online prediction of uncertainty which is beyond the scope of our work.

Generalization of predictiveness to other corpora

At the time of this investigation, one other ITSPOKE corpus was available: the S05 corpus (2.2.2). This corpus was an ideal dataset to test the generalization of the predictiveness: it uses similar versions of the system (the S05 ITSPOKE version has some minor bug fixes) and similar experiment procedure but the user populations are different (2 years difference) and two speech output versions are used (synthesized, S05SYN, versus prerecorded, S05PR, output).

We had to use a different correctness annotation since human correctness annotation (HSEM - 2.3.2.1) is not available in the S05 corpus. The closest one is the transcript correctness (TSEM – see 2.3.1.4). The only difference between HSEM and TSEM is who does the semantic interpretation: a human annotator for HSEM or a system component for TSEM. Thus, in order to be labeled correct, a user answer has to be correct but also be formulated so that the system can interpret it (e.g. use the right terminology and lexical choice).

Table 12 shows the correlations between the two bigrams and learning on all 3 corpora. As in previous correlation tables, we present the average and standard deviation, the best Pearson's Correlation Coefficient (R) and the corresponding significance. Correlation information for each parameter is available in Appendix B.1.1, Table 40. Since we use a different form of correctness we rerun the correlations on F03 with TSEM.

We observe that the predictiveness of the PopUp–Incorrect bigram generalizes to all corpora (at a trend level of significance). First, we see that even if we use TSEM correctness, PopUp–Incorrect remains negatively correlated with learning in F03 (trend). Although the average number of PopUp–Incorrect is somewhat larger in S05PR and S05SYN, the bigram remains negatively correlated with learning (trend for S05PR and significant for S05SYN).

Table 12. Predictiveness generalization to other corpora

	PopUp-Incorrect			NewTopLevel-Incorrect		
Corpus	Avg (StdDev)	Best R	p	Avg (StdDev)	Best R	p
F03	3.3 (2.0)	*-0.43*	*0.07*	2.5 (1.9)	**0.59**	**0.01**
S05PR	3.6 (2.5)	*-0.32*	*0.08*	3.8 (1.9)	-0.20	0.30
S05SYN	4.5 (2.5)	**-0.50**	**0.00**	4.0 (1.9)	-0.13	0.49

In contrast, NewTopLevel–Incorrect predictiveness does not generalize. Although we still find a strong correlation on F03 when using TSEM correctness, the correlations with learning are not significant in the other two corpora. In fact, even the signs of the correlations are different.

These findings suggest that our best option is to test the informativeness of the PopUp–Incorrect bigram. We will pursue this investigation in Section 4.5.

Summary

Two of the several predictive discourse structure bigrams have informativeness potential: PopUp–Incorrect and NewTopLevel–Incorrect. Only the PopUp–Incorrect bigram generalizes to other corpora and, consequently, we will investigate its informativeness.

4.5 INFORMATIVENESS

In Section 4.4, we saw that the PopUp–Incorrect bigram is one of the predictive bigrams with a great potential to be informative (4.4.2). We first discuss the relationship between PopUp–Incorrect events and system performance (4.5.1). Based on this relationship, we propose a modification of the system: a new strategy for handling PopUp–Incorrect events (i.e. give additional explanations after such events - 4.5.2). Next, we describe a user study designed to investigate the utility of the new strategy (4.5.3). Finally, we present our analyses on the collected corpus (4.5.4). Our results validate the utility of the new strategy and as a result the informativeness of the PopUp–Incorrect bigram.

4.5.1 PopUp–Incorrect – Performance relationship

In Section 4.4.1, we saw that the PopUp–Incorrect bigram is negatively correlated with system performance (i.e. learning) in the F03 corpus. In Section 4.4.2, we show that this finding generalizes across experiments (present in S05) and versions of the system (synthesized S05SYN vs. prerecorded S05PR). This generalization suggests that there is something behind PopUp–Incorrect events that is linked with system performance.

To demonstrate this link we need 3 things: an interpretation of the correlation, a new system strategy and a validation of the new strategy. An interpretation of the correlation will allow us to propose a change in the system, i.e. a new strategy. If we prove that the new strategy is successful, then this validates the interpretation, the strategy and consequently the link.

As mentioned, our interpretation of the correlation between PopUp–Incorrect events and learning is that PopUp–Incorrect events signal failed learning opportunities. A PopUp–Incorrect event is when users go through a remediation dialogue and then give an incorrect answer when being asked again the question that triggered the remediation dialogue (the same question which they answered incorrectly the first time). The remediation dialogue is the failed learning opportunity: the system had a chance to correct user's lack of knowledge and failed to achieve that. The more such events we see, the lesser the system performance.

How can we change the system in light of this interpretation? We propose to *give additional explanations after a PopUp–Incorrect event* as the new strategy. To arrive to this strategy, we hypothesized why the failed opportunity has occurred. The simplest answer is that the user has failed to absorb the information from the remediation dialogue. Responsible for this is probably that the user did not understand the remediation dialogue and/or failed to make the connection between the remediation dialogue and the original question. The current ITSPOKE strategy after a PopUp–Incorrect is to give away the correct answer and move one. The negative correlations indicate that this strategy is not working. Thus, maybe it would be better if the system will engage in additional explanations to correct the user. If we can make the user understand, then we transform the failed learning opportunity into a successful learning opportunity. This will be equivalent to a PopUp–Correct event which we have seen that it is *positively* correlated with learning (4.4.1).

There are other ways we can change the system. For example, maybe the reason why the user exhibits a failed learning opportunity is his/her lack of interest/involvement. In this case, then a strategy that motivates the user might be a better strategy (Aist et al., 2002; Kumar et al., 2007). Another way is to be proactive and prevent occurrence of PopUp–Incorrect events. However it is not very clear how this can be achieved.

How can we prove that the new strategy is successful? There are three ways and we will explore all of them. We can show that the new system outperforms the old system. However, this might not be the best way as the new PopUp–Incorrect strategy directly affects only people with PopUp–Incorrect events. In addition, its effect might depend on how many times it was activated. Indeed, our results (4.5.4.2) find no significant difference between the two systems. Secondly, we can investigate if the correlation between PopUp–Incorrect and learning is broken in the new system. Our results (4.5.4.4) show that this is true. An even stronger result is to show that the new strategy has a positive effect on users that "needed" it the most: users with more PopUp–Incorrect events. Our data shows that learning is improved for this category of users (4.5.4.4).

4.5.2 The new PopUp–Incorrect strategy

In the previous section, we proposed a new PopUp–Incorrect strategy: instead of simply giving away the correct answer, the new strategy will provide additional explanations. This section discusses the implementation of the new strategy.

To implement the new strategy, we had to find a way to deliver the new explanations. One way would have been to engage in an additional subdialogue. However, this was complicated by two factors: an instructional factor and a technological factor. Firstly, we did not know exactly what information to convey and/or what questions to ask. It was crucial that the information and/or the questions were on target due to the extra burden of the new subdialogue. Secondly, implementing new dialogues would have required many changes to the system: implementing the subdialogue in the Why2-Atlas backend, updating the semantic interpretation component for the new questions, building new language models for the automated speech recognition component and recording new system prompts with the voice talent (prerecorded version of ITSPOKE was used in later experiments).

We opted for a different implementation of the strategy: interrupt the conversation at PopUp–Incorrect events and offer the additional explanations in form of a *webpage* that the user will read. Each potential PopUp–Incorrect event had an associated webpage that was displayed in case the event was detected. This implementation did not have the problems mentioned above. Because the information was presented visually, users can choose which part to read, which meant that we did not have to be right on target with our explanations. Moreover, implementing this in the system was much easier than creating new subdialogues.

What information should be included in the webpage? We present and motivate below our choice of information snippets.

1. *Tutor question* – we believed it was important for the user to see the current question primarily to address cases where the user had trouble concentrating on the question. For PopUp transitions, the system uses variations of "Back to the original question" and then plays the question in addition to its feedback to the previous user answer. This can result in a relatively long system turn which users might have trouble following.
2. *Correct answer* – we believed that it was important to ground the correct answer with the user. If users were correct but the system had trouble understanding them, this will be a fast confirmation of their correctness. If users were incorrect, knowing the correct answer might help users follow the explanations available in the webpage.
3. *Dialogue summary* – we presented the relevant part of the Navigation Map (our graphical representation of the discourse structure discussed in Section 3.4). Our previous studies showed that users think they can identify the tutoring plan better when they have the Navigation Map during instruction (6.3.2).
4. *What did we learn so far?* – we provided a very short recap of the information discussed in the problem before this question. Only relevant bits of information were included. We hoped that this information will setup the right informational context for the user.

5. *How did we try to find the correct answer to this question?* – this part is a rehash of the information presented in the remediation dialogue. Where possible, references to the reading material were made and examples used there were adapted to or related to the current question.

These 5 information snippets are primarily based on the instruction and the reading material. Since it is possible that users have problems understanding these, we provided another section with information not discussed by the system. The section was labeled "Additional information" and was captioned "If you still feel unsure about the answer to this question, the material below provides additional information". The section included as many as possible of the following:

6. *Intuition* – We tried to provide intuitions and examples from real life that are similar to the situation described in the question. We hoped that if users have trouble following the conceptual presentation, these intuitions will provide some form of scaffolding.
7. *What is the purpose of this question?* – We tried to make explicit the connection between the current questions and the solution plan. This information offered a forward-looking perspective of the question (i.e. why it is useful to know this information). Careful attention was paid here not to provide answers for future questions.
8. *How does this question relate to the other problems we discussed?* – We tried to make explicit the connection between the situation discussed in the current question and similar situations from previously discussed problems. We included a discussion of the similarities and differences here.
9. *Possible pitfalls* – We discussed possible pitfalls that can lead to an incorrect answer. Primarily, we discussed here how the current situation is different from situations users observe in the real world (e.g. air resistance is ignored in all ITSPOKE problems but it is present in real world – this difference is a fruitful source of wrong answers).

All webpages and their content were designed by the author. All potential places for PopUp–Incorrect events (i.e. questions) were identified in the instruction and a webpage was authored for each question. All webpages had the same presentation structure. Information was presented in the order described above and the same formatting was used in each webpage. There were a total of 96 system questions that the system may ask during the "walkthrough" instruction we planed to use in the user experiment (2.2.4). Out of these, 24 questions (25%) could result in PopUp–Incorrect events and a webpage was authored for each of them. A sample webpage is available in Appendix A.4.5.

The information we included in the PopUp–Incorrect webpages has a "reflective" nature. For example, items 3 and 4 above summarize and discuss the relevant instruction. Item 8 takes this even further by reflecting on the connection between the current problem and previous problems. The value of “reflective” information has been established by previous studies. For example, (Albacete and VanLehn, 2000) improve the performance of an interactive tutor (natural language is not used) when a series of mini-lessons are activated for incorrect answers. These mini-lessons explain links between concepts, address misconceptions or summarize knowledge for users (more details in (Albacete, 1999)). (Katz et al., 2003) show that post-instruction reflective dialogues, either using a human tutor or canned text, lead to more learning.

To detect PopUp–Incorrect events at runtime we used the only form of correctness available at runtime – system correctness (ASEM - 2.3.1.4). As a result, a number of factors can result in an activation of the new strategy. We discuss these factors and the contribution of the new PopUp–Incorrect strategy for each of them.

- *Misunderstanding* – these are cases where the remediation dialogue has failed to correct user’s misunderstanding. For example, if an object is moving in a direction and no forces act on it then the speed will be constant. Many users have a misconception in this case and think that the speed will decrease since no force acts on the object. These situations are the failed learning opportunity events behind our interpretation of the negative correlation between PopUp–Incorrect and learning (4.5.1) and are directly targeted by our new PopUp–Incorrect strategy.
- *Imprecise linguistic choice* – these are cases where the student has the right intuition but his formulation of the answer does not use the right/correct domain concepts. For example, users might refer to the force of impact that acts on two object that collide as collision, neglecting that there are two equal and opposite forces involved in a collision. From the physics perspective, these are incorrect answers but they are not far from a correct answer. These are also situations where the new strategy can have an effect as it will enforce the right linguistic choice.
- *Unresponsive users* – these are cases where the user refuses to give a relevant answer (e.g. swearing, making fun of the system, etc.). In these cases, the strategy can be used by users as a timeout to recuperate.
- *Unhandled correct answers* – these are cases where the student has a correct answer but the semantic interpretation component (2.3.1.4) is unable to interpret the answer. For example, the system might understand “it is smaller” but not “it is less”. For these cases, the new strategy helps users by confirming visually the correct answer.
- *Speech errors* – these are cases where although the student answer is correct, it is interpreted as incorrect due to problems with speech processing. This includes semantic misrecognitions (i.e. correctness of the human transcript is different from correctness of

the system transcript – see Section 2.3.1.5) and clipping of the user speech (e.g. users that start speaking before the system listens or cases where the system thinks users are done talking). For these cases, the new strategy helps users by confirming visually the correct answer.

When users are done reading, they return to the conversation by pressing a button. Two buttons (Yes/No) were available next to the question “Was this information useful?”. We opted for this feedback because we wanted to know if the additional explanations were indeed useful for the user. The factors we described above can also influence user feedback (e.g. users with a correct answer that was interpreted incorrectly are more likely to give a negative feedback).

4.5.3 User study

To test the effect of the new PopUp–Incorrect strategy, we designed and performed a between-subjects study with 2 conditions. In the control condition we used the regular version of ITSPOKE with the old PopUp–Incorrect strategy (i.e. give the current answer and move on), which will be identified as ***R***. In the experimental condition, we had the regular version of ITSPOKE with the new PopUp–Incorrect strategy (i.e. give additional information) which will be identified as ***PI***.

There were two confounding factors we wanted to avoid. To reduce the effect of the quality of the speech output, we used a version of the system with human prerecorded prompts. We also wanted to account for the amount of instruction as in our system the instruction is tailored to what users write in the essay. Thus the essay analysis component was disabled; for all users, the system went through the “walkthrough” instruction which assumed no information in the user 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 actual experiment had an additional experimental condition (*NM*) used to test the objective utility of the Navigation Map (our graphical representation of the discourse structure - 3.4) compared to *R*. Details for that part of the study are available in Section 6.4.1. The experiment was designed for two pairwise comparisons: *NM* vs. *R* and *PI* vs. *R* (i.e. *R* was shared). For more information about the actual experiment and the collected corpus, see Section 2.2.4.

4.5.4 Results

In the following sections we investigate differences between *R* and *PI*. We have already shown in Section 2.2.4.2 that the two conditions are balanced in terms of user abilities (PRE score). We begin with several statistics related to the activation of the new PopUp–Incorrect strategy in *PI* (4.5.4.1). To investigate if the new strategy has any effects, we compare the two conditions in terms of system performance overall (4.5.4.2) or on specific subpopulations (4.5.4.3 and 4.5.4.4). Next, we look at subjective metrics (4.5.4.5) and several dialogue efficiency metrics (4.5.4.6 and 4.5.4.7).

4.5.4.1 Statistics for the new PopUp–Incorrect strategy

For our 27 *PI* users[17], the new strategy was activated 103 times in total, with an average of 3.8 times per user. For each user, we computed how many times the strategy was activated. A distribution of these values is available in Figure 16. 2 users did not have any PopUp–Incorrect events and 1 user had the largest number of PopUp–Incorrect events (8). For users where the strategy was activated at least once, the average time spent by a user with the strategy (i.e. total time spent by user divided by the number of activations for that user) ranges from a minimum of 4.4 seconds to a maximum of 97 seconds with a mean of 26 seconds. As described in 4.5.2, we collected user feedback based on which button they clicked to return to the conversation with the system. 69% of the total activations were considered useful by users. A similar ratio is obtained if we compute usefulness at the user level.

[17] For the 25 *R* users, the strategy would have been activated in 109 cases, with an average of 4.4 times per user (a minimum of 1 per user and a maximum of 8 per user).

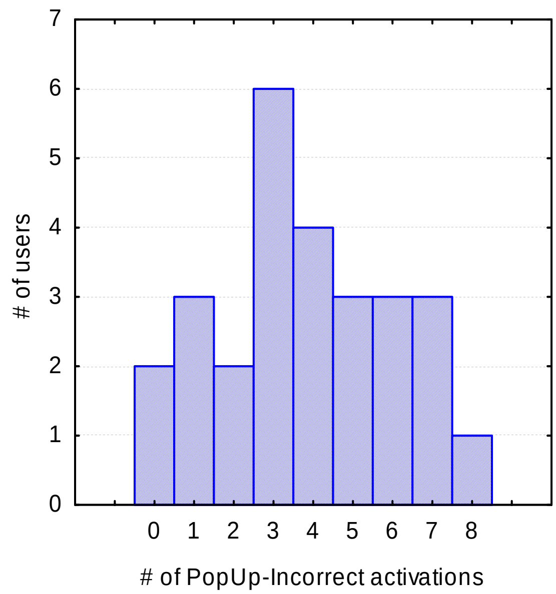


Figure 16. Distribution of the total number of PopUp–Incorrect activations per user (X axis: number of activations, Y axis: number of users with a given number of activations)

The PopUp–Incorrect strategy is activated based on system correctness (ASEM - 2.3.1.4). To measure the impact of speech recognition problems on the activation of the new strategy, we looked at the human transcript correctness (TSEM - 2.3.1.4) for the PopUp–Incorrect events. We found that 18% of these events would have been labeled as correct if it wasn't for speech recognition problems. Surprisingly, users still gave a positive feedback on 68% of these cases (the same percentage with the overall percentage of useful activations).

We also investigated where the new PopUp–Incorrect strategy was activated out of the total of 24 possible places/questions. We find that users cover on average 16% of them (recall 3.8 activations per user on average). The strategy was never activated in our corpus for 4 of the 24 questions. The strategy was activated for individual questions 5.7 times on average with a minimum of 1 and a maximum of 25 activations per question.

In the open question interview at the end of the experiment (2.2.4), we asked users if they had any activations of the strategy and if yes, if it was helpful and which part they found useful. Although this is not a formal evaluation, several of items included in the additional explanations webpage (4.5.2) were mentioned by the user. The graphical dialogue summary (item 3 – the Navigation Map) was mentioned by several users, followed by the recap summary (item 4 - "What did we learn so far?") and the restatement of the information from the remediation dialogue (item 5 – "How did we try to find the correct answer to this question?"). Very few users mention the Additional Information part (items 6-9) while others indicate that they did not even look at it. One user mentioned that he/she found the webpage intimidating because of the length.

Discussion

We saw that the presence of a speech recognition problem does not change the percentage of activations users find useful. It is possible that in these cases, users find the activation useful as a way of confirming the correct answer. This observation raises an interesting question: will a random activation of the new strategy (regardless of the correctness) produce the same effect? Although only another user study can answer this question, we believe that a random activation will be less effective as it will inherently miss some of the incorrect answers and provide unnecessary instruction for correct answers (note that for misrecognized correct answers, the new strategy confirms for users that they were indeed correct).

It is also interesting to observe that the Navigation Map (item 3, the dialogue summary) was mentioned as useful by several users. This complements our other results on the subjective (6.3) and objective (6.4) utility of the Navigation Map.

Summary

We find that the new strategy was activated on average 3.8 times per user. In more than two thirds of these activations users find the additional explanations useful, a percentage which is not influenced by problems from the automated speech recognition. Several items from the additional explanations webpage are indicated to be useful, primarily in the first part.

4.5.4.2 Performance (Learning)

Our hypothesis is that *PI* users will perform better on a number of objective metrics. The most important metric on which we would like to see the impact of the new PopUp–Incorrect strategy is the system performance metric – that is, learning. As mentioned in Section 2.3.1.2, there are various ways of measuring learning, each giving a different perspective on the learning phenomena. We will present the results on all of them.

Table 13 shows the average and standard deviation for each condition on our 2 learning metrics (POST and NLG). We find that *PI* users perform similarly but slightly better than *R* users on both metrics. We run a one-way ANOVA with POST and NLG as dependent variables and the Condition (*PI* vs. *R*) as the categorical factor. Results show no significant differences between the two conditions on any learning metrics (for POST: $F(1, 50)=0.13$, $p<0.72$; for NLG: $F(1,50)=0.06$, $p<0.80$). The corresponding NLG effect size is 0.07.

Table 13. Average and Standard Deviation for learning metrics

Condition	PRE	POST	NLG
PI	12.5 (4.4)	19.4 (2.8)	0.50 (0.20)
R	12.6 (4.2)	19.1 (3.1)	0.49 (0.20)

To test if students learn in the two conditions, we run a repeated-measure ANOVA with the Test Phase as the within-subjects factor (PRE vs. POST scores) and the Condition (*PI* vs. *R*) as the between-subjects factor. We find a significant effect from the Test Phase ($F(1,50)=211.82$, $p<0.0001$) and, as expected, no significant effect from the Condition or the combination Condition × Test Phase (for Condition: $F(1,50)=0.01$, $p<0.91$; for Condition × Test Phase: $F(1,50)=0.16$, $p<0.69$). Posthoc tests show a significant difference between the PRE and POST scores on each condition ($p<0.0001$). In other words, students learn significantly in both conditions but learning is not dependent on the condition.

As in many other tutoring studies, the pretest and posttest scores are positively correlated for the combined *R-PI* population ($R=0.631$, $p<0.001$). We wanted to know if this PRE-POST relationship differs between conditions. Thus, we run a one-way ANCOVA with POST as the dependent variable, PRE as the covariate and the Condition (*PI* vs. *R*) as the independent variable. As expected from the POST/NLG results, we find that the condition has no effect on the PRE-POST covariation ($F(1,49)=0.26$, $p<0.61$). A visual presentation of this result is available in Appendix B.1.2.

Restricting the analyses to the RELNLG subset (2.3.3.2) does not impact the results. Although the differences between *PI* and *R* become bigger (NLG effect size of 0.10), the ANOVA and ANCOVA effects continue to be insignificant.

Summary

There are no differences between *PI* users and *R* users in terms of learning (main system performance metric). In both conditions, users learn significantly but their learning patterns do no differ between conditions.

4.5.4.3 Aptitude – Treatment interaction on Performance

It is a common analysis practice in tutoring research to investigate the relationship between user aptitudes and his/her performance while working with the system. Many studies have shown that the treatment condition can produce effects only on specific subsets of the populations and that, in some cases, the treatment has opposite effects depending on the subset. Here, we will use the initial physics knowledge aptitude (PRE score - 2.3.1.1) and the subsets created by the mean split on this aptitude (PRE Split - 2.3.3.1). Although there are no results for the other aptitude measured in this experiment (working memory span), for completeness we present that analysis in Appendix B.1.3.

To investigate the effect of the PRE Split and condition on performance (learning), we run a factorial ANOVA with POST and NLG as dependent variables and two factors: the aptitude split (low vs. high pretesters) and the Condition (`PI` vs. *R*). We look at the effect of the aptitude factor and the effect of the aptitude-condition combination. The effect of the condition was discussed in the previous section.

Table 14 shows the average and standard deviation for the two relevant learning metrics (POST and NLG) for PRE Split and the combination PRE Split × Condition. We find a strong significant effect of PRE Split on POST ($F(1,48)=15.32$, $p<0.0003$) with high pretesters reaching a significantly higher POST score compared with low pretesters (21.0 on average versus 18.1). However, the effect on NLG, which accounts for PRE score, is far from significance ($F=0.20$, $p<0.65$).

Table 14. Learning performance based on PRE Split and Condition

PRE Split	Condition	# of users	POST	NLG
L		31	18.1 (2.4)	0.51 (0.15)
H		21	21.0 (2.8)	0.48 (0.25)
L	*PI*	16	18.2 (2.4)	0.52 (0.16)
L	*R*	15	17.9 (2.4)	0.50 (0.15)
H	*PI*	11	21.1 (2.5)	0.48 (0.25)
H	*R*	10	20.8 (3.3)	0.48 (0.26)

The combination PRE Split × Condition has no effect on both learning metrics (for POST: $F(1,48)=0.01$, $p<0.99$; for NLG: $F=0.04$, $p<0.83$). The averages are very similar with *PI* having slightly better values in most cases. In terms of NLG effect size, the new PopUp–Incorrect strategy has a very small, insignificant 0.15 effect size for high pretesters and a zero effect size for low pretesters. Restricting to the RELNLG subset (2.3.3.2) produces no significant results.

Summary

We find that the effect of treatment on learning does not depend on user aptitudes collected in this experiment.

4.5.4.4 PopUp–Incorrect – Performance interaction

We have seen in Section 4.5.4.2 that *PI* users learn slightly more than *R* users but the difference is far from being significant. Thus, the new PopUp–Incorrect strategy does not have a strong enough effect to change the learning of the population. However, it is possible that the new PopUp–Incorrect strategy effect is more apparent for users where it was engaged (more).

There are two factors behind our new PopUp–Incorrect strategy. First, there is the correlation factor: the negative correlation between PopUp–Incorrect bigram parameters and learning (4.4.1).

Second, there is the interpretation factor: our hypothesis that a PopUp–Incorrect event signals a failed learning opportunity (4.5.1). We will investigate if the new PopUp–Incorrect strategy has any interactions with the two factors.

If the new PopUp–Incorrect strategy has an effect of the correlation factor, we should find that PopUp–Incorrect bigram parameters[18] are still correlated with learning for *R* students but the correlations are weaker for *PI* students. Table 15 shows the best correlation coefficient and its significance[19] for both correct and incorrect answer after a PopUp transition on each of the two conditions. Correlation information for each parameter is available in Appendix B.1.1, Table 41. For the *R* users, we find that PopUp–Correct bigrams are positively correlated with learning while PopUp–Incorrect bigrams are negatively correlated with learning. Only the PopUp–Incorrect bigram results in a trend correlation. These correlations are in line with our investigations from other corpora (see Section 4.4.1, Table 9 and Section 4.4.2, Table 12). We did not expect very strong correlations as we used only the walkthrough dialogues in this experiment[20].

---

[18] To compute the PopUp-Correctness bigrams we use the transcript correctness (TSEM - 2.3.1.4). The same correctness annotation was used to investigate if the predictiveness of PopUp–Incorrect generalizes to other corpora (4.4.2). Please note that this correctness will ignore activations of the new PopUp–Incorrect strategy that are solely due to speech recognition problems. 4-way correctness was used (i.e. Correct, Incorrect, Partially Correct and Unable to Answer).

[19] As in our analysis of predictiveness (4.4), we only report the best partial correlation coefficient (R) out of the correlation coefficients computed for each of the three parameters derived from a bigram (e.g. for PopUp–Incorrect bigram, we compute the count, percentage and relative percentage parameters - recall Section 4.3). When comparing R and PI in terms of the best R we are not comparing on the same parameter. However, the comparison still makes sense because when we choose the parameter that produces the best R for the R condition, its corresponding R for the PI condition is equal or worse than the best R reported for PI. For example, for the PopUp-Incorrect bigram, the best R (-0.34) comes from the percentage parameter for R; the corresponding R for the percentage parameter in PI is -0.16 ($p<0.46$).

[20] We repeated the predictiveness generalization investigation from Section 4.4.2 on walkthrough dialogues only and the correlations between PopUp–Incorrect and learning are still negative but weaker (trend on F03, near trend on S05PR and very weak on S05SYN). However, some of the users from these corpora did not go through all walkthrough dialogues as all users in the Main experiment did.

Table 15. Partial correlations for PopUp–Correctness bigram parameters

	PopUp-Correct			PopUp-Incorrect		
	Avg (StdDev)	Best R	p	Avg (StdDev)	Best R	p
R	6.7 (2.6)	0.30	0.15	2.6 (1.1)	*-0.34* ^{t}	*0.10*
PI	5.1 (2.3)	0.24	0.23	2.5 (1.6)	-0.23	0.27

For the *PI* users, the correlations follow the same pattern but are weaker: the correlation coefficient is 20% weaker for PopUp–Correct and 35% weaker for PopUp–Incorrect. This suggests that our new PopUp–Incorrect strategy reduces the connection between user behavior after a PopUp transition and learning.

Restricting to PopUp–Incorrect events in the last 2 problems

We argue that in ITSPOKE *later PopUp–Incorrect events are more important* for our purpose (i.e. to signal failed learning opportunities). First, even if users answer incorrectly after a PopUp transition in one of the earlier problems, they still have other chances to correct their misconceptions in the later problems especially if the same physics laws or concepts are applied in a similar fashion. ITSPOKE discusses increasingly more complex problems with each problem building on previous ones (see Appendix A.2). For example, the second problem discusses a situation in which an object is in freefall; in the fourth problem, we have an object that is first accelerated up and then is in freefall. Second, during the authoring of the new PopUp–Incorrect strategy for each remediation dialogue, we noticed that the information taught in some of the remediation dialogues is less important than in other remediation dialogues or is being repeated in later problems. For example, the first problem has four remediation dialogues that discuss inferences that will not be reused (i.e. they demonstrate that the freefall acceleration is equal to gravitational acceleration, but students can simply remember this as a rule). In contrast, most remediation dialogues in later problems discuss about applications of physics law in specific contexts. Some remediation dialogues from earlier problems are repeated in later problems (e.g. second and fourth problems discuss similar situations). In contrast, there is a minimal overlap between the remediation dialogues in the last two problems.

Since an annotation of the "importance" of each remediation dialogue is subjective, we opted for an objective way to approximate importance based on the observations discussed in the previous paragraph: *we only counted PopUp–Incorrect events in the last 2 problems*[21]. This restriction applies to the reminder of this section.

[21] When computing the percentage and relative percentage parameters for each bigram, only turns from the last 2 problems are used (it is as if only these two last problems were discussed).

Table 16 shows the bigram-learning correlations for the last two problems (correlation information for each parameter is available in Appendix B.1.1, Table 42). We find that for *R* users the connection between user behavior after a PopUp and learning is even stronger if we restrict to the last two problems. There is a significant correlation for both PopUp–Correct and PopUp–Incorrect bigrams. In contrast, for *PI* users the correlation is even weaker than the one presented in Table 15. These findings suggest that our hypothesis that later PopUps are more important is true and also that, indeed, the new PopUp–Incorrect strategy has some effect.

Table 16. Partial correlations for PopUp–Correctness bigram parameters (last 2 problems only)

	PopUp-Correct			PopUp-Incorrect		
	Avg (StdDev)	Best R	p	Avg (StdDev)	Best R	p
R	2.9 (1.1)	**0.52** *	0.01	1.5 (0.8)	**-0.52** *	0.01
PI	2.4 (1.3)	0.10	0.63	1.5 (0.8)	-0.19	0.35

If in addition we restrict to the RELNLG subset (2.3.3.2), the correlations become even stronger for *R* (Best R is 0.60 and 0-.65 respectively) and remain the same for *PI* users (Best R is 0.18 and -0.18 respectively). One way to visualize the partial correlations is to look at the correlations with NLG (this metric also accounts for the PRE score). Figure 17 shows a scatter plot of the PopUp–Incorrect relative percentage parameter and NLG for each *PI* and *R* user. The regression lines for the correlation between PopUp–Incorrect and NLG for *PI* and *R* are shown: PopUp–Incorrect is significantly negatively correlated with NLG for *R* users (R=-0.65, p<0.001); for *PI* users we find only a very weak correlation (R=-0.11, p<0.60). The graph shows that users with less PopUp–Incorrect events (e.g. less than 30% relative) tend to have a higher NLG (0.5 or higher). However, for users with more PopUp–Incorrect events, the behavior depends on the condition: *R* users (crosses) tend to have lower NLG (0.5 or lower) while *PI* users (circles) tend to cover the whole NLG spectrum (0.2 to 0.73). Our next analysis will provide objective support for this observation.

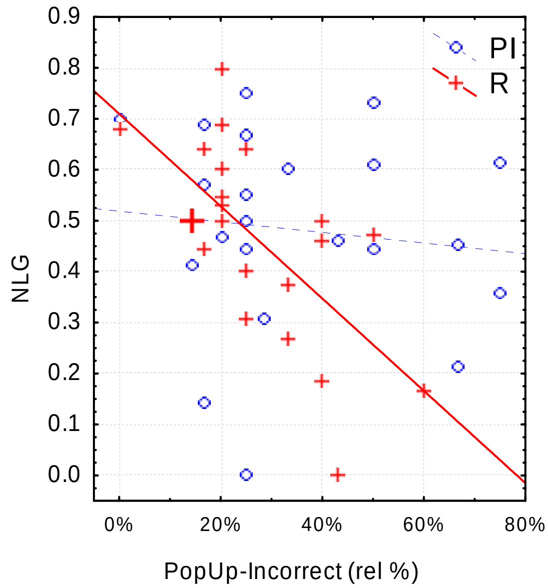


Figure 17. Correlations between PopUp–Incorrect and NLG
(RELNLG subset, PopUp–Incorrect events from last 2 problems)

Next we turn our attention to the interpretation factor: our hypothesis that a PopUp–Incorrect is an instance of a failed learning opportunity. If this is true and our new PopUp–Incorrect strategy is effective, then `PI` users with a higher number of PopUp–Incorrect events should achieve more learning than similar *R* users. To test this, we computed a mean split based on user PopUp–Incorrect (**PI Split**) behavior in a high and low subset[22]. We look only at the last two problems as the connection between PopUp–Incorrect and learning is clearer in this case (recall the correlations results from Table 15 and Table 16).

To see if the PI Split has any effect on learning, we run a two-way ANCOVA with POST as the dependent variable, PRE as the independent covariate and two factors: the Condition (`PI` vs. *R*) and the PI Split (Low vs. High). We find a trend effect of the PI Split ($F(1,47)=3.71$, $p<0.06$) and a trend effect of the combination PI Split × Condition ($F(1,47)=2.80$, $p<0.10$). Figure 18 shows how the POST-PRE covariation is affected by the PI Split. The lines show the regression line for each split. The trend effect of the PI Split in ANCOVA tells us that the two regression lines are different (recall 2.4.4). We see that the regression line for low PopUp–Incorrect users has a higher intercept and a smaller slope than the regression line for high PopUp–Incorrect users. A regression line with a higher intercept and a smaller

[22] We used the mean split based on the PopUp–Incorrect relative percentage parameter for the R users computed in the last 2 problems (average value is 29.9%). We chose this parameter because its correlation with learning for R users is the strongest.

positive slope is more desirable since the small slope is an indication of the fact that the pretest score has a smaller effect on the posttest score (i.e. most people will reach a similar posttest score); a higher intercept means that the posttest scores achieved are higher. Indeed, the average adjusted posttest score (2.4.4) is higher for low PopUp–Incorrect users (19.7) than for high PopUp–Incorrect users (18.6). This interaction confirms again the fact that an increase in PopUp–Incorrect events is linked with reduced learning. On the RELNLG subset (2.3.3.2) the effect of PI Split becomes significant (F(1,42)=7.62, p<0.009). For this subset, the regression lines for low/high PI Split are almost parallel but the intercept of the low PI Split is higher.

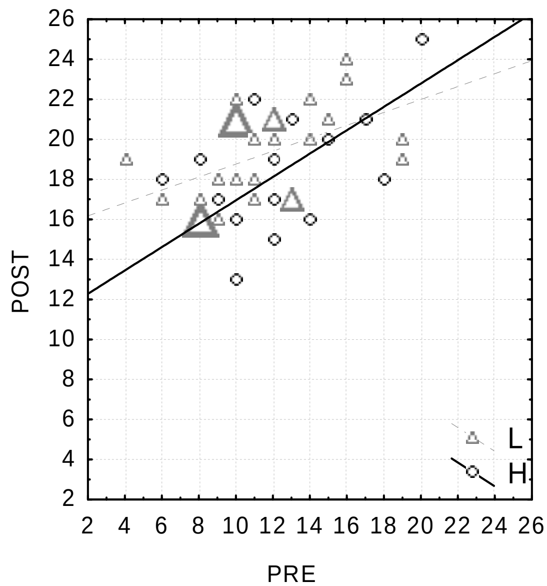


Figure 18. PI Split effect on POST-PRE covariation
(PopUp–Incorrect events from last 2 problems)

Figure 19 summarizes the effect of the combination PI Split × Condition. The graph on the left shows the regression lines for POST-PRE covariation for high PI Split users in the *PI* and *R* condition. We find a "better" regression line in the *PI* condition than in the *R* condition: a smaller slope and a higher intercept. For low PI Split users, the *PI* and *R* regression lines are almost similar (graph not shown). These findings indicate that the new PopUp–Incorrect strategy has a positive effect on learning for high PopUp–Incorrect users but no effect for low PopUp–Incorrect users. In other words, people that need and are exposed to the new strategy benefit the most. A similar thing can be observed if we look at the adjusted posttest scores (2.4.4) in the right side of Figure 19. We see that the adjusted posttest score for *PI* users is similar regardless of the PI Split. In contrast, for *R* users, the adjusted posttest score is significantly different between the high and low PI Split ($p<0.03$). Restricting to the RELNLG subset (2.3.3.2) makes the effect of the combination PI Split × Condition significant ($F(1,42)=4.47$, $p<0.04$).

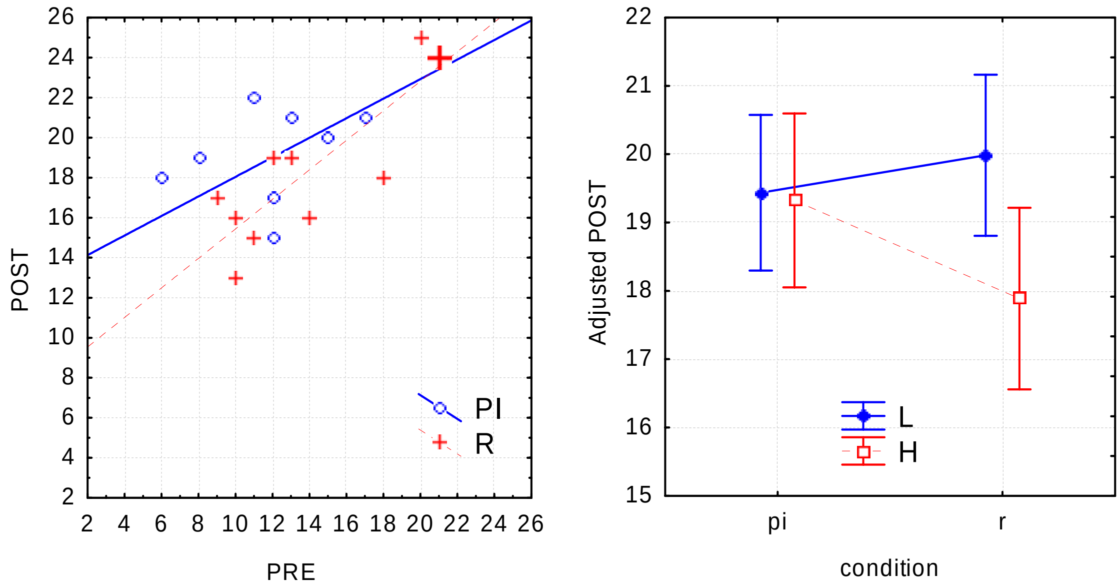


Figure 19. PI Split × Condition effect on POST-PRE covariation

(PopUp–Incorrect events from last 2 problems; 95% confidence intervals shown in right graph)

To make the ANCOVA results easier to interpret, we also looked at the effect on learning as measured by NLG. We run a factorial ANOVA with NLG as the dependent variable and two factors: PI Split (Low vs. High) and Condition (*PI* vs. *R*). We present the results on the RELNLG subset[23] (2.3.3.2). We find a significant effect of the PI Split ($F(1,43)=6.29$, $p<0.02$) and the PI Split × Condition interaction ($F(1,42)=5.13$, $p<0.03$). Figure 20 shows the effect of the PI Split × Condition combination. The findings are similar to the ANCOVA findings: *PI* users have a similar NLG regardless of their PopUp–Incorrect behavior while for *R*, high PopUp–Incorrect users learn less than low PopUp–Incorrect users. Posthoc tests indicate that high PopUp–Incorrect *R* users learn significantly less than low PopUp–Incorrect *R* users ($p<0.01$) and both categories of *PI* users ($p<0.05$). Interestingly, low PopUp–Incorrect *R* users learn better than both categories of *PI* users but the difference in not significant. The corresponding NLG effect size is 0.96 for high PopUp–Incorrect users and negative 0.52 for low PopUp–Incorrect users. The fact that in the *PI* condition there are no trend/significant differences between low and high PopUp–Incorrect users suggests that the new PopUp–Incorrect strategy is paying off.

[23] If we do not eliminate these users, PI Split * Condition has a trend interaction on NLG ($F(1,48)=2.87$, $p<0.10$)

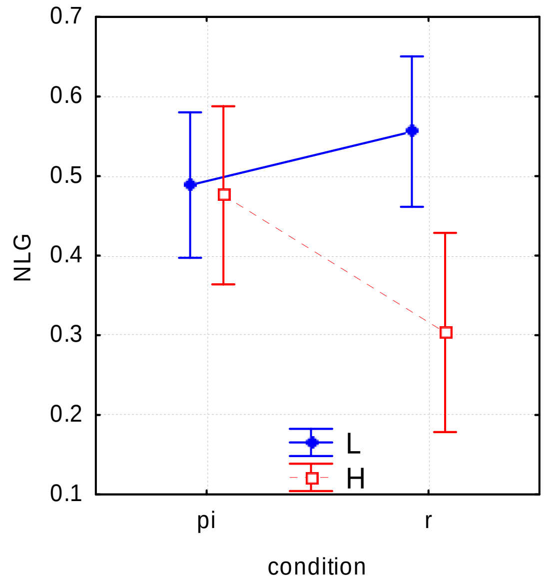


Figure 20. PI Split × Condition effect on NLG

(RELNLG subset, PopUp–Incorrect events from last 2 problems; 95% confidence intervals shown)

Discussion

It is interesting to speculate why the performance improvements observed for certain users (i.e. users with a larger number of PopUp–Incorrect) do not translate in improvements at the population level (recall Section 4.5.4.2). In Figure 20 we saw that for high PopUp–Incorrect users, `PI` significantly outperforms `R` (effect size of 0.96). However, for low PopUp–Incorrect users, `PI` performs non-significantly lower than `R` (negative effect size of 0.52). As a result, when we look at the population level we see a very small non-significant improvement from `PI` (effect size of 0.10). One hypothesis is that the new PopUp–Incorrect strategy has a negative effect for users with a low number of PopUp–Incorrect events. However, these users have little exposure to the modification because they have a lower than average number of PopUp–Incorrect events (recall definition of the PI Split). Coupled with the fact that users can dismiss the additional information webpage with ease (recall Section 4.5.2), we believe that it is very unlikely for this hypothesis to be true.

Other hypotheses seem to be better at explaining this situation. One of them is that we have a small sample and that the difference between `R` and `PI` for low PopUp–Incorrect users will go away once we increase the sample size. A new, larger experiment can be run to test this hypothesis. Another hypothesis is that not all learning issues are signaled by PopUp–Incorrect events: a user might still have low learning even if he/she does not exhibit any PopUp–Incorrect events. Indeed, there are two `PI` user with a single PopUp–Incorrect event but with very low learning (PRE=19 → POST=19 and PRE=19 → POST=20 respectively). It is very likely that other things went wrong for these users rather then the activation of the new PopUp–Incorrect strategy (e.g. they might have other misconceptions that are not addressed with remediation subdialogues). In fact, removing these two users results in identical NLG averages for *PI* and *R* users with a low number of PopUp–Incorrect events (i.e. the NLG effect size becomes 0).

Summary

We find that PopUp–Incorrect events continue to be associated with decreased learning for *R* users, but the same does not hold for *PI* users. The effect of the new PopUp–Incorrect strategy is even more visible on the correlations when we acknowledge that some PopUp events are more important than others by looking at such events in the last 2 problems only. By splitting based on the PopUp–Incorrect events, we find that *PI* users with a high number of such events learn more than their *R* counterparts and similarly to *PI* users with a low number of such events. This confirms that the new PopUp–Incorrect strategy is effective where it is most needed (users with a high number of PopUp–Incorrect events).

4.5.4.5 Subjective metrics

While there are no differences between the PI and *R* user population in terms of learning, we wanted to know if their perception of the system was changed by the PI. We find no trend/significant differences between the ratings of the two populations. The ranking of the two versions based on the average rating depends on the question but there are no significant differences. We also looked at grouping questions based on timing (see Section 6.4.2.3). We again find no significant/trend differences. In fact, the averages of the ALL ratings are very similar for *PI* and *R* (3.48 for *PI* and 3.52 for *R*).

Discussion

The fact that there are no differences between *PI* and *R* is actually a positive result. *PI* is not very different from *R*: the same graphical interface is being used and the only difference is that the dialogue is interrupted after a PopUp–Incorrect and a webpage with additional explanations is presented in the *PI* condition. The fact that there are no differences confirms the stability of the subjective metrics and strengthens the subjective differences we will find between *NM* and *R* in Section 6.4.2.3.

Summary

PI users and *R* users rate the system similarly on all questions from the system questionnaire. This result is not surprising given that the *PI* version changes the system very little over *R* version.

4.5.4.6 Dialogue time

We also wanted to know if the new PopUp–Incorrect strategy has an effect on measures of dialogue efficiency. The new PopUp–Incorrect strategy delivers additional explanations which can result, for example, in an increase in the time users spend with the system (users spend time reading the new instruction). Also, when designing tutoring systems researchers strive for learning efficiency: deliver increased learning as fast as possible. We look at two shallow dialogue metrics: dialogue time and number of turns.

Section 2.3.3.3 defines and discusses the dialogue duration metric. The metric was computed for each problem (**P1Time-P5Time**) and as a total (**TotalTime**). Table 17 shows the average and standard deviation of the 6 dialogue time metrics (in seconds) in each condition. We find that on most metrics *PI* users spend less time conversing with the system however the differences are not significant. ANOVA tests find no differences between conditions on this metric. The most important difference is on the 4th problem (for P4Time: $F(1,50)=1.05$, $p<0.31$). In total, *PI* users spend on average 1 minute less with the system though the difference is not significant.

Table 17. Average and standard deviation for dialogue time

Condition	P1Time	P2Time	P3Time	P4Time	P5Time	TotalTime
PI	681 (190)	439 (148)	135 (17)	646 (163)	693 (153)	2594 (418)
R	726 (199)	427 (148)	130 (13)	688 (126)	701 (167)	2671 (368)

Since user initial knowledge (the PRE score) is an important factor in how much time the user spends with the system, we wanted to know if the differences are influenced by this factor. Indeed, the PRE score is significantly negatively correlated with the TotalTime for the combined *R* and *PI* population ($R=-0.46$, $p<0.001$). To account for PRE in our comparison, we run an ANCOVA with dialogue time metrics as dependent variables, PRE as the covariate and the condition as the independent variable (*PI* vs. *R*). The results are similar to ANOVA tests: no significant effect of the condition.

Discussion

It is interesting that on average *PI* users spend less time than *R* users even though *PI* users have to spend an extra time reading the information activated by the new PopUp–Incorrect strategy. Thus the new content does not result in users spending more time with the system. We hypothesize that 2 factors are at play here. First, the additional information activated by the new PopUp–Incorrect strategy might have a positive effect on users' correctness for future system questions especially on questions that discuss similar topics[24]. As a result, the system has to correct the user less and, consequently, finish faster. Second, the average total time *PI* users spend reading the additional information is very small (about 2 minutes) compared to the average dialogue time (about 45 minutes for *R* users).

Summary

In spite of having to go through the additional information activated by the new PopUp–Incorrect strategy, on average *PI* users spend 1 minute less conversing with the system than *R* users though the difference is not significant. No significant differences are found on the dialogue time for individual problems.

4.5.4.7 Number of turns

As another metric that measures the dialogue length, we also looked at the number of system turns. Section 2.3.3.4 defines and compares this metric with the dialogue duration metric. The metric was computed for each problem (**P1Tut-P5Tut**) and as a total (**TutTotal**). Table 18 shows the average and standard deviation of these metrics. The results are more conclusive than for the dialogue time. On most metrics, *PI* users have smaller averages[25] and ANOVA tests find a significant difference on the first problem (F(1,50)=4.87, p<0.04) and a trend for the total number of tutor turns (F(1,50)=3.98, p<0.06). In total, *PI* spend have on average almost 5 turns less than *R* users.

Table 18. Average and standard deviation for the number of system turns

Condition	P1Tut	P2Tut	P3Tut	P4Tut	P5Tut	TotalTut
PI	22.8 (4.0)	16.5 (4.4)	7.4 (0.9)	19.6 (4.0)	19.0 (1.9)	85.3 (7.7)
R	25.6 (5.2)	16.8 (5.2)	7.1 (0.3)	21.4 (4.0)	19.2 (2.4)	90.1 (9.7)

Since PRE score is also significantly negatively correlated with total number of system turns (R=-0.43, p<0.001), we run an ANCOVA to discount for PRE in our analyses. We find that the results are strengthened: significant effect of condition on the first problem and also in total (for TutTotal: F(1,49)=5.08, p<0.03). There is also a trend effect on the fourth problem (F(1,49)=2.87, p<0.10).

[24] Note that our system questions are not coded for tutoring topics. Nonetheless, it would be an interesting direction for future work.

[25] We do not expect any differences on the third problem: it is the smallest problem with no remediation dialogues. Any differences in the number of system turns are solely due to speech problems (i.e. rejections and timeouts).

Summary

Although there are no significant differences in the dialogue time, there is a strong trend that *PI* users have fewer system turns that *R* users.

4.6 RELATED WORK

Designing robust, efficient and usable SDS is a very complex process that is still not well understood by the SDS research community (Möller and Ward, 2008). As a first step towards understanding this process, researchers have looked at ways to evaluate and compare systems. Knowing how to evaluate and compare systems allows one to prescribe design methodologies (e.g. which strategies work better than others). One of the early approaches to evaluation was based on the notion of reference answers (Hirschman et al., 1990). In this approach the system answer is compared to a number of reference answers (similar to the BLEU score for evaluating machine translation systems (Papineni et al., 2002)). However this approach fixes the evaluation to a particular dialogue strategy. To compare systems with different dialogue strategies a variety of other metrics have been proposed. Some of the metrics take a global view: e.g. task completion, user satisfaction (Singh et al., 2002; Walker et al., 1997). Others look at specific aspects of the dialogue: e.g. inappropriate utterance ratio, turn correction ratio, implicit recovery (Danieli and Gerbino, 1995; Smith and Gordon, 1997).

Once we have a battery of evaluation/performance metrics, we can compare multiple systems or multiple versions of a system. But what do these metrics and the resulting comparisons tell us about designing SDS? There are several approaches to answering this question, each requiring a different level of supervision. One approach that requires little human supervision is to use reinforcement learning. In this approach, the dialogue is modeled as a (partially observable) Markov Decision Process (Levin et al., 2000; Roy et al., 2000; Singh et al., 2002). A reward is given at the end of the dialogue (i.e. the evaluation metric) and the reinforcement learning process propagates back the reward to learn what the best strategy to employ at each step is. Other semi-automatic approaches include machine learning and decision theoretic approaches (Levin and Pieraccini, 2006; Paek and Horvitz, 2004). However, these semi-automatic approaches can in general be applied only to small and limited domains though recent work has shown how more complex domain can be modeled (Young et al., 2007).

Another approach is through performance analysis: finding factors that affect the performance as measured by the performance metric. This approach was pioneered by the PARADISE framework (Walker et al., 1997). In PARADISE, the SDS behavior is quantified in form of *interaction parameters*.

These parameters are then used in a multivariate linear regression to predict the target performance metric. Note that certain factors are intuitively related to performance. For example, problems with automated speech recognition can derail a dialogue from the normal course: e.g. non-understandings, misunderstandings, end-pointing, etc. However, new factors can be discovered by looking at the parameters that are included in the final performance model (e.g. in our work, the PopUp–Incorrect bigram was one such factor).

A critical ingredient in performance analysis is the relevance of the interaction parameters for the SDS success. A number of parameters that measure the dialogue efficiency (e.g. number of system/user turns, task duration) and the dialogue quality (e.g. recognition accuracy, number of rejections or help requests) have been shown to be successful (Walker et al., 2000). An extensive set of parameters can be found in (Möller, 2005; Walker et al., 2001). Several information sources are being tapped to devise parameters classified by (Möller, 2005) in several categories: dialogue and communication parameters (e.g. dialogue duration, number of system/user turns), speech input parameters (e.g. word error rate, recognition/concept accuracy) and meta-communication parameters (e.g. number of help request, cancel requests, corrections).

In the intelligent tutoring community (recall ITSPOKE is a speech-based computer tutor), correlations between interaction parameters and learning are the common approach to performance analysis (2.4.2). For these systems, the primary evaluation/performance metric is learning (2.3.1.2). Besides many of the parameters used in the SDS work, parameters that model various aspects of learning are used: evidence of cohesion (Ward and Litman, 2006), number of words and duration of user turns and ratios to equivalent metrics in system turns (Litman et al., 2004).

However, most of these parameters do not take into account the discourse structure information. A notable exception is the DATE dialogue act annotation from (Walker et al., 2001). The DATE annotation captures information on three dimensions: speech acts (e.g. acknowledge, confirm), conversation domain (e.g. conversation- versus task-related) and the task model (e.g. subtasks like getting the date, time, origin, and destination). All these parameters can be linked to the discourse structure but flatten the discourse structure. Moreover, the most informative of these parameters (the task model parameters) are domain dependent.

In Section 4.3 we propose using the hierarchical aspect of discourse structure to derive interaction parameters. We exploit this information to measure two types of events: transition–phenomena events and transition–transition events. Our results show discourse structure-based parameters are predictive of system performance (4.4). In addition, we also prove the informativeness of one such parameter by showing that the modification it suggests leads to improvements on a number of objective metrics (4.5).

Our work extends over previous work on several dimensions. First, we exploit in more detail the hierarchical information in the discourse structure through the notion of discourse structure transitions. Second, in contrast to previous work (Walker et al., 2001), our usage of discourse structure is domain independent. Third, we exploit discourse structure as a contextual information source. To our knowledge, previous work has not employed parameters similar to our transition–phenomena bigram parameters (4.3). Fourth, via the transition–transition bigram parameters, we exploit trajectories in the discourse structure as another domain independent source of information for performance modeling. Finally, similar to (Forbes-Riley and Litman, 2006), we are tackling a more problematic performance metric: learning. While the requirements for a successful information access SDS are easier to spell out, the same can not be said about tutoring SDS due to the currently limited understanding of the human learning process.

In addition, in our work we also look at the informativeness of interaction parameters that are predictive. Most of the previous work stops at the predictiveness step. A notable exception is the work by (Litman and Pan, 2002). The factor they look at is user's having multiple speech recognition problems in the dialogue. This factor is well known in the SDS field and it has been shown to be predictive of system performance by previous work (e.g. (Walker et al., 2000)). To test the informativeness of this factor, Litman and Pan propose a modification of the system in which the initiative and confirmation strategies are changed to more conservative settings (initiative: user→mixed→system, confirmation: no confirmation→implicit→explicit) whenever the event is detected. Their results show that the modified version leads to improvements in terms of system performance (task completion). We extend over their work by looking at a factor (PopUp–Incorrect) that was not known to be predictive of performance beforehand. We discover this factor through our empirical analyses of existing dialogues and we show that by addressing it (the new PopUp–Incorrect modification) we also obtain performance improvements (at least for certain users). In addition, we are looking at a performance metric for which significant improvements are harder to obtain with smaller system changes (e.g. (Graesser et al., 2003)).

4.7 SUMMARY & FUTURE WORK

We find that discourse structure information is useful for performance analysis. We use the transition information (introduced in Section 3.3) that is part of discourse structure in combination with other dialogue phenomena to derive a number of factors (i.e. transition–phenomena and transition–transition bigrams/interaction parameters). Our results show that these factors are both predictive and informative. In terms of predictiveness, we find that several of these factors are correlated with performance. The predictiveness of one of these factors (PopUp–Incorrect) generalizes to other corpora. This factor also informs a promising modification of our system: offer additional explanations after PopUp–Incorrect events. We implement this modification and we compare it with the original version of the system by running a user study. Analysis of the data from the user study reveals several interesting findings. We find that the modification breaks down the negative correlation between the PopUp–Incorrect factor and system performance. In addition, users that need the modification the most (i.e. users with a larger number of PopUp–Incorrect events) show significant improvement in performance in the modified version of the system over corresponding users in the old version of the system. However, this improvement is not strong enough to generate significant differences at the population level. Even though the additional explanations that are part of the new modification add extra time to the dialogue, overall we see no increase in dialogue length in terms of dialogue time or number of system turns for the new version of the system.

Our definition of discourse structure transitions is domain independent. Coupled with our approximation of the discourse structure hierarchy, this makes replicating our experiments on other domains a relatively straightforward process. First, a similar automatic annotation of the discourse structure can be performed in SDS that rely on dialogue managers inspired by the Grosz & Sidner theory of discourse (e.g. (Bohus and Rudnicky, 2003; Rich and Sidner, 1998)). Discourse transitions can be automatically computed using the procedure described in Section 3.3. The transition–transition bigrams are already domain independent and can be easily computed once we have the transition information. Finally, for the transition–phenomena bigrams, researchers have only to identify relevant dialogue phenomena (e.g. speech recognition problems, corrections, out-of-domain utterances, semantics of user input, etc). In addition, it would be interesting to replicate our analyses of predictiveness for other metrics in other systems: e.g. task completion, user satisfaction, metrics that include business costs (Levin and Pieraccini, 2006; Paek and Horvitz, 2004), etc.

The predictiveness of the discourse structure-based parameters suggests their potential for PARADISE modeling (Walker et al., 1997). In fact, work in collaboration with other colleagues (Forbes-Riley et al., 2008) has already made use of our transition–user affect and transition–transition parameters. Although the purpose of that study was to show that the inclusion of parameters that use user affect increases the quality and generality of PARADISE models, results show that the resulting PARADISE models make heavy use of the discourse structure-based parameters. It would be interesting to do a similar study that looks at the relative utility of the discourse structure-based parameters for PARADISE modeling.

Our work suggests that the position in the dialogue and the structure of the dialogues are two important information sources. We define the position in the dialogue based on the current discourse transition and we approximate the dialogue structure through transition trajectories of length two. Both definitions are domain independent and can be computed automatically. In our tutoring system, each discourse structure transition corresponds to important events in the tutoring process (e.g. a Push indicates that the student has an important knowledge gap which is being addressed through a remediation dialogue). Thus, our definitions are tied to these important events. It would be interesting to study how these interpretations change in other systems and domains and if additional transitions are needed to capture behavior of interest. Visualization of dialogue corpora (e.g. (Abella et al., 2004)), is another domain where discourse transitions and the two concepts can have a positive impact.

In terms of design methodology for SDS in tutoring, our results suggest the following design principle: "do not give up on students but try other approaches" (i.e. in our case, we do not give up after a PopUp-Incorrect but we provide additional explanations). Note that our experiment design provides no direct support for this design principle. This is because our user study lacks a third condition typically used in adaptation studies in SDS (e.g. (Pon-Barry et al., 2006)). This third condition will activate the additional explanations randomly instead of using the appropriate trigger (i.e. the PopUp–Incorrect event). For example, in (Pon-Barry et al., 2006) the authors investigate a third condition where the modification is not triggered by student certainty and it is always activated. This third condition allows them to link any improvement to certainty and not only to the additional explanations. This condition is not required in this work due to our performance analysis perspective: we are given a system, we perform an offline analysis of the system using the discourse structure information, we propose a modification and then we test if the modification improves the system. However, it would be interesting to run additional user studies to test this design principle.

5 APPLICATIONS FOR CHARACTERIZATION OF DIALOGUE PHENOMENA

5.1 INTRODUCTION

Human-human and human-computer dialogues are more than just a simple, structured exchange of verbal information. A number of dialogue phenomena are likely to occur during any conversation: lexical and semantic ambiguity, misrecognitions, non-understandings, interruptions, overlapping, affect etc. Detecting and handling such phenomena has a big impact on the robustness and usability of a SDS. For example, detecting and handling speech recognition problems is crucial for a SDS. One approach towards detecting and handling dialogue phenomena is to first try and characterize phenomena: why do phenomena occur and where. Answers to these questions can help us build predictive models for detecting the dialogue phenomena and formulate strategies to handle them.

We will use discourse structure information to characterize two dialogue phenomena: user affect and speech recognition problems.

Detecting and adapting to **user affect**[26] is currently being pursued by many researchers as a method of improving the quality of SDS (Batliner et al., 2003; Lee et al., 2002). This direction has received a lot of attention in the tutoring domain where affective reasoning is explored as a method of closing the performance gap between human tutors and current machine tutors (Aist et al., 2002; Forbes-Riley and Litman, 2005; Forbes-Riley et al., 2008; Pon-Barry et al., 2006). In this domain, there is particular interest in detecting and responding to *student uncertainty*. Researchers hypothesize that student uncertainty creates an opportunity for constructive learning to occur (VanLehn et al., 2003) and studies have shown a positive correlation between uncertainty and learning (Craig et al., 2004). (Forbes-

---

[26] We use the term "affect" loosely to cover both affects and attitudes that can impact performance or that can affect how users communicate their answers. Although some argue that "affect"' should be distinguished from "attitude"', some speech researchers have found that the narrow sense of "affect" is too restrictive because it excludes states in speech where emotion is present but not full-blown, including arousal and attitude (Cowie and Cornelius, 2003). Some tutoring researchers have also found it useful to take a combined view of affect and attitude (Bhatt et al., 2004).

Riley and Litman, 2005) show that student certainty interacts with a human tutor's dialogue decision process (i.e. the choice of feedback).

As a first step in detecting and adapting to student uncertainty it is important to understand where and why uncertainty occurs in a dialogue. Our intuition is that uncertainty is not uniformly distributed across the dialogue but occurs more than frequently depending on the position in the dialogue. We use discourse structure transitions to define the notion of "position in the dialogue".

Previous work has highlighted the impact of **speech recognition problems** (**SRP -** 2.3.1.5) on the dialogue flow. In reaction to system misrecognitions, users try to correct the system by employing strategies that work in human-human interactions. They tend to correct the system by switching to a prosodically marked speaking style (Levow, 1998) in many cases consistent with hyperarticulated speech (Swerts et al., 2000). Since most recognizers are not trained on this type of speech (Soltau and Waibel, 2000), these attempts lead to further errors in communication (Levow, 1998; Swerts et al., 2000). The resulting "chaining effect" of recognition problems can affect the user emotional state; a frustrated and irritated user will lead to further recognition problems (Boozer et al., 2003). Ultimately, the number of recognition problems is negatively correlated with the overall user satisfaction (Walker et al., 2001).

Given the negative impact of SRP, there has been a lot of work in trying to understand this phenomenon through predictive models (Gabsdil and Lemon, 2004; Hirschberg et al., 2004; Litman et al., 2000; Walker et al., 2000). Acoustic, prosodic and lexical features are commonly used in these models. However, usage of the discourse structure information is limited to local features (e.g. dialogue act sequencing information (Gabsdil and Lemon, 2004)) or flattens the discourse structure (e.g. the number of confirmation subdialogues (Walker et al., 2000)). We investigate if discourse structure transition information can be used to characterize SRP.

The main question behind this investigation is: "Are there places in the dialogue prone to more SRP?". While it is commonly believed that the answer is "yes", the main obstacles in answering this question are defining what "places in the dialogue" means and finding those problematic "places". We propose using the discourse structure transition information to define the notion of "places in the dialogue", extending over previous work that ignores this information (Gabsdil and Lemon, 2004; Walker et al., 2000).

5.2 PROBLEM STATEMENT

Statement: We investigate the utility of the discourse structure information for characterizing two dialogue phenomena: user affect and speech recognition problems.

Hypothesis: Discourse structure information offer insights regarding to where and why dialogue phenomena occur.

Intuition: Dialogue phenomena are not uniformly distributed across a dialogue but occur more frequently depending on the "*position in the dialogue*" (Interaction intuition). Discourse structure transitions can be used to define "position in the dialogue"

Approach: We look at statistical dependencies between discourse structure transitions and dialogue phenomena. We use the Chi Square (χ^2) test to mine dependencies in a corpus of dialogues. Analyses of the dependencies tell us where dialogue phenomena occur more/less than expected and allow us to formulate hypotheses behind this increase/decrease.

Results: We find dependencies between discourse structure transitions and the two dialogue phenomena. Several transitions in the discourse structure are associated with an increase or decrease of uncertainty and have specific interaction patterns with speech recognition problems.

A summary of our analyses, results and other work items is available below in Table 19. For each item we show the section where more details are available, a brief description, the statistical confidence in the outcome and whether it is a positive or negative outcome.

Table 19. Summary of results: characterization of dialogue phenomena

Result	Confidence	+/-
(5.4) Characterization of user affect (uncertainty)		
Statistical dependencies between discourse structure transitions and uncertainty	**significant**	**+**
Statistical dependencies between discourse structure transitions and uncertainty even after we discount for user correctness		
• Dependencies exist for correct user turns	**significant**	**+**
• Dependencies exist for incorrect user turns	**significant**	**+**
(5.5) Characterization of Speech Recognition problems		
Statistical dependencies between discourse structure transitions and ASR misrecognitions	**significant**	**+**
Statistical dependencies between discourse structure transitions and Rejections	**significant**	**+**
Statistical dependencies between discourse structure transitions and Semantic Misrecognitions	*trend*	**+**

5.3 METHOD

We mine statistical dependencies between discourse structure transitions and dialogue phenomena in the F03 corpus (2.2.1). We use the Chi Square (χ^2) test to measure the dependency strength. To apply the test, we define variables for discourse structure transition, user affect and speech recognition problems.

For discourse structure transition information we define the variable **TRANS** with six values corresponding to each type of transition (*Advance*, *NewTopLevel*, *PopUp*, *PopUpAdv*, *Push* and *SameGoal*). For uncertainty, we define the **UNCERT** variable with two values[27]: *Uncert* (uncertain) and *Other* (certain, neutral and mixed collapsed together – 2.3.2.2). We define one variable for each type of SRP described in Section 2.3.1.5 (Rejections – **REJ**, ASR Misrecognitions – **ASRMIS**, Semantic Misrecognitions - **SEMMIS**). REJ variable has two values: *Rej* (a rejection occurred in the turn) and *noRej* (no rejection occurred in the turn). The **ASRMIS** variable also has two values: *AsrMis* (difference between the human transcript and the system transcript) and *noAsrMis*. Similarly, the SEMMIS variable

[27] We collapse the other 3 uncertainty labels (certain, mixed and neutral) because of interest for affect adaptation is the presence of uncertainty.

has two values: *SemMis* (difference between the correctness interpretation of the system transcript (ASEM) and the correctness interpretation of the human transcript (TSEM)) and *noSemMis*. Table 20 shows the distribution of the variables in the F03 corpus.

Table 20. TRANS, UNCERT and SRP variable distribution in F03 corpus

Variable	Values	Distribution
Discourse structure transition		
TRANS	Advance	53.4%
	NewTopLevel	13.5%
	PopUp	9.2%
	PopUpAdv	3.5%
	Push	14.5%
	SameGoal	5.9%
User affect – certainty		
UNCERT	Uncert	19.1%
	Other	79.9%
Speech recognition problems		
ASRMIS	AsrMis	25.4%
	noAsrMis	74.6%
SEMMIS	SemMis	5.7%
	noSemMis	94.3%
REJ	Rej	7.0%
	noRej	93.0%

To discover the dependencies between our variables, we apply the χ^2 test. We explain the test and the interpretation of the test results on the interaction between TRANS and UNCERT (Table 21). χ^2 is a non-parametric test of the statistical significance of the relationship between two variables. The χ^2 value assesses whether the differences between observed and expected counts are large enough to conclude a statistically significant dependency between the two variables. The observed counts are computed from the data. The expected counts are the counts that would be expected if there were no relationship at all between the two variables. χ^2 value would be 0 if observed and expected counts were equal. To account for a given table's degree of freedom and one's chosen probability of exceeding any sampling error, the χ^2 value has to be larger than the critical χ^2 value. When looking at the TRANS–UNCERT interaction, which has five degrees of freedom ((6-1)*(2-1)), the critical χ^2 value at a $p<0.05$ is 11.07. Our χ^2 value of 30.71 exceeds this critical value and the interaction is significant at $p<0.00001$ (Table 21 row 1 – note that we will use only two decimals for the p values in the tables). We thus conclude that there is a statistically significant dependency between the discourse structure transition and the uncertainty in the following student answer. In other words, knowledge of the current discourse structure transition influences the distribution of certainty we see in the following user answer.

Table 21. TRANS–UNCERT interaction on all student turns

Combination		Obs.	Exp.	χ^2	p
TRANS – UNCERT				**30.71**	**0.00**
Advance – Uncert	-	216	237	**5.22**	**0.03**
PopUp – Uncert	-	26	40	**7.44**	**0.01**
PopUpAdv – Uncert	+	26	15	**8.81**	**0.00**
Push – Uncert	+	90	64	**14.71**	**0.00**

To better understand how two variables interact, we can look more deeply into this overall interaction by investigating how particular variable values interact with each other. For our TRANS-UNCERT interaction, we compute a binary variable for each value of TRANS and study dependencies between these variables and UNCERT. For example, for the value 'Advance' of variable TRANS we create a binary variable with two values: 'Advance' and 'Anything Else' (the other five transition labels). By studying the dependency between these binary variables we can understand how the interaction works.

Table 21 reports in rows 2-5 all *significant* interactions between the values of variables TRANS and UNCERT. Each row shows: 1) the value for each original variable, 2) the sign of the dependency, 3) the observed counts, 4) the expected counts, 5) the χ^2 value and 6) the level of significance of the interaction (p value). For example, in our data there are 26 uncertain turns after a PopUp transition. This value is smaller than the expected counts (40); the dependency between Advance and Uncert is significant with a χ^2 value of 7.44 and $p<0.007$. A comparison of the observed counts and expected counts reveals the direction (sign) of the dependency. In our case we see that after a PopUp transition student answer is uncertain *less* than expected; consequently, it means that all other types of certainty (i.e. neutral, mixed and certain) occur more than expected after such transition (not shown in the table). On the other hand, there is no significant interaction between NewTopLevel and uncertainty (thus, we do not show it in the table).

5.4 CHARACTERIZATION OF USER AFFECT

To investigate if discourse structure information can be used to characterize uncertainty[28] (our choice of user affect - 5.1), we look at the dependencies between TRANS and UNCERT. We have already presented and partially discussed the interaction results in Table 21 from Section 5.3.

[28] Publications: A similar investigation but with a different focus is reported in (Forbes-Riley et al., 2007).

The interactions that we find in Table 21 highlight connections between discourse structure transitions and student uncertainty and allow us to formulate hypotheses behind these connections. We find that after an Advance transition there are fewer than expected uncertain student answers. An Advance transition captures cases where the student has answered correctly the previous question or where the student answered incorrectly but the question was simple enough to be remediated without a remediation subdialogue. The interaction indicates that such cases are followed by a decrease in uncertainty. We hypothesize that in such cases because the student knew how to answer the previous system question or in case she was incorrect she understood the correct answer from the tutor explanation, the student *thinks* she knows the answer to the current question thus exhibiting less uncertainty. A similar behavior can be observed after a PopUp transition. Such transitions occur when the system has finished the remediation dialogue and is asking the original question again. We find that such situations are followed by less uncertainty than expected. We hypothesize that in such cases the student *thinks* she understood the information from the remediation dialogue and knows how to answer the original question.

In contrast, the PopUpAdv and Push transitions are followed by more uncertain student answers than expected. A PopUpAdv transition occurs after the system has finished with the remediation dialogue and moves to the next question without asking again the question that triggered the remediation dialogue. In such cases, we find more uncertain student answers than expected. We hypothesize that this interaction may be related to students loosing track of the original question and the connection between the current question and the previous instruction. Making explicit these transitions by showing how a subtopic fits in the larger topic may help reduce the amount of student uncertainty. Our graphical representation of the discourse structure, the Navigation Map (3.4), is one way of making these connections explicit. A Push transition occurs after the user has given an incorrect answer which triggers a remediation dialogue and the system asks the first question in the remediation dialogue. We find that in such cases there are more uncertain student turns than expected. We hypothesize this is because Push transitions correspond to deeper student knowledge gaps about the basic topics in the problem solution. The increased uncertainty after Pushes may also be related to the perceived lack of cohesion between the subtopic (i.e. the remediation dialogue) and the larger topic. As with the PopUpAdv transition, making explicit the connection between the subtopic and the larger topic can have positive results.

Because student correctness and student uncertainty are intertwined, we wanted to investigate if discourse structure transitions can still be used to characterize uncertainty if we discount the correctness. Indeed, our data shows that there is a highly significant interaction between a binary version of system correctness (ASEM: Correct vs. Others, recall Section 2.3.1.4) and uncertainty (UNCERT: Uncertain vs. Others): χ^2 value of 121.23 with $p<10^{-27}$; more incorrect answer are uncertain than expected (295 observed versus 191 expected). We use system correctness because this information is available at runtime for an online prediction of uncertainty.

To discount correctness, we rerun the interaction analysis on two complementary subsets: only correct answers (57% of all student turns) and only incorrect answers (43% of all student turns). Table 22 shows the TRANS–UNCERT interaction on correct student answers only. We find that the strength of the interaction has decreased (χ^2 goes down from 30.71 to 18.79) and that two of the value interaction are not significant anymore (Advance and PopUpAdv). The other two value interactions have the same sign and slightly reduced significance. However, the Push–Uncert interaction has an interesting implication in this case. According to this interaction, correct student answers after a Push transition are more uncertain than expected. The increased uncertainty after Pushes even for correct answer is likely related to the perceived lack of cohesion between the subtopic (i.e. the remediation dialogue) and the larger topic.

Table 22. TRANS–UNCERT interaction on **correct** student turns only

Combination		Obs.	Exp.	χ^2	p
TRANS – UNCERT				**18.79**	**0.00**
PopUp – Uncert	**-**	5	12	**5.31**	**0.03**
Push – Uncert	**+**	33	19	**13.16**	**0.00**

Table 23 shows the TRANS–UNCERT interaction on incorrect student answers only. The overall interaction has increased in significance (χ^2 increases from 30.71 to 44.04) and two new value interaction become significant while the Push–Uncert interaction is not significant anymore. The lack of significance for the Push–Uncert interaction for incorrect turns coupled with a significant interaction in case of correct turns further supports our hypothesis that Pushes are related to a perceived lack of cohesion between the subtopic and the larger topic. The Advance–Uncert and PopUp–Uncert interaction which we observed when we looked at all turns have interesting interpretations for incorrect student turns. We find that for incorrect answers we see a decrease of uncertainty after Advance and PopUp transitions. We hypothesize that the reduced uncertainty is explained by the student failing to realize his answer is incorrect. This probably happens because the student can not make the connection between the current question and the previous question at the same level (the Advance transition) or the connection between the remediation dialogue and the original question (the PopUp transition). Thus techniques that explicitly or implicitly make the connection between the tutor questions can help the student be more aware of the correctness of her answers. Our graphical representation of the discourse structure, the Navigation Map (3.4), is one such technique. The other interaction we observed before account for correctness (i.e. PopUpAdv–Uncert) has the same interpretation: a perceived lack of coherence between the remediation dialogue and the current question.

Table 23. TRANS–UNCERT interaction on **incorrect** student turns only

Combination		Obs.	Exp.	χ^2	p
TRANS – UNCERT				**44.04**	**0.00**
Advance – Uncert	**-**	131	147	**4.95**	**0.03**
NewTopLevel – Uncert	**+**	42	24	**20.58**	**0.00**
PopUp – Uncert	**-**	21	31	**5.2**	**0.03**
PopUpAdv – Uncert	**+**	24	13	**13.06**	**0.00**
SameGoal – Uncert	**-**	20	29	**4.68**	**0.04**

Two new significant interactions are found for incorrect turns: NewTopLevel–Uncert and SameGoal–Uncert. We find an increase in uncertainty after a NewTopLevel transition if the answer was incorrect. In such cases, the system starts a new dialogue based on essay analysis. We hypothesize that these are cases of the system discovering deep student knowledge gaps thus the observed increase in uncertainty for incorrect answers after this transition. We also find that after a rejection (SameGoal transition) if the answer is incorrect there is a decrease in uncertainty. We hypothesize that for these cases frustration and hyperarticulation takes priority over uncertainty: our previous analysis on the same corpus shows that students exhibit an increase in frustration and hyperarticulation after rejections (Rotaru and Litman, 2005, 2006).

Summary

Our results indicate that the discourse structure information can be used to characterize user uncertainty over and above correctness. We find that specific transitions in the discourse structure are associated with an increase or decrease of uncertainty. If we discount for correctness, which is interacts significantly with uncertainty, we find additional interactions. These interactions allow us to formulate hypotheses that can explain the phenomenon. We hypothesize several interactions are explained by students failing to make the connection between the instruction so far and the current system question. Our graphical representation of the discourse structure, the Navigation Map (3.4), is one way of making these connections explicit. The observed interaction also suggest the discourse structure transition information as a useful feature for predicting uncertainty (Ai et al., 2006).

5.5 CHARACTERIZATION OF SPEECH RECOGNITION PROBLEMS

To investigate if discourse structure information can be used to characterize speech recognition problems[29] (**SRP**), we look at the dependencies between TRANS and the 3 SRP variables defined in Section 5.3 (ASRMIS, SEMMIS, REJ).

Table 24. TRANS–ASRMIS interaction

Combination		Obs.	Exp.	χ^2	p
TRANS – ASRMIS				**23.88**	**0.00**
NewTopLevel – AsrMis	**-**	61	79	**6.75**	**0.01**
PopUp – AsrMis	**+**	74	54	**10.56**	**0.00**
Push – AsrMis	**+**	106	85	**7.3**	**0.01**

We find that TRANS interacts with all three types of SRP: ASR MIS (Table 24), REJ (Table 25) and SEMMIS (Table 26). We find that the student answer to the first system question after an essay (NewTopLevel) have less AsrMis than expected. In contrast, going down (Push) or going up (PopUp) in the discourse structure is correlated with more AsrMis. One hypothesis is that while entering or exiting remediation subdialogues, students have emotional and correctness states that are correlated with more AsrMis (Rotaru and Litman, 2006). Another explanation is that students are more confused by Push and PopUp transitions since our system employs a minimal number of lexical markers and no prosodic markers to signal these transitions (Hirschberg and Nakatani, 1996). Our graphical representation of discourse structure, the Navigation Map (3.4), explicitly signals these transitions through graphical means. Interestingly, Push and PopUp interact with AsrMis but do not interact with Rej (see Table 25).

Table 25. TRANS–REJ interaction

Combination		Obs.	Exp.	χ^2	p
TRANS – REJ				**383.15**	**0.00**
Advance – Rej	**-**	45	87	**46.95**	**0.00**
NewTopLevel – Rej	**-**	12	21	**5.58**	**0.02**
SameGoal – Rej	**+**	66	9	**376.63**	**0.00**

In terms of rejections (Table 25), we find that starting a new tutoring dialogue (NewTopLevel) or advancing at the same level (Advance) in the discourse structure reduces the likelihood of a rejection. In contrast, if the system repeats the same goal (i.e. due to a previous rejection) then the subsequent student

[29] Publications: The work presented in this section was published in (Rotaru and Litman, 2006).

turn will be rejected more than expected. The SameGoal–Rej interaction is another way of looking at the rejection chaining effect we reported in our previous work (Rotaru and Litman, 2005): rejections in the previous turn are followed more than expected by rejections in the current turn. The new TRANS–REJ interaction refines this chaining effect by pointing out situations that will make rejections less likely: cases when the user is advancing without major problems in the dialogue (NewTopLevel and Advance). This observation provides additional support for the rejection handling strategy we proposed in (Rotaru and Litman, 2005, 2006) for our domain: do not reject but keep the conversation going. This strategy is on par with observations on human-human dialogues (Skantze, 2005).

Table 26. TRANS–SEMMIS interaction

Combination		Obs.	Exp.	χ^2	p
TRANS – SEMMIS				*9.3*	*0.10*
NewTopLevel – SemMis	-	11	18	*3.35*	*0.07*
PopUp – SemMis	+	18	12	*3.1*	*0.08*

The interaction between TRANS and SEMMIS is weaker (only a trend, Table 26) but offers additional insights. We find a decrease in the number of SemMis when starting a new dialogue (a NewTopLevel transition). Not only are there fewer AsrMis after a NewTopLevel transition (recall NewTopLevel–AsrMis in Table 24) but if they happen they are less likely to cause problems in terms of interpretation. In Section 4.4.2 we proposed a modification of the system based on the NewTopLevel–Incorrect bigram: disable the current essay interpretation component and try all authored essay update dialogues; for each dialogue, based on the correctness of the first student answer either continue with the rest of the dialogue (incorrect answer) or skip the rest of the dialogue (correct answer). However, that analysis was based on the human correctness. The NewTopLevel–SemMis interaction suggest that there are fewer problems due to recognition errors in the first student answer, thus the modification is likely to work even with the much noisier system correctness.

In contrast, after a remediation dialogue (PopUp transition) we see an increase of SemMis. In other words, when trying to answer the original question after the remediation dialogue, the student answer has more correctness interpretation problems due to recognition errors than expected. We hypothesize that this happens because the student lacks the appropriate technical language that the system expects at that point; even if these technical terms show up in the remediation dialogue, the student has problems reusing them. This suggest that making visible the correct answers and the language used in the remediation dialogues can reduce the semantic problems caused by SRP, at least after a PopUp transition. The graphical representation of the discourse structure we proposed in Section 3.4 was designed to include this information.

Summary

Our results indicate that the discourse structure information can be used to characterize SRP. We find that certain discourse structure transitions have specific interaction patterns with SRP (e.g. Push and PopUp transitions have problematic interactions with AsrMis). These findings suggest that discourse structure transitions can be an informative feature for predictive models of SRP.

5.6 RELATED WORK

We discuss below previous work that looks at the characterization of the two dialogue phenomena we explored: user affect and speech recognition problems.

Affective computing (Picard, 2003) is a relatively new research direction that investigates computer systems that detect, react to and/or exhibit emotions. Central to affective computing is the fact that human conversational partners detect and respond to the speaker's or listener's emotional state and that humans extend this behavior and their expectations to interactions with non-human entities like computers, media, television, etc. (Reeves and Nass, 1996). As a result, detecting and adapting to user affect is currently being pursued by many researchers as a method of improving the quality of spoken dialogue systems (Batliner et al., 2003; Lee et al., 2002). The main idea is that SDS should react not only to what users say but also to *how* they speak. This direction has also received a lot of attention in the tutoring domain where it is hypothesized that human tutors respond to student affective states and similar capabilities are explored for computer tutors (Aist et al., 2002; Craig et al., 2004; Forbes-Riley and Litman, 2005; Forbes-Riley et al., 2008; Pon-Barry et al., 2006).

Leaving aside issues like what affective states are and how to represent them (Cowie and Cornelius, 2003), one of the first steps in affective computing is to detect user affect. Characterizing user affect is an important tool for this task as it indicates how affective speech differs from normal speech. Previous studies have found specific acoustic-prosodic correlates for user affect (see (Scherer, 2003) for a review). As a result, many studies in automatic prediction of user affect have employed a variety of acoustic-prosodic features and other types of context-independent features: pitch features (e.g. mean, max, slope), amplitude features, tempo features (e.g. speaking rate, amount of silence), duration features, spectral features, lexical features (e.g. words, part-of-speech), identification features, etc (Ang et al., 2002; Batliner et al., 2003; Lee et al., 2002; Litman and Forbes-Riley, 2006). Other non-speech features like facial expressions and posture patterns have also been explored (D'Mello et al., 2005; Swerts and Krahmer, 2005).

The context in which the affect occurs has also been used to produce features for affect prediction. Several context-dependent features have been explored: number of turns in the dialogue, the length of the dialogue, number of user corrections/repetitions, dialogue acts (Ai et al., 2006; Ang et al., 2002; Batliner et al., 2003). However, most context-dependent parameters do not take into account the discourse structure information. The dialogue act parameters used in (Batliner et al., 2003) exploit the discourse information but ignore the hierarchical aspect.

We extend over previous work by exploiting the hierarchical aspect of the discourse structure to characterize user affect. We exploit the discourse segment hierarchy through our six discourse structure transitions (3.3). Our results show strong interactions between discourse structure transitions and user affect (uncertainty in our case) validating our intuition that affect does not occur uniformly across the dialogue. As a result, (Ai et al., 2006) use our discourse structure transitions as features in their affect prediction experiments; however, their work does not investigate directly the contribution of these features. Similar to the dialogue act features (Batliner et al., 2003), discourse structure transitions are domain independent and can be easily applied in other domains.

Speech recognition problems (**SRP**) occur when the automated speech recognition component of a SDS fails to produce the correct recognition of the user turn. Several types of SRP were described in Section 2.3.1.5. Previous work has highlighted the impact of SRP on various dialogue phenomena. In reaction to system misrecognitions, users try to correct the system by employing strategies that work in human-human interactions. They tend to correct the system by switching to a prosodically marked speaking style (Levow, 1998) in many cases consistent with hyperarticulated speech (Swerts et al., 2000). Since most recognizers are not trained on this type of speech (Soltau and Waibel, 2000), these attempts lead to further errors in communication (Levow, 1998; Swerts et al., 2000). The resulting "chaining effect" of recognition problems can affect the user emotional state; a frustrated and irritated user will lead to further recognition problems (Boozer et al., 2003). Ultimately, the number of recognition problems is negatively correlated with the overall user satisfaction (Walker et al., 2001).

Given the negative impact of SRP, there has been a lot of work in trying to understand this phenomenon through predictive models (Gabsdil and Lemon, 2004; Hirschberg et al., 2004; Walker et al., 2000) and in terms of strategies for handling SRP (Bohus and Rudnicky, 2005). Acoustic, prosodic and lexical features are commonly used in these models. However, usage of the discourse structure information is limited to local features (e.g. dialogue act sequencing information (Gabsdil and Lemon, 2004)) or flattens the discourse structure (e.g. the number of confirmation subdialogues (Walker et al., 2000)).

We extend over previous work by exploiting the hierarchical aspect of the discourse structure to characterize SRP. We exploit the discourse segment hierarchy through our six discourse structure transitions (3.3). Our results find several significant and trend interactions between discourse structure transitions and SRP. These findings identify problematic transitions (e.g. Push, PopUp) in the dialogue in terms of SRP and allow us to formulate hypotheses to address increases of SRP after certain transitions (e.g. PopUp). In terms of investigating which tutor states lead to more SRP, discourse structure transitions allow us to deal with the data sparsity problem by providing a level of abstraction over individual system states.

5.7 SUMMARY & FUTURE WORK

We find that discourse structure information can be used to characterize two dialogue phenomena: user affect and speech recognition problems. We use the transition information (introduced in Section 3.3) that is part of discourse structure and we find statistical dependencies between transitions and dialogue phenomena in the F03 corpus. Several transitions in the discourse structure are associated with an increase or decrease of uncertainty and have specific interaction patterns with speech recognition problems. The interactions we find allow us to formulate hypotheses behind the presence/absence of the phenomena, which in turn suggest ways to modify our system. At the very least, these dependencies suggest that discourse structure transition information can be used as an informative feature for predicting dialogue phenomena.

The main contribution of discourse structure transitions for this application was to provide a level of abstraction over individual system states. Our system has 254 unique states (i.e. system questions) but there are only 6 transitions. Ideally, we would like to see how each system state is associated with dialogue phenomena. However, given the relatively small size of our corpora (e.g. F03 has 2964 system turns), this type of analysis will result in data sparsity issues and overfitting (e.g. we have to repeat the analysis for each system state). In contrast, by providing a level of abstraction over individual system states, the discourse structure transitions allow us to perform a meaningful analysis on our corpus with interesting results. Furthermore, an in-depth analysis of the interactions indicates that the observed behavior is attributable to a set of system states as a whole rather than to specific system states. For each significant interaction, the number of unique system states involved in the interaction is between 15 and 47 with no system state from this set being repeated in our corpus more than 10-15 times.

From the dialogue designer perspective, the dependencies we observe in our corpus offer insights into system behavior. The transition–SRP interactions suggest that particular attention should be paid to specific locations in the discourse structure. For example, for our system, the interactions between

Push/PopUp and ASRMIS suggest that increasing student awareness of the discourse structure through lexical and prosodic means (Hirschberg and Nakatani, 1996) might be beneficial (i.e. Push and PopUp transition signal crossing of discourse segment boundaries). Also, the PopUp–SemMis interaction suggests that grounding the technical terms used and the correct answers might reduce the number of semantic errors caused by SRP. Our graphical representation of the discourse structure, the Navigation Map (3.4), explicitly signals our transitions through graphical means and displays the technical terms and the correct answers; consequently, it will be interesting to see if this feature will affect the TRANS–SRP interaction patterns.

The dependencies between transitions and user affect offer insights into what contexts are associated with increase/decrease in user affect. Knowledge of these contexts is an important step towards context-dependent affect adaptation. This knowledge allows us to formulate hypotheses to explain these dependencies which in turn can be used to design appropriate system responses. In fact, this direction is currently being pursued by other colleagues (e.g. (Forbes-Riley et al., 2007)).

Our results also suggest that discourse structure transitions might be an informative feature for prediction of dialogue phenomena. Previous work by other colleagues (Ai et al., 2006) has already used this information source to predict user affect however the study does not explicitly look at the contribution of the discourse transitions. Empirical studies can be run to test if indeed discourse transition features are informative for predicting dialogue phenomena.

It is also interesting to note that this application of discourse structure makes use of discourse structure transitions in a conceptually different way than the performance analysis application (Section 4). For performance analysis, we use transitions in collaboration with dialogue phenomena to model system performance. For characterization of dialogue phenomena, we use transitions to model dialogue phenomena by looking at statistical dependencies between them.

6 APPLICATIONS FOR GUI: THE NAVIGATION MAP

Acknowledgments: This line of research was inspired by an earlier unpublished work performed during author's 2004 summer internship at IBM T.J. Watson Research Center under the supervision of Shimei Pan.

We explore the applications of discourse structure for the graphical output of a SDS. We first motivate why presenting discourse structure information graphically might have a positive impact (6.1). Our graphical representation of discourse structure, the Navigation Map, was described in Section 3.4. We formulate our research problem (6.2) and describe our findings on the subjective (6.3) and objective (6.4) utility of the Navigation Map. We discuss related work (6.5) and then we summarize our findings and present avenues for future work (6.6).

6.1 INTRODUCTION

There are several observations and previous findings that motivated us to investigate a graphical representation of discourse structure: characteristics of the interaction with SDS in complex domains, findings from previous work and our results from previous sections. We will discuss them in detail here.

As mentioned in the introduction (Section 1), complex-domain SDS are characterized by increased task complexity. In addition, two other key properties are present: user's lack of or limited task knowledge and longer system turns. Take for example the tutoring domain. User's lack of or limited knowledge of the tutoring topic is a general artifact of the domain: it is the purpose of a tutoring SDS to convey this knowledge to the user. Longer system turns are also common in the tutoring domain as the system has to discuss concepts, laws and relationships and to engage in complex subdialogues to correct user misconceptions. $Tutor_1$ in Figure 14 is a good example of a long system turn in our system[30].

These properties manifest in an increased effort on the user side while interacting with these systems. Because users have limited knowledge or no knowledge about the domain, in these systems the information flows primarily from the system to the user. Coupled with the complexity of the underlying task, users are in a situation where they have to assimilate a lot of new information conveyed by the system. The speech-only interaction commonly used in SDS puts additional burdens on the user's working memory as users have no other alternatives to store and access the information discussed so far. Listening to long system turns and processing new information concurrently can also prove to be challenging for users.

One method to reduce the effort on the user side is to use a complementary communication channel: the visual channel. This channel can be used to provide additional support information and as a repository for information that can be easily accessed by users. Previous studies have shown that if both modalities are offered, users know how to choose the appropriate modality (e.g. (Oviatt et al., 2004)). From a learning perspective (learning is the main performance metric in our system), combining the audio and visual channel has been shown to be helpful as the working memory doubles since each modality channel has its own independent processor/memory (Mousavi et al., 1995). However, the best way to combine them is still an open question due to issues like the split-attention effect (e.g. (Sweller et al., 1998)).

But what to communicate via the visual channel? The ITSPOKE interface has a dialogue history text box which displays all the system and user turns so far. However, we hypothesize that displaying two other pieces of information will be more beneficial: the purpose of the current topic and how it relates to the overall discussion. The purpose of the current topic provides a digested view of the local conversation and can help with user's listening of long system turns. The relationship between the current topic and the overall discussion can help users manage the complexity of the task.

The information described in the previous paragraph is implicitly encoded in the intentional structure that is part of discourse structure. Consequently, *we propose using a graphical representation of*

---

[30] While some long tutor turns in ITSPOKE might be an artifact of the Why2-Atlas backend which was designed for text based interaction, long tutor turns are common in human-human tutoring too. In a parallel human-human study which used the same graphical interface as ITSPOKE as well as speech communication, the human tutor spoke 26 words per turn on average (ITSPOKE words per turn average is 43). Here is an example of the speech transcript of a long human tutor turn: "and also on the force that is being applied now uh this statement in the last sentence here uh of the problem which says that both uh mm are uh they are in both cases it is in front of the stop sign and it is in first gear all that is trying to tell you is that the force applied remains the same in both cases".

the discourse structure as a way of improving the performance of complex-domain SDS. We will use our graphical representation of discourse structure (the Navigation Map) described in Section 3.4.

Previous tutoring research provides additional support. A number of previous studies have shown that classroom instruction that includes a graphical representation of information is beneficial (e.g. graphical organizers (Marzano et al., 2000)). In addition, some educational psychology studies argue for a guided instruction instead of a unguided or minimally guided instruction though there is no consensus in the field regarding this issue (e.g. (Kirschner et al., 2006)). A graphical representation of discourse structure will act as an implicit guide for the information being tutored. It is similar in spirit to process worksheets described in (Nadolski et al., 2005). However all these studies are done in non-computer tutoring settings; we are investigating if the same holds in the case of speech-based computer tutors for a graphical representation of the discourse structure.

Our analysis of the applications of the discourse structure on the system side provides additional support for this proposed work. In Section 5.4, we found an increase in uncertainty after Push and PopUpAdv transitions. We hypothesized that this observation is due to a perceived lack of coherence between the current system question and the discussion so far. A graphical representation of the discourse structure can be beneficial in this case as it implicitly encodes how instruction topics relate to each other. We also found a decrease in uncertainty for *incorrect* answers after an Advance and PopUp transition. We hypothesized that these events are linked to users who fail to realize how the instruction so far (the previous question at the same level for Advance and the remediation dialogue for PopUp) contradicts their incorrect answer. Also, in Section 4.4.1 we found that incorrect answers after key transitions (PopUp and PopUpAdv) are associated with less learning. Consequently, improving the correctness of the student answers after these transitions might lead to increased learning. These findings suggest that a graphical representation of the discourse structure can be beneficial in such situations as students will have direct access to a summary of the instruction while making the inference and, thus, have more chances to realize they are incorrect (i.e. exhibit more uncertainty) or even provide the correct answer. Moreover, in Section 5.5 we find that certain transitions have problematic interactions with speech recognition problems and we hypothesized that a graphical representation of the discourse structure might be beneficial as it makes clear the transitions through means other than lexical and prosodic.

6.2 PROBLEM STATEMENT

Statement: We investigate the utility of integrating the discourse structure information in the graphical output of spoken dialogue systems. This integration is achieved through a graphical representation of discourse structure which we call the Navigation Map (**NM** – see Section 3.4). In effect, we are investigating the utility of the NM for spoken dialogue systems.

Hypothesis: The NM improves the perceived quality and the performance of SDS in complex domains.

Intuition: Users can follow the conversation easier with the NM as it provides a digested view of the conversation and of the relationships between various conversation topics.

Approach: Two directions are pursued: an investigation of the subjective utility (6.3) and an investigation of the objective utility (6.4). For the subjective utility, we hypothesize that users will rate a NM version of the system better than a non NM version of the system. For the objective utility, we hypothesize that we will see an improvement on a number of objective metrics (e.g. system performance, dialogue efficiency) for users which have the NM over users that do not have it.

Results: The NM has a positive subjective and objective utility. We find that users prefer the NM version and rate it better on a number of dimensions (e.g. ability to identify and follow tutoring plan, integrate instruction, concentration, etc.). Users' preference for the NM is also reflected in a number of objective metrics (e.g. increased system performance for certain users but decreased performance for others and shorter dialogues).

Details

As a first step towards understanding the utility of the NM, we look at whether users like having the NM (subjective utility). If the intuition presented above is true than if we expose users to both a version with the NM and a version without the NM, we expect users to "like" more the version with the NM. The rationale for this is that if indeed users follow the conversation easier with the NM, then they will prefer the version of the system that requires the least effort on their part. To measure if users "like" the NM, we ask them to rate the two versions of the system on a variety of dimensions. Our results indicate that this is true and users rate the NM version better on a number of dimensions.

The positive outcome from the subjective utility investigation motivated us to pursue the objective utility of the NM. To measure the objective utility, we compare users that have the NM with users that do not have the NM on a number of objective metrics (e.g. system performance, dialogue efficiency metrics). If the intuition presented above is true, then it should be easier for NM users to integrate the instruction as they spend less effort following the conversation and can focus more on understanding and integrating the information. As a result, they will give more correct answers resulting in more efficient dialogues (dialogue is driven by user correctness). Overall, they will learn better resulting in better system performance. Our results indicate that this is true and we see an increase in dialogue efficiency (e.g. dialogue time) and system performance (i.e. learning - but only for certain users).

Note that, since in our first NM experiment (NMPrelim – subjective utility) we were still exploring the utility of the NM, the NM information highlight feature was enabled in that study. As discussed in Section 3.4, it is arguable whether this feature relates to discourse structure or not. To eliminate the impact of this issue on our claims about the utility of discourse structure, in our second NM experiment (Main – objective utility) the information highlight feature was disabled. Even with information highlight disabled, the NM has a positive impact on objective and subjective metrics.

A summary of our analyses, results and other work items is available below in Table 27. For each item we show the section where more details are available, a brief description, the statistical confidence in the outcome and whether it is a positive or negative outcome.

Table 27. Summary of results: the Navigation Map

Result	**Confidence**	**+/-**
(3.4) Implementation of the Navigation Map for ITSPOKE		**+**
(6.3) Subjective utility		
(6.3.1) Run a within-subjects user study		**+**
(6.3.2) Comparison with a control condition with no graphical support (NM vs noNM)		
(6.3.2.1) Presence of the NM has a positive effect on how users rate several questions in the following categories:		
• Overall impression (easy to learn and concentrate, expectation, reuse)	**trend/significant**	**+**
• During the conversation (identify and follow the tutoring plan, integrate the instruction)	**trend/significant**	**+**
• After the conversation (understand tutor's main point)	*trend*	**+**
(6.4) Objective utility		
(6.4.1) Run a between-subjects user study		**+**
(6.4.2) Comparison with a control condition with graphical support (*NM* vs. *R*)		
• (6.4.2.1) System performance: very small decrease (- 0.06 NLG effect size)	non-significant	**-**
• (6.4.2.2) Aptitude-treatment interaction: pretest split (PRE Split)		
• PI Split x Condition: interaction with learning	*trend*	
• + 0.46 NLG effect size for high pretesters	non-significant	**+**
• - 0.66 NLG effect size for low pretesters	non-significant	**-**
RELNLG subset (2.3.3.2)		
• PI Split x Condition: interaction with learning	**significant**	
• + 0.85 NLG effect size for high pretesters	**significant**	**+**
• - 0.66 NLG effect size for low pretesters	non-significant	**-**
• (6.4.2.5) Dialogue time: reduction in dialogue time	**significant**	**+**
• (6.4.2.6) Number of system turns: reduction in the number of system turns	non-significant	**+**
• Subjective metrics:		
• (6.4.2.3) Positive effect on average rating in the Conversation category (Conv)	*trend*	**+**
• (6.4.2.4) Correlation between average user rating on every category (Overall, Conv, Essay and ALL) and learning for *NM* users but not for *R* users	**significant**	**+/-**

6.3 SUBJECTIVE UTILITY OF THE NAVIGATION MAP

Since we did not know if the Navigation Map (**NM**) will have any impact on users, our first study was focused on how users perceive the NM[31]. We first describe a user study designed to investigate this subjective utility (6.3.1). Next, we present our analyses on the collected corpus (6.3.2). Our results indicate that indeed the NM has a positive effect on users' perception. In addition, preliminary findings suggest a positive effect on several objective metrics.

6.3.1 User study

As a first step towards understanding the utility of the NM, we wanted to know if users prefer having the NM over not having it. Consequently, we designed and performed a within-subjects user study where users received instruction both with and without the NM. We will describe here the important design choices. For more details about the actual experiment and the collected corpus see Section 2.2.3.

Since the main focus of this user study is not the actual learning but users' perception, we used a downsized version of ITSPOKE with only 2 problems[32]. The NM was enabled in *only one* problem. The information highlight NM feature was also enabled in this experiment. After each problem users filled a **system questionnaire** (2.2.3.1) in which they rated the system on various dimensions; these ratings were specifically designed to cover dimensions the NM might affect. While the system questionnaire implicitly probed the NM utility, at the end of the experiment, users filled a **NM survey** (2.2.3.2) which explicitly asked the users whether the NM was useful and on what dimensions.

Note that in both problems, users did not have access to the dialogue transcript. The original ITSPOKE interface has a dialogue history text box which displays all system and user turns so far (2.1). We chose to disable the dialogue history box because, if the NM has any effect, we would be able to see it easier if the dialogue history text box is disabled than if it was enabled. In addition, previous work (Litman et al., 2004) shows that on the same ITSPOKE interface with a human wizard, the speech without transcript condition was better than the text chat condition with transcript. Nonetheless, we run another NM study (6.4.1) in which the dialogue history box was enabled in the control condition; that control condition can be seen as a "higher" baseline for the NM than the one used in this experiment.

---

[31] Publications: The work presented in this section was published in (Rotaru and Litman, 2007).

[32] Nonetheless, users still learned in our experiment: there is a significant difference between pretest and posttest scores (p<0.01, see Table 5).

To account for the effect of the tutored problem on the user's questionnaire ratings, users were randomly assigned to one of two conditions: *F* or *S*. 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. Figure 21 shows the ITSPOKE interface with the NM enabled (*F* students – 1st problem, *S* students – 2nd problem) while Figure 22 shows the ITSPOKE interface with the NM disabled (*F* students – 2nd problem, *S* students – 1st problem).

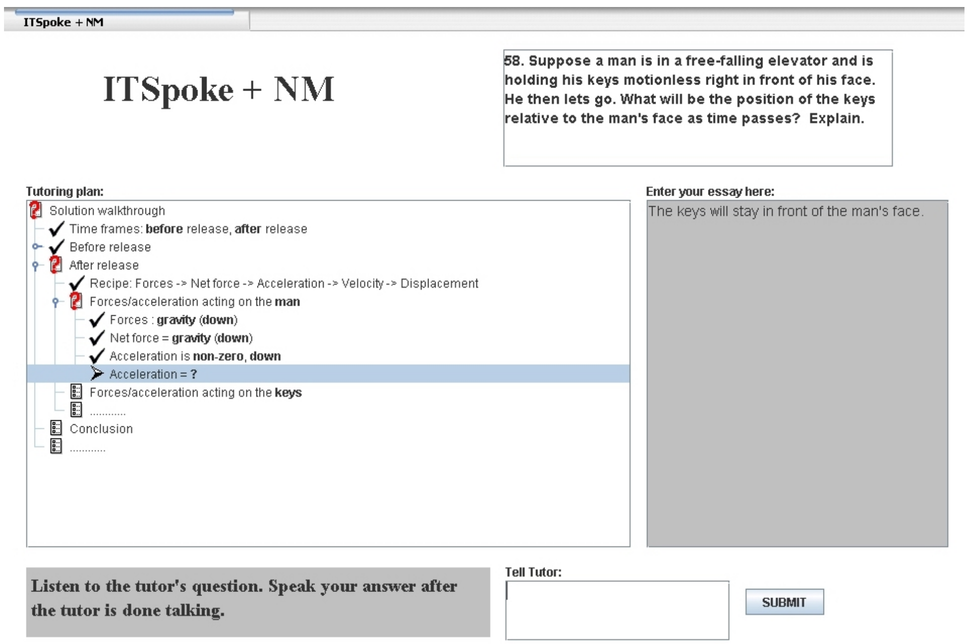


Figure 21. ITSPOKE interface with the NM enabled

There were two confounding factors we wanted to account for. To reduce the effect of the quality of the speech output, we used a version of the system with human prerecorded prompts. We also wanted to account for the amount of instruction as in our system the instruction is tailored to what users write in the essay. Thus the essay analysis component was disabled; for all users, the system went through the "walkthrough" instruction which assumed no information in the user 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.

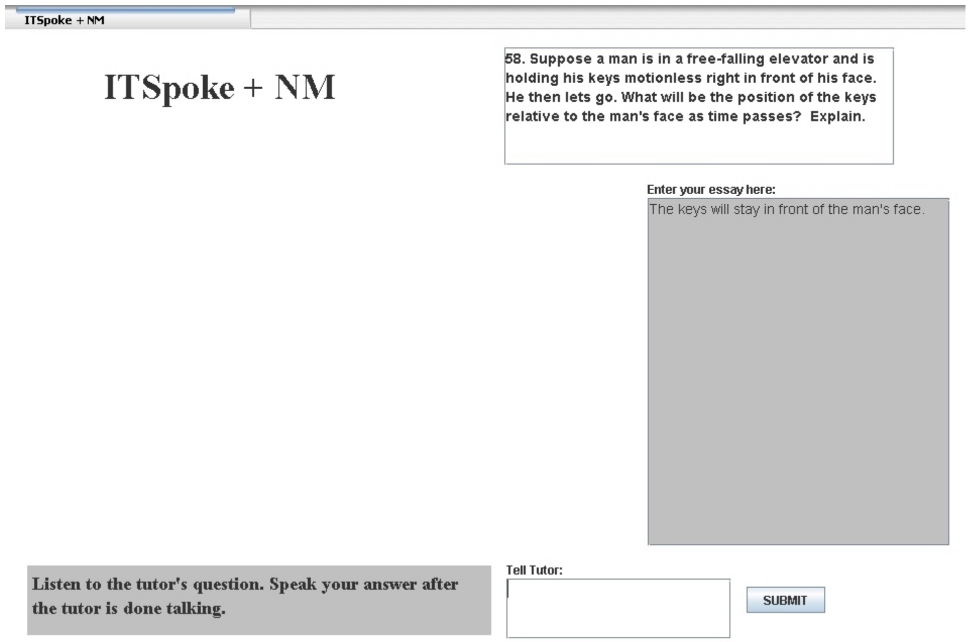


Figure 22. ITSPOKE interface with the NM disabled

6.3.2 Results

We focus primarily on the effect of the NM on a number of subjective metrics (6.3.2.1). We also run a number of preliminary analyses on objective metrics (6.3.2.2).

6.3.2.1 Subjective metrics

Our main resource for investigating the effect of the NM was the system questionnaires given after each problem (2.2.3.1). 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 on the NM problem and the **noNM** problem (i.e. the NM is disabled).

Table 28 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 28) 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.

Table 28. System questionnaire results

Question	ANOVA			Average rating: F condition			Average rating: 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

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**)[33]. The significance of the effect of each factor and their combination (NMPres×Cond) is listed in the table with significant/trend effects highlighted in bold/italics (see columns 2-4). Post-hoc t-tests between the NM and noNM ratings were run for each condition. Significant/trend differences are marked with an "s/t" after the comparison sign.

We discuss our results for each question group based on the connection with the NM. We first look at the Conversation group (Q7-Q13), then at the Post-conversation group (Q14-Q16) and finally users overall perception of the system (Q1-Q6).

[33] 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.

Results for the Conversation group (Q7-Q13)

Q7-Q13 relate directly to our hypothesis that users benefit from access to the discourse structure information. These questions probe user's perception of ITSPOKE during the dialogue. We find that for 6 out of 7 questions the NM presence has a significant/trend effect (Table 28, column 2). We discuss the results based on the dimension that the question(s) is investigating.

Structure. Users perceive the system as having a structured tutoring plan significantly[34] 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 trend that it is easier for users to know the correct answer if the NM is present (Q13). While not significant, in the NM condition users think it is easier to know if they were correct or not (Q12). We hypothesize that speech recognition and language understanding errors are responsible for the reduced NM effect on this dimension.

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 easier on what the system is uttering. In many of the open question interviews

---

[34] We refer to the significance of the NMPres factor (Table 28, column 2). When discussing individual experimental conditions, we refer to the post-hoc t-tests.

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 the Post-conversation group (Q14-Q16)

Questions Q14-16 were included to probe user's post-conversation 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.

Results for Overall group (Q1-Q6)

Questions Q1-Q6 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 also a trend that it was easier for them to learn from the NM enabled version of the system (Q2). Non-significant differences in the same direction exist in terms of user's enjoyment (Q5) and perceived learning (Q1).

Other results

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).

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 (2.2.3.2) 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 offer further support for the conclusions from the system questionnaire analysis.

Summary

The presence of the NM has a positive effect on how users rate our system. 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.

6.3.2.2 Objective metrics

The design of this experiment limited the range and scope of our analyses on objective metrics. This is because the effect of the NM can be reliably measured only in the first problem as in the second problem the NM is toggled. 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 the instruction in different forms (with the NM for *F* users, without the NM for *S* users); thus any differences in the second problem have to be attributed to the NM presence/absence in this problem as well as to the NM absence/presence in the previous problem.

As a result, we can only look at objective metrics that involve the first problem: dialogue time (P1Time - 2.3.3.3), number of system turns (P1Tut - 2.3.3.4), correctness (ASEM - 2.3.1.4) and speech recognition problems (2.3.1.5). Differences in learning (i.e. differences between pretest and posttest - 2.3.1.2) can not be investigated for the same reason.

Our preliminary investigation[35] found several dimensions on which the two conditions differed in the first problem (*F* users had NM, *S* users did not). We find that while there are no differences in terms of dialogue time, if the NM was present there was one tutor turn less in average and users gave more correct answers; however these differences are far from significance (Table 29). In terms of speech recognition performance, we looked at two metrics: AsrMis and SemMis (ASR Misrecognition and Semantic Misrecognition – 2.3.1.5). 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 shows 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. only 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 but to the NM presence. These results suggest

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

that using the NM might lead to fewer speech recognition problems (hypothesis which was suggested by our interaction analysis from Section 5.5).

Table 29. Objective metric results

(Average, standard deviation and significance for the two conditions – first problem only)

Metric	*F* (NM)	*S* (noNM)	p
Dialogue time (P1Time)	666 (180)	668 (177)	0.97
# tutor turns (P1Tut)	23.8 (5.3)	24.8 (6.5)	0.67
% correct turns (ASEM)	72% (18%)	67% (22%)	0.59
AsrMis	37% (27%)	46% (28%)	0.46
SemMis	5% (6%)	12% (14%)	*0.09*

Summary

Our limited scope analysis on objective metrics suggests that users' preference for the NM is reflected in more correct answers and fewer speech recognition problems in the NM version of the system though the differences are not significant.

6.4 OBJECTIVE UTILITY OF THE NAVIGATION MAP

In Section 6.3, we showed that the Navigation Map (**NM**) has a subjective utility: users rate the system better on a number of dimensions if the NM is present. Here, we investigate the objective utility of the NM: i.e. if there are effects on objective metrics (i.e. system performance, dialogue efficiency metrics etc.) associated with the presence of the NM. We first describe a user study designed to investigate the objective utility of the NM (6.4.1). Next, we present our analyses on the collected corpus (6.4.2). Our results indicate that indeed the NM has an effect on learning but it is dependent on user aptitude. In addition, we find a number of positive effects on other objective metrics.

6.4.1 User study

We designed and performed a between-subjects study with 2 conditions. In the control condition we used the regular version of ITSPOKE, which will be identified as ***R***. In the experimental condition, we used a version of ITSPOKE where we replaced the dialogue history box with the NM, version which will be identified as ***NM***. The information highlight NM feature was disabled as this information source is not part of discourse structure information. Also, note that the previous NM experiment (6.3.1) used a "weaker" baseline: in that experiment, the dialogue history was not present in the control condition. For that version, users could only listen to the system speech. In contrast, in *R* users can listen to system speech but also read this information from the screen.

There were two confounding factors we wanted to avoid. To reduce the effect of the quality of the speech output, we used a version of the system with human prerecorded prompts. We also wanted to account for the amount of instruction as in our system the instruction is tailored to what users write in the essay. Thus the essay analysis component was disabled; for all users, the system went through the "walkthrough" instruction which assumed no information in the user 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 actual experiment had an additional experimental condition (*PI*) that was investigating the utility of a new PopUp–Incorrect strategy (4.5.2) compared to *R*. Details for that part of the study are available in Section 4.5.3. The experiment was designed for two pairwise comparisons: *NM* vs. *R* and *PI* vs. *R* (i.e. *R* was shared). For more information about the actual experiment and the collected corpus, see Section 2.2.4.

6.4.2 Results

In the following sections we investigate differences between *R* and *NM*. We have already shown in Section 2.2.4.2 that the two conditions are balanced in terms of user abilities (PRE score). To investigate if the NM has any objective effects, we compare the two conditions in terms of system performance overall (6.4.2.1) and on specific subpopulations (6.4.2.2). Next we look at subjective metrics (6.4.2.3 and 6.4.2.4) and several dialogue efficiency metrics (6.4.2.5 and 6.4.2.6).

6.4.2.1 Performance (Learning)

One of our NM hypotheses is that users with access to the NM will perform better on a number of objective metrics. The most important metric on which we would like to see the NM impact is the system performance metric – that is, learning. As mentioned in Section 2.3.1.2, there are various ways of measuring learning, each giving a different perspective on the learning phenomena. We will present the results on all of them.

Table 30 shows the average and standard deviation for each condition on our 2 learning metrics (POST and NLG). On both metrics we find that *NM* users perform similarly but slightly worse than *R* users. We run a one-way ANOVA with POST and NLG as dependent variables and the Condition (*NM* vs. *R*) as the categorical factor. Results show no significant differences between the two conditions on any learning metrics (for POST: $F(1, 50)=0.10$, $p<0.76$; for NLG: $F(1,50)=0.05$, $p<0.82$). The corresponding NLG effect size is negative 0.06.

Table 30. Average and Standard Deviation for learning metrics

Condition	PRE	POST	NLG
NM	12.6 (4.1)	18.8 (3.9)	0.48 (0.21)
R	12.6 (4.2)	19.1 (3.1)	0.49 (0.20)

To test if students learn in the two conditions, we run a repeated-measure ANOVA with the Test Phase as the within-subjects factor (PRE vs. POST scores) and the Condition (*NM* vs. *R*) as the between-subjects factor. We find a significant effect from the Test Phase ($F(1,50)=199.24$, $p<0.0001$) and, as expected, no significant effect from the Condition or the combination Condition × Test Phase (for Condition: $F(1,50)=0.03$, $p<0.87$; for Condition × Test Phase: $F(1,50)=0.11$, $p<0.74$). Posthoc tests show a significant difference between the PRE and POST scores on each condition ($p<0.0001$). In other words, students learn significantly in both conditions but learning is not dependent on the condition.

As in many other tutoring studies, the pretest and posttest scores are positively correlated for the combined *R-NM* population ($R=0.649$, $p<0.001$). We wanted to know if this PRE-POST relationship differs between conditions. Thus, we run a one-way ANCOVA with POST as the dependent variable, PRE as the covariate and the Condition (*NM* vs. *R*) as the independent variable. As expected from the POST/NLG results, we find that the condition has no effect on the PRE-POST covariation ($F(1,49)=0.16$, $p<0.69$). A visual presentation of this result is available in Appendix B.2.1.

Restricting the analyses to the RELNLG subset (2.3.3.2) does not impact the results. The POST averages become the same (18.4) but there is a swap in the NLG average ranking (0.48 for *NM*, 0.46 for *R*). The ANOVA and ANCOVA effects are similar.

Summary

There are no differences between *NM* users and *R* users in terms of learning (main system performance metric). In both conditions, users learn significantly but their learning patterns do no differ between conditions.

6.4.2.2 Aptitude – Treatment interaction on Performance

It is a common analysis practice in tutoring research to investigate the relationship between user aptitudes and his/her performance while working with the system. Many studies have shown that the treatment condition can produce effects only on specific subsets of the populations and that, in some cases, the treatment has opposite effects depending on the subset. Here, we will use the initial physics knowledge aptitude (PRE score - 2.3.1.1) and the subsets created by the mean split on this aptitude (PRE Split - 2.3.3.1). Although there are no results for the other aptitude measured in this experiment (working memory span), for completeness we present that analysis in Appendix B.2.2.

To investigate the effect of the PRE Split and condition on performance (learning), we run a factorial ANOVA with POST and NLG as dependent variables and two factors: the aptitude split (low vs. high pretesters) and the Condition (*NM* vs. *R*). We look at the effect of the aptitude factor and the effect of the aptitude-condition combination. The effect of the condition was discussed in the previous section.

Table 31 shows the average and standard deviation for the two learning metrics (POST and NLG) for PRE Split and PRE Split × Condition. We find a strong significant effect of PRE Split on POST ($F(1,48)=31.58$, $p<0.0001$) with high pretesters reaching a significantly higher POST score compared with low pretesters (21.5 on average versus 17.2). However, the effect on NLG, which accounts for PRE score, does not reach significance ($F(1,48)=2.41$, $p<0.13$).

Table 31. Learning performance based on PRE split and condition

PRE Split	Condition	# of users	POST	NLG
L		31	17.2 (2.7)	0.45 (0.18)
H		21	21.5 (2.8)	0.54 (0.23)
L	*NM*	16	16.4 (2.9)	0.40 (0.19)
L	*R*	15	17.9 (2.4)	0.50 (0.15)
H	*NM*	11	22.2 (2.3)	0.59 (0.19)
H	*R*	10	20.8 (3.3)	0.48 (0.26)

The combination PRE Split × Condition has a trend effect on both learning metrics (for POST: $F(1,48)=3.53$, $p<0.07$; for NLG: $F(1,48)=3.24$, $p<0.08$). This effect is best viewed graphically: Figure 23. For POST, we find that the NM is beneficial for high pretesters but detrimental for low pretesters. We find that *NM* high pretesters obtain a 1.4 POST points higher than *R* high pretesters though the difference is not significant (22.2 vs. 20.8, $p<0.25$). In contrast, *NM* low pretesters fare 1.5 POST points worse than *R* pretesters though again the difference is not significant (16.4 vs. 17.9, $p<0.13$). For NLG, we find that high and low *R* pretesters perform similarly while the NM presence boosts the average NLG for high pretesters (0.48 vs. 0.59, $p<0.23$) but decreases it for low pretesters (0.49 vs. 0.40, $p<0.19$). *NM* high pretesters perform significantly better than *NM* low pretesters (0.59 vs. 0.40, $p<0.02$). In terms of NLG effect size, for high pretesters the NM has a positive effect of 0.42, while for low pretesters the NM has a negative effect of 0.66.

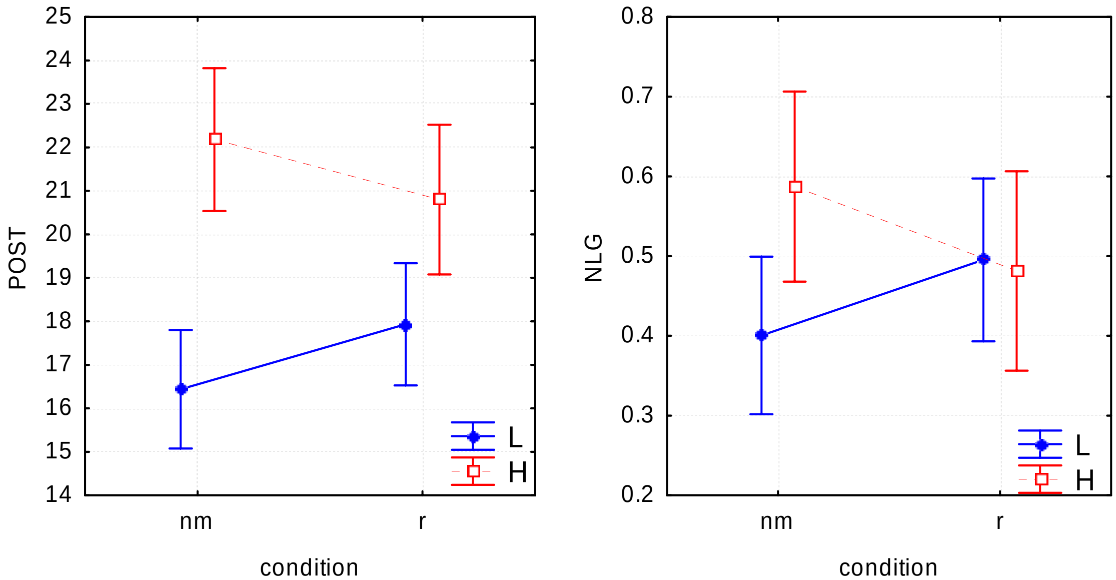


Figure 23. PRE Split × Condition effect on learning

(Average and 95% confidence intervals are shown)

On the RELNLG subset (2.3.3.2), the impact of the PRE Split × Condition combination is even clearer and reaches significance on both learning metrics (for POST: $F(1,43)=6.43$, $p<0.01$; for NLG: $F(1,43)=6.94$, $p<0.01$). Elimination of these users changes only the high pretest population used in the analysis. The results are shown in Figure 24. For high pretesters, there is a trend that *NM* works better as measured by POST (21.9 vs. 19.3, $p<0.06$) and the difference is now significant for NLG (0.62 vs. 0.40, $p<0.03$). As a result, the effect size of the NM for high pretesters almost doubles to .85.

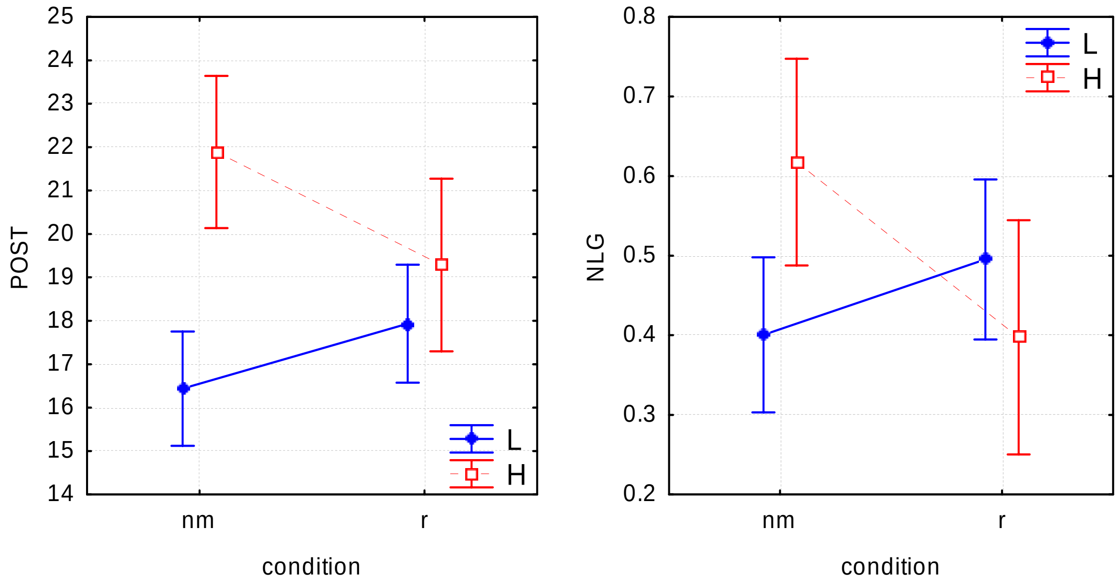


Figure 24. PRE Split × Condition effect on learning (RELNLG subset)

(Average and 95% confidence intervals are shown)

Since the effectiveness of the treatment depends on the PRE Split, whenever we compare *NM* and *R* on a metric, we will also investigate if there are differences in that metric for low/high pretesters in *NM* and *R*. For example, we will look to see if there a differences in the subjective metrics between low/high pretesters in the two condition. This type of analyses will allow us to hypothesize explanations for why *NM* worked better for high pretesters and not that well for low pretesters.

Discussion

The aptitude-treatment interactions suggest that the NM is an "expert tool": it is effective for "expert" users (high pretesters) but it has a negative effect for "novice" users (low pretesters). One explanation of this effect is that "expert" users (high pretesters) can easily absorb the information from the discourse segment intentions displayed in the NM. In contrast, "novice" users (low pretesters) have problems following and integrating this information with the tutor speech. This can be linked to the work of (McNamara and Kintsch, 1996) on the effect of text coherence. In their study, they manipulate the instructional text coherence to create two versions: an "expert" text and a "novice" text, with the "novice" text having semantic relationships more explicit (i.e. more coherent). Their results show that "expert" text is beneficial for "expert" users and detrimental for "novice" users while "novice" text has the opposite effect. In our case, the discourse segment intentions from the NM are the equivalent of the "expert" text while the system turns from the dialogue history are the equivalent of the "novice" text. This might also explain why *R* high pretesters perform less than *R* low pretesters in terms of NLG (on the RELNLG subset). The learning-coherence relationship was also investigated in ITSPOKE by other colleagues (Ward and Litman, 2006) but they look at tutor-student cohesion.

Another possible explanation was informed from the open-questions interview with users at the end of the experiment. When asked about the usefulness of the dialogue history, several *R* users indicated that it helped them when they "zoned out". They indicate that sometimes they will loose concentration and not listen to the entire tutor turn. In order to get back on track or to catch up, users indicated that they read the tutor turn from the dialogue history. However, this alternative is not available for *NM* users as they only see a "condensed" version of the tutor turn in form of dialogue segment intentions. In addition, when asked what information they used to update the essay after the dialogue with the system, some *R* users indicate that they reread the dialogue transcript from the dialogue history. Again, this option is not available for the *NM* users who have to make sense of the NM in order to "reread" the conversation. Updating the essay is an important step towards learning: it can be seen as a "reflection" step in which users self-express the assimilated knowledge. Previous work shows that other post-instruction activities are correlated with increased learning (e.g. (Katz et al., 2003)).

These hypotheses can be linked to the two differences between *NM* and *R* condition: *NM* users have the NM *but* they do not have the dialogue history. If the hypotheses are true, then an improvement might come from a new version of the NM that will also show the text of the current tutor (e.g. whenever the user clicks on a discourse segment). Human-computer interaction issues need to be considered for this implementation as it is not clear where the best place to display the tutor text is. This is particularly important as (Sweller, 1988) indicates that the visual and textual information should be placed together and that placing them separately has an effect on the cognitive load and can impact performance.

Summary

We find that the effect of treatment on learning depends on user's PRE aptitude: *NM* works better for high pretest users while *R* works better for low pretest users.

6.4.2.3 Subjective metrics

While there are no differences between the *NM* and *R* user population in terms of learning, we wanted to know if their perception of the system was changed by the NM. As described in Section 6.3.2, we found that when exposed to both *NM* and *R* (but without the dialogue history) users rate the *NM* version significantly higher on several dimensions. We investigate again the ratings from the system questionnaire. We first look at ratings for individual questions and then at ratings for groups of questions (2.2.3.1).

Table 32 shows the average rating in each condition for each question in the system questionnaire (similar in spirit to Table 28). We find that for all questions *NM* is rated better than or at least equal to *R* though none of the differences is significant. Note that for Q7 and Q11 (italicized in Table 32) a lower rating is better as they gauge negative factors (high level of concentration and task disorientation, respectively). The averages are similar only for Q3 and the last group of questions (Q14-Q16).

Table 32. System questionnaire results by question

Question	Average rating		
(Overall)	***NM***		***R***
1. The tutor increased my understanding of the subject	4.1 (0.8)	>	3.9 (0.8)
2. It was easy to learn from the tutor	3.7 (1.0)	>	3.5 (0.8)
3. The tutor helped me to concentrate	3.1 (1.1)	~	3.1 (1.1)
4. The tutor worked the way I expected it to	3.4 (1.2)	>	3.1 (0.9)
5. I enjoyed working with the tutor	3.3 (1.1)	>	3.1 (1.0)
6. Based on my experience using the tutor to learn physics, I would like to use such a tutor regularly	3.0 (1.2)	>	2.8 (1.1)
(Conv) During the conversation with the tutor:			
7. ... a high level of concentration is required to follow the tutor	*3.5 (0.9)*	<	*3.6 (1.0)*
8. ... the tutor had a clear and structured agenda behind its explanations	4.4 (0.9)	>	4.1 (0.8)
9. ... it was easy to figure out where the tutor's instruction was leading me	4.2 (1.0)	>	4.0 (0.8)
10. ... when the tutor asked me a question I knew why it was asking me that question	4.0 (1.0)	>	3.8 (0.9)
11. ... it was easy to loose track of where I was in the interaction with the tutor	*2.3 (1.0)*	<	*2.6 (0.9)*
12. ... I knew whether my answer to the tutor's question was correct or incorrect	3.8 (1.1)	>	3.4 (1.2)
13. ... whenever I answered incorrectly, it was easy to know the correct answer after the tutor corrected me	4.0 (1.0)	>	3.6 (0.8)
(Post) At the end of the conversation with the tutor:			
14. ... it was easy to understand the tutor's main point	4.0 (0.9)	~	4.0 (0.7)
15. ... I knew what was wrong or missing from my essay	3.9 (1.1)	~	3.9 (0.8)
16. ... I knew how to modify my essay	3.9 (1.0)	~	3.9 (1.0)

To get an overview of the differences between *NM* and *R* in terms of subjective metrics, we grouped individual questions in 3 categories corresponding to the timing of the perception being probed: overall system perception (**Overall**: Q1-Q6), during the conversation (**Conv**: Q7-Q13) and after the conversation (**Post**: Q14-Q16) (see Table 32). For each student we averaged the rating in each group. In addition, we also averaged for each user its ratings to produce an average rating per user (**ALL**: Q1-Q16).

Note that for the "negative" questions (Q7 and Q11) the rating was inversed (new value equals 6 minus old value) so that better ratings are in the same direction when averaging.

Table 33 shows in the second column the average for each question group. As expected, we find that *NM* fares better in the Overall and Conv category and fares similarly in the Post category. This leads to an overall rating of 3.7 for *NM* compared to 3.5 for *R*. However, none of the differences is significant according to a one-way ANOVA test. The only difference that is closer to a trend is for the Conv group (F(1,49)=1.89, p<0.18).

Table 33. System questionnaire results by question group (significant/trend differences marked with "*"/"t")

	All students			Minus outlier			Minus qNM>=4		
	NM		*R*	*NM*		*R*	*NM*		*R*
# users	26		25	25		25	20		25
Overall	3.4 (0.8)	>	3.3 (0.8)	3.5 (0.7)	>	3.3 (0.8)	*3.6 (0.7)*	*>^{t}*	*3.3 (0.8)*
Conv	3.8 (0.7)	>	3.5 (0.6)	*3.8 (0.6)*	*>^{t}*	*3.5 (0.6)*	**3.9 (0.6)**	**>***	**3.5 (0.6)**
Post	4.0 (0.9)	~	4.0 (0.7)	4.0 (0.8)	~	4.0 (0.7)	4.1 (0.8)	>	4.0 (0.7)
All	3.7 (0.7)	>	3.5 (0.5)	3.8 (0.6)	>	3.5 (0.5)	*3.9 (0.6)*	*>^{t}*	*3.5 (0.5)*

A quick look at the ALL rating indicates that the combined *NM* and *R* population has an outlier in terms of rating (see Section 2.4.1 for more information about outliers). Figure 25 shows the distribution of the ALL rating for the combined *NM* and *R* population. The outlier is in the *NM* condition and has an ALL rating of 1.6 (z-score=3.16), PRE=19 and POST=24. The user indicated in the free text comment box that it was hard for him to concentrate solely on the tutor voice and that he needed pictures and animations. He gave a rating of 4 ("Agree") for the NM-related questionnaire question (qNM: "it was hard for me to follow the information on the screen while listening to the tutor speech" - 2.2.4.3). The user rated with a very low rating most questions in the Overall and Conv categories.

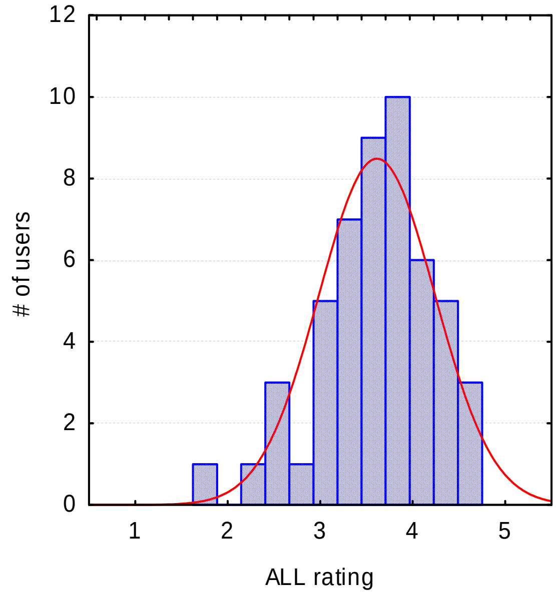


Figure 25. Distribution of the ALL ratings

The third column of Table 33 compares *NM* and *R* after the removal of the outlier. The difference between *NM* and *R* is increased in most categories. Statistical tests find a trend increase in rating for *NM* users on the Conv category (ANOVA for Conv: $F(1,48)=3.42$, $p<0.07$). The effect size on Conv category for *NM* is .55.

In the preliminary NM study we found that some users had problem concentrating on the speech and the NM in the same time and preferred not to have the NM. For *NM* users, an extra question was added to the questionnaire to identify these users (qNM – see Section 2.2.4.3). To see how our results change if we eliminate these users, we discarded the *NM* users that gave a rating of 4 or 5 (Agree or Strongly Agree) to this question. 6 users were removed by this criterion (the outlier is also covered by this elimination process). The resulting comparison is available in the fourth column of Table 33. Removal of these users[36] further increases the difference in rating between *NM* and *R* resulting in significant differences for the Conv category (ANOVA: $F(1,43)=5.00$, $p<0.03$) and a trend difference for Overall and ALL (ANOVA for Overall: $F(1,43)=2.74$, $p<0.10$; for ALL: $F(1,43)=3.90$, $p<0.06$). The effect size on the Conv category is .72 and for the ALL category it is .61.

[36] Note that since qNM was presented only to *NM* users, there was no way to identify similar users in the R condition. Since the R population might contain "similar" users, the comparison is an upper bound on the differences between *NM* and R when these users are removed..

Aptitude – Treatment interaction on Subjective metrics

In Section 6.4.2.2, we found that the NM works better for high pretesters while R works better for low pretesters. To gain more insight into why this happens, we look at the ratings for each category of users. Since the differences between NM and R users are clearer when we remove the NM outlier, these analyses will be done without the NM outlier[37].

We run a two-way factorial ANOVA with ratings as dependent variable and two factors: the PRE Split and the Condition. We are interested in trend/significant effects of the combination PRE Split × Condition as this combination has effects on learning. We find several trend/significant effects for individual questions but no effects for questions categories.

We find a significant effect on Q7 ($F(1,46)=5.73$, $p<0.03$) which is shown in Figure 26. We find that while the average rating for low and high R pretesters are similar, the average is higher for low NM pretesters and considerably smaller for high NM pretesters. Posthoc tests indicate that the rating of high NM pretesters is significantly smaller than the other three groups of users. Note that Q7 is a "negative question": lower ratings are better. In other words, the category of users that learns the most (NM high pretesters) thinks that a lower level of concentration is needed to follow the tutor. In contrast, the category with the least learning (NM low pretesters) agrees that a high level of concentration is required. We hypothesize that the reduced learning of NM low pretesters is linked to the fact that they need to concentrate more when the NM is present. Indeed, all users that gave a rating of 4 or more to qNM are low NM pretesters (with the exception of the NM outlier).

[37] If the NM outlier is included, the strength of the effects is reduced: the significant interaction becomes a trend, while the trend interactions we present loose significance.

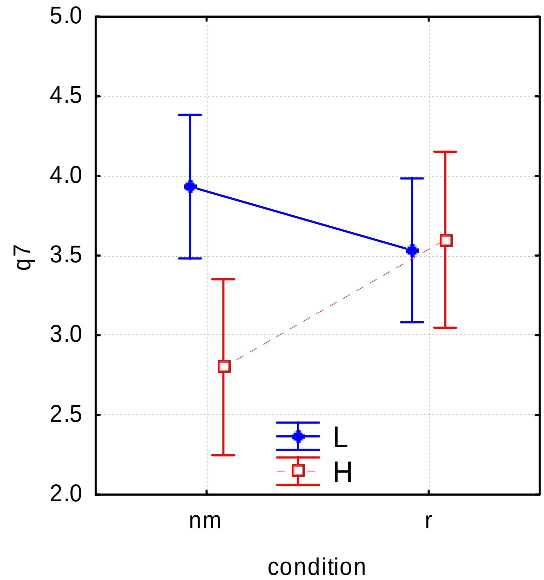


Figure 26. PRE Split × Condition effect on Q7
("…a high level of concentration is required to follow the tutor") (outlier removed)

We also find a trend effect on Q8 ($F(1,46)=2.97$, $p<0.09$), Figure 27. Posthoc tests find significant/trend differences between *R* high pretesters and the other 3 categories of users. In other words, it is harder for *R* high pretesters to see the structure behind instruction compared to *NM* pretesters. We hypothesize the since high *R* pretesters have already considerable knowledge, they are confused by the point of the system instruction. In contrast, it is easier for *NM* high pretesters (which also require less concentration) to identify the tutoring plan.

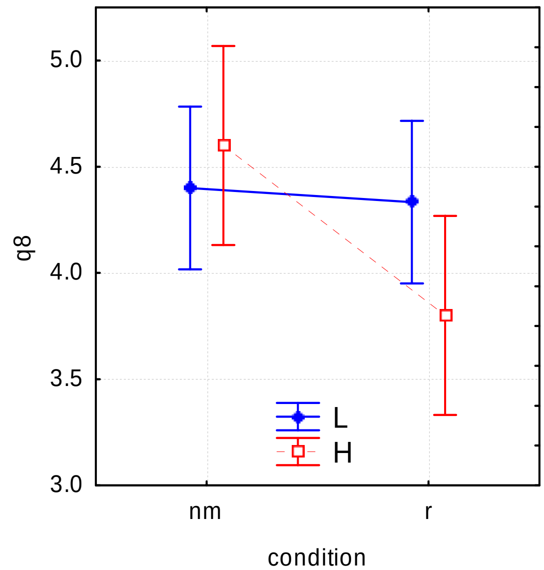


Figure 27. PRE Split × Condition effect on Q8
("… the tutor had a clear and structured agenda behind its explanations") (outlier removed)

An interaction similar to Q7 is found for Q9 where we find a trend effect ($F(1,46)=3.24$, $p<0.08$) - Figure 28. Posthoc tests indicate that it is significantly easier for *NM* high pretesters to figure out where the instruction is going (forward integration of the current question – see Section 6.3.2).We hypothesize that since it is easier for them to concentrate on the NM, *NM* high pretesters can take advantage of the limited horizon feature. By knowing what the system will discuss next, they perceive that it is easier to know where the instruction is going.

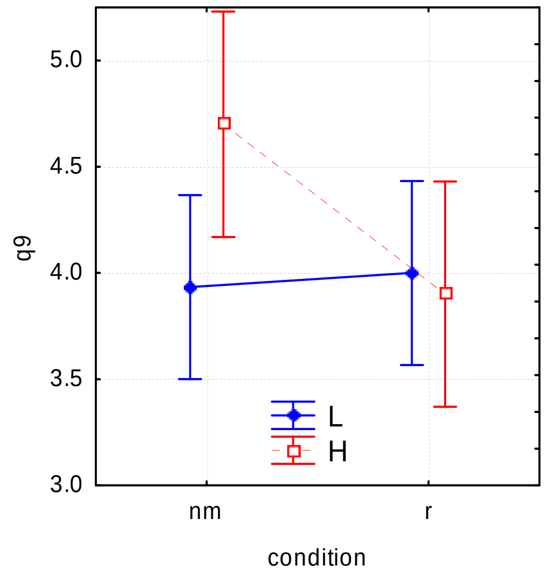


Figure 28. PRE Split × Condition effect on Q9

("…... it was easy to figure out where the tutor's instruction was leading me") (outlier removed)

Finally, we find a trend effect on Q1 ($F(1,46)=3.32$, $p<0.08$), Figure 29, which is similar to Q8. Posthoc tests indicate that when asked to self assess the effect of the system on their learning, *NM* high pretesters significantly outperform *R* low pretesters. This is probably linked to the answer on Q8 where *R* high pretesters have a harder time figuring out what the system wants from them.

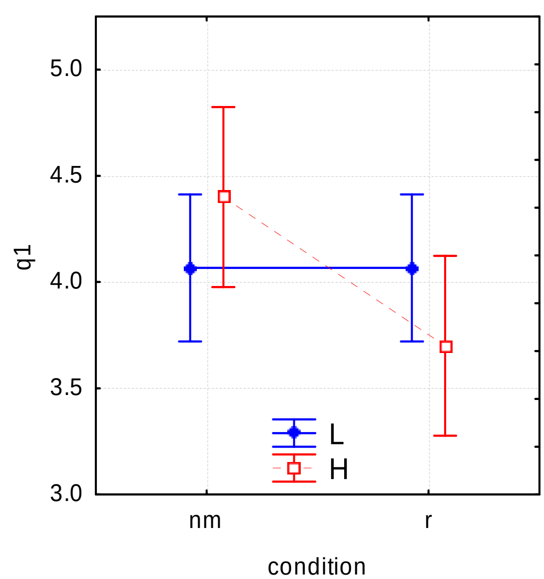


Figure 29. PRE Split × Condition effect on Q1

("The tutor increased my understanding of the subject") (outlier removed)

To summarize, subjective metrics offer us some insight on why the NM worked or failed. *NM* high pretesters perceive that a lower level of concentration is required (Q7) and that it is easier to identify the structure behind the conversation (Q8) and to figure out where the conversation is going (Q9). As a result, their self assessment of learning is improved. These perceptions are significantly different primarily from *R* high pretesters. *NM* low pretesters are similar in perception with their *R* counterparts with a bigger difference in terms of concentration but this difference is not significant.

Discussion

Although in the PrelimNM study we found significant/trend differences for several questions (recall Table 28), we did not expect to find any significant/trend differences in this study. The main difference between the two studies is that PrelimNM is a within-subjects study while this study is a between-subjects study. In PrelimNM users interact with both a NM version of the system and a version without the NM and rate each version of the system. In contrast, in this study users do not have a reference point when rating. Moreover, in PrelimNM the version without the NM does not have the dialogue history and it is an easier baseline compared with *R*. In addition, in the PrelimNM study a comparison of ratings for the first problem (which corresponds to a NM vs. noNM between-subjects comparison) does not produce any trend/significant differences.

We were not surprised by the lack of differences in the Post category between *NM* and *R*. All questions in this category capture the quality of the adaptation of the tutoring instruction to the student initial essay. Since the essay analysis component was disabled and all users received the same instruction regardless of their initial essay, it is not surprising that *NM* and *R* do not differ on this dimension.

The outlier we found in our data is a good example of users for which more information should be conveyed on the visual channel. This is in opposition with users that can only concentrate on a single communication channel (divided attention problem - see Section 6.3.2).

The aptitude-subjective interactions offer additional insight into the NM as an “expert tool” hypothesis we discussed in Section 6.4.2.2. It is indeed easier for expert *NM* users (high pretesters) to concentrate than for other users. In particular, expert *R* users when faced with the full dialogue history (the “novice” text as in the parallel we made with (McNamara and Kintsch, 1996)) have more problems than *NM* experts to concentrate, to identify the tutoring plan and to figure out where the instruction is going.

Summary

We find that on average users rate *NM* higher than *R* on most subjective questions however the differences are not significant. Grouping questions by timing of the perception further confirms these differences but still lacks statistical significance. Removal of an outlier results in trend differences and the trend becomes significant if in addition we remove *NM* users who indicate that they have problems following the information on the screen. Aptitude-subjective interactions indicate that *NM* high pretesters differ primarily from *R* high pretesters on several dimensions: ease of concentration, identification of tutoring plan, forward integration and their perception of the effectiveness of the system.

6.4.2.4 Subjective metric – performance relationship

As another way of investigating the effect of the NM, we looked for any relationship between user ratings and system performance (i.e. learning in our system). Previous work has also looked at this relationship. For example, in several information-access SDS, (Walker et al., 2000) finds a positive correlation between user satisfaction and user's perceived task completion (one of the measures for system performance - 4.6). However this relationship is not surprising in information access domains: users are satisfied when they *think* they got what they wanted from the system. For the tutoring domain, the relationship is more complex as the main focus is on how much users learn. For example, in one of the user comments reported in (Shelby et al., 2001), the user mentions that working with the system "was a pain to get done, but it definitely helped to understand the material by *forcing* certain method use and thought processes." Nonetheless, a study by (Atkinson et al., 2005) shows that a version of a system that was rated higher also produces higher learning. Note that user satisfaction/rating is measured in different ways across domains and systems although there are efforts for creating standardized questionnaires for SDS (e.g. the SASSI questionnaire (Hone and Graham, 2000)).

To investigate this relationship we looked at the correlation between various subjective metrics and various learning metrics in each of the conditions (2.4.2). For subjective metrics we used the group ratings (Overall, Conv, Essay and ALL). For learning metrics we used the POST score and the NLG. As another way to look at correlations with learning, we also looked at partial correlations between the subjective metrics and POST accounting for PRE (2.4.2).

Table 34. Subjective metrics – learning correlations for *NM* users

(outlier is removed)

	POST		NLG		POST partial PRE	
	R	p	R	p	R	p
Overall	**0.469**	0.018	**0.424**	0.035	0.384	0.064
Conv	**0.708**	0.001	**0.677**	0.001	**0.709**	0.001
Essay	**0.397**	0.049	**0.438**	0.029	**0.421**	0.040
ALL	**0.637**	0.001	**0.614**	0.001	**0.611**	0.002

We find no trend/significant correlations for the *R* population. In contrast, for the *NM* population minus the *NM* outlier, we find mostly significant positive correlations and partial correlations between all subjective metrics and all learning metrics. Table 34 shows the correlation coefficients and their significance. We find that users that rate the system higher in any of the three groups or overall tend to have higher learning. The correlation is very strong for the Conv group and as a result for the overall rating (ALL). Figure 30 displays the regression line for the correlation ALL rating – NLG. We observe that users that rate the system lower tend to have smaller learning gains while users with high ratings tend to have high learning gains. The results are the same on the RELNLG subset (2.3.3.2). If the *NM* outlier is included, the correlations remain trend/significant only for Conv and ALL.

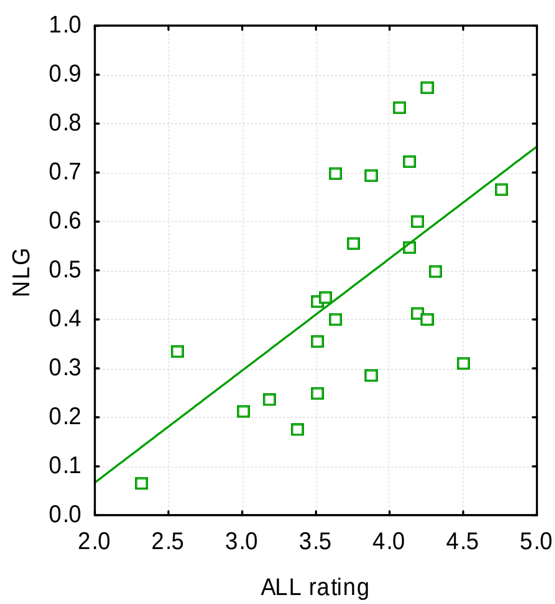


Figure 30. Correlation between the ALL rating and NLG for *NM* users

(outlier removed)

Discussion

We hypothesize that the NM produces a "hand glove" reaction in users: if the NM "fits" the user, he/she tends to obtain higher learning gains; if the NM does not "fit" the user, he/she tends to obtain lower learning gains. In contrast, the *R* version produces higher learning even for people with a lower rating and vice-versa: users can like the system but that does not necessarily translate into higher learning (scatter plot available in Appendix B.2.3 - Figure 41).

It is important to study if there are other factors that can explain the observed effect for *NM* users and its lack for *R* users. As discussed in Section 2.2.4.2, the *NM* and *R* user population are similar in terms of their pretest aptitude. In addition, the pretest score is not significantly correlated with NLG and ALL rating in both conditions (*R*: PRE-NLG ($R=0.07$, $p<0.70$), PRE-ALL ($R=0.01$, $p<0.96$); *NM*: PRE-NLG ($R=0.16$, $p<0.44$), PRE-ALL ($R=0.28$, $p<0.16$)). Thus the pretest distribution can not be responsible for the observed effect. But maybe the aptitude-treatment interaction from Section 6.4.2.2 can explain it: maybe because the NM does not work that well for low pretesters, these users learn less and rate lower. A two-way factorial ANOVA with ALL rating as dependent variable and two factors (the PRE Split and the Condition) is not statistically significant ($F(1,46)=2.00$, $p<0.17$) but not very far from a trend. However, if we look at the interaction graphically (Figure 31), we observe that if there was any interaction, it would have come from high pretesters: the averages are almost identical for low pretesters. So it is possible the NM makes high pretesters learn more and rate higher and thus produce the observed NLG-ALL correlation. However, our data shows that some of the low pretesters in the *NM* condition do achieve higher learning (4 out of the 16 low *NM* pretesters have an NLG of 0.6 or higher (see Figure 39)) and also rate the system high (ALL rating of 3.6 or higher)). More analyses are needed in the future to better understand the observed relationship between subjective metrics and learning.

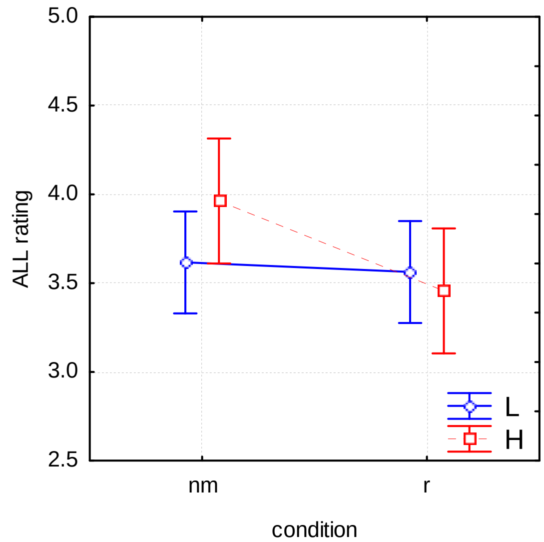


Figure 31. PRE Split x Condition non-significant effect on ALL rating (outlier removed)

Summary

We find a relationship between user perception of the system and their learning performance for the *NM* users but not for *R* users. For *NM* users, their ratings are positively correlated with learning: users with higher ratings also tend to have higher learning (note that no causal relationship is claimed).

6.4.2.5 Dialogue time

We also wanted to know if the Navigation Map has an effect on measures of dialogue efficiency. When designing tutoring systems researchers strive for learning efficiency: deliver increased learning as fast as possible. We look at two shallow dialogue metrics: dialogue time and number of turns.

Section 2.3.3.3 defines and discusses the dialogue duration metric. The metric was computed for each problem (**P1Time-P5Time**) and as a total (**TotalTime**). Table 35 shows the average and standard deviation of the 6 dialogue time metrics (in seconds) in each condition. We find that for all usable[38] metrics the *NM* users spent less time conversing with the system. ANOVA tests find a trend difference in the fourth problem (P4Time: $F(1,50)=3.82$, $p<0.06$) and significant differences in the last problem and the total time (P5Time: $F(1,50)=8.89$, $p<0.004$; TotalTime: $F(1,50)=8.31$, $p<0.006$). We observe that even on

[38] We do not expect any differences in the third problem: it is the smallest problem with no remediation dialogue and very few system questions. There is little place for variation for this metric as indicated by the very small standard deviations (12 and 13 seconds respectively).

the first problem *NM* users finish the dialogue with the system faster. There continues to be a difference in the second problem but a very small one. Once we get to the last two problems the differences between *NM* and *R* become more and more significant. As a result, on average *NM* users spend 5 minutes less conversing with the system.

Table 35. Average and standard deviation for dialogue time

Condition	P1Time	P2Time	P3Time	P4Time	P5Time	TotalTime
NM	642 (171)	407 (177)	134 (12)	624 (109)	588 (100)	2395 (323)
R	726 (199)	427 (148)	130 (13)	688 (126)	701 (167)	2671 (368)

Since users' initial knowledge (the PRE score) is an important factor in how much time users spend with the system, we wanted to know if the differences are influenced by this factor. Indeed, the PRE score is significantly negatively correlated with the TotalTime for the combined *R* and *NM* population (R=-0.50, p<0.001). To account for PRE in our comparison, we run an ANCOVA with dialogue time metrics as dependent variables, PRE as the covariate and the Condition as the independent variable (*NM* vs. *R*). The results are similar to ANOVA tests: we find trend differences for P1Time and P4Time and significant differences for P5Time and TotalTime.

Aptitude – Treatment interaction on Dialogue Time

In Section 6.4.2.2, we found that *NM* works better for high pretesters while *R* works better for low pretesters. To gain more insight into why this happens, we looked at the dialogue time for each category of users. We run a factorial ANOVA with dialogue time as dependent variables and two factors: the PRE Split (Low vs. High) and the Condition (*NM* vs. *R*). Since PRE is negatively correlated with dialogue time, it is not surprising that PRE Split has a significant effect on the TotalTime (F(1,48)=11.07, p<0.002) with high pretesters having shorter dialogue time than low pretesters. We also find that the combination PRE Split × Condition has a significant effect for the fifth problem (F(1,48)=4.73, p<0.04) and a trend effect on total dialogue time (F(1,48)=3.43, p<0.07). The effects are presented graphically in Figure 32. The effect of the NM is seen on high pretesters: their dialogue time is reduced significantly compared with both *R* low and high pretesters for both P5Time and Total time. *NM* low pretesters also have shorter dialogue time than their *R* counterparts but the difference is very small. Interestingly, for high *R* pretesters, the dialogue time is longer than low *R* pretesters on the fifth problem though overall they spend less time.

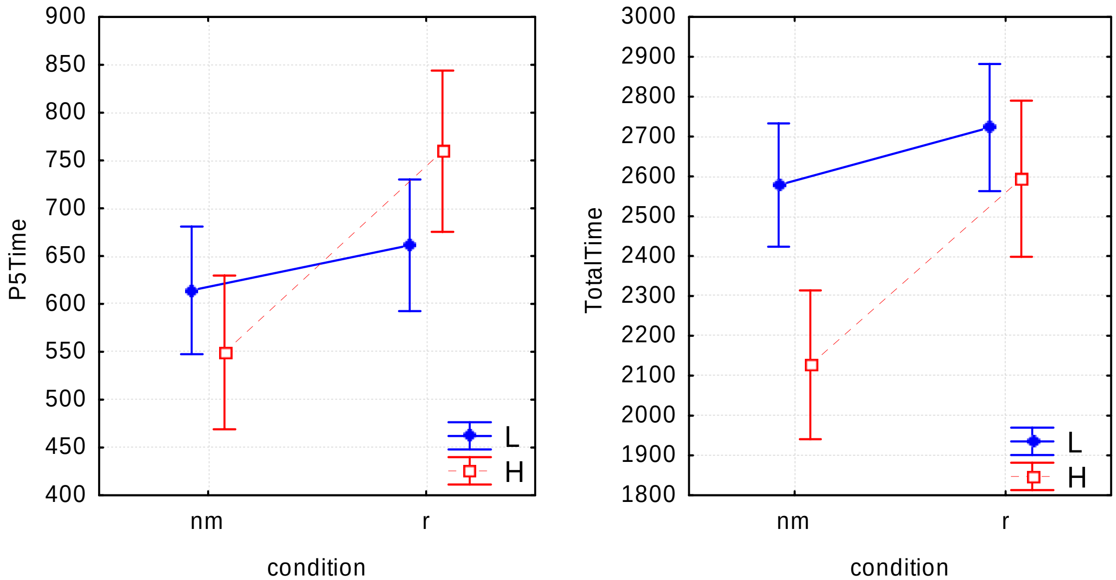


Figure 32. PRE Split × Condition effect on dialogue time (P5 and Total)

(Averages and 95% confidence intervals are shown)

Discussion

In Section 6.4.2.1, we saw that *NM* and *R* users do not differ in terms of learning. Our results show that even if *NM* users learn overall the same as *R* users, they have shorter dialogues. Thus, although the NM did not have the anticipated learning effect on the entire population, it does reduce the dialogue time. A closer look at low/high pretesters shows that *NM* high pretesters benefit both in terms of higher learning and in terms of dialogue time. *NM* low pretesters also have slightly shorter dialogue time, but their learning is reduced compared to *R* low pretesters.

We hypothesize that the reduction in dialogue time is due to the NM influencing users' answer conciseness and lexical choice. In the future, we plan to test this hypothesis by comparing *NM* and *R* in terms of number of words in user turns and speech recognition problems (particularly semantic misrecognitions - 2.3.1.5).

It is also interesting to interpret the dialogue time differences in terms of the knowledge used in each of the problems: problem 4 is a direct follow-up of problem 1 and 2 while problem 5 combines the knowledge from all previous four problems (Appendix A.2). We found that *NM* users spent less time on these two last problems. We hypothesize that this might be related to one or both the following factors: *NM* users give more correct answers and/or have fewer "detours" caused by speech recognition problems.

We plan to test these hypotheses in the future. It is also interesting to note that the aptitude-treatment interactions are visible only in Problem 5 which combines the knowledge from all previous problems

Summary

Although in terms of learning *NM* users do not differ from *R* users, we find that the dialogue time of *NM* users is significantly shorter than *R* on several problems and in total. The same effect holds when we discount for the PRE score. In addition, we find that high pretesters benefit the most from the NM in terms of reduction of dialogue time.

6.4.2.6 Number of system turns

As another metric that measures the dialogue length, we also looked at the number of system turns. Section 2.3.3.4 defines and compares this metric with the dialogue duration metric. The metric was computed for each problem (**P1Tut-P5Tut**) and as a total (**TutTotal**). Table 36 indicates a result similar to dialogue time: on all usable[39] metrics *NM* have fewer system turns with a trend difference for the fourth problem ($F(1,50)=2.79$, $p<0.10$). Unlike dialogue time, no trend/significant difference is present in the fifth problem. In total, *NM* users have 4 tutor turns fewer than *R* users but this difference is not significant ($F(1,50)=2.30$, $p<0.14$).

Table 36. Average and standard deviation for the number of system turns

Condition	P1Tut	P2Tut	P3Tut	P4Tut	P5Tut	TotalTut
NM	23.8 (5.0)	16.6 (5.7)	7.3 (0.7)	19.6 (3.5)	18.8 (2.1)	86.1 (9.4)
R	25.6 (5.2)	16.8 (5.2)	7.1 (0.3)	21.4 (4.0)	19.2 (2.4)	90.1 (9.7)

Since PRE score is also significantly negatively correlated with total number of system turns ($R=-0.431$, $p<0.001$), we run an ANCOVA to discount for PRE in our analyses and we found similar results (a trend on the 4th problem and a trend for total number of tutor turns).

Aptitude – Treatment interaction on Number of system turns

We also run a factorial ANOVA with PRE Split and condition as factors. We find a trend effect of the PRE Split × Condition combination on the first problem and the fifth problem (for P1Tut: $F(1,48)=3.40$, $p<0.08$, for P5Tut: $F(1,48)=3.32$, $p<0.08$). Figure 33 shows these effects graphically. We find that *NM* high pretesters benefit from a significant reduction in system turns in the first problem as

[39] As for dialogue time, we did not expect any differences on the third problem. There is little place for variation for this metrics as indicated by the very small standard deviations (0.7 and 0.3 turns respectively). Variations are solely due to speech problems (i.e. rejections and timeouts).

compared with other *NM* and *R* users whose behavior is similar. We again find *R* high pretesters (*NM* worked better for high pretesters) going through more system turns in last problem. It is interesting to compare Figure 32 and Figure 33 on the fifth problem. Although *NM* low pretesters have a similar number of turns with *NM* high pretesters, their dialogue time is longer which suggest that they give more incorrect answers to questions that do not require remediation dialogues. Similarly, when we compare *NM* and *R* low pretesters, we find a similar number of system turns but *R* low pretesters have more dialogue time.

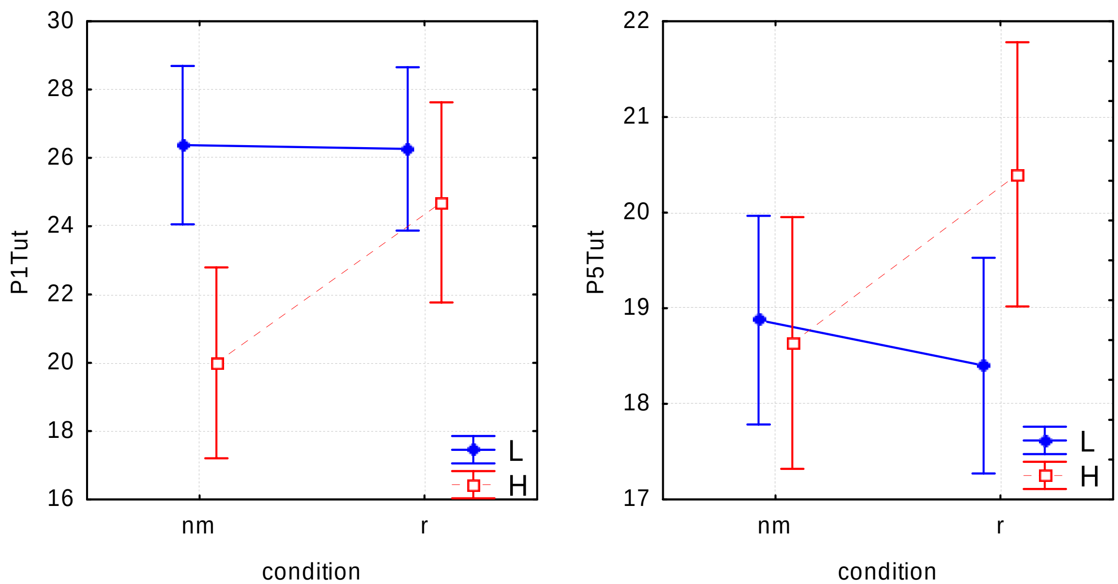


Figure 33. PRE Split × Condition effect on number of tutor turns (P1 and P5)

(Averages and 95% confidence intervals are shown)

Summary

The results for the number of system turn metric confirm the results for the dialogue time metric: *NM* users have fewer system turns than *R* users though the difference in not significant. The same effect holds when we discount for the PRE score.

6.5 RELATED WORK

Most SDS research has focused on the speech-only condition: the communication between users and the SDS is done via voice only. Factors like the ubiquity of the land and mobile phone services and commercial applications (e.g. call centers) have contributed to this. In these settings, systems can signal to users information about discourse structure using lexical means (e.g. (Hovy, 1993; Passonneau and Litman, 1997)) and/or prosodic means (e.g. (Hirschberg and Nakatani, 1996; Möhler and Mayer, 2001; Pan, 1999).

We focus on systems that use multiple modalities on the output side. With the increase in performance of desktop computer systems and mobile devices (e.g. laptops, PDAs), multimodal SDS have gained in popularity. These systems employ multiple modalities to support communication: pen/gesture input and/or text/graphical output (Allen et al., 2000; Gruenstein et al., 2006; Oviatt et al., 2000). While in certain domains graphical output is part of the underlying task (e.g. geographical applications (Gruenstein et al., 2006; Oviatt et al., 2004)), it has been used in other systems to increase usability: animated talking heads (Graesser et al., 2003; Graesser et al., 2001), segmented interaction history (Rich and Sidner, 1998), timeline layout of the current plan in a planning assistant (Allen et al., 2000).

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: an indented text representation of the interaction history augmented with plan purpose information. We extend over their work in several areas. The most salient difference is that we investigate the benefits of displaying the discourse structure information for users. In contrast, (Rich and Sidner, 1998) never test the utility of the segmented interaction history. 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 simpler than our tutoring task. To our knowledge we are the first to experiment with a graphical representation of the discourse structure in tutoring SDS. In addition, the NM displays only the purpose information while the segmented interaction history also displays the conversation transcript reducing the amount of information users can see at one time on the screen. Also, the segmented interaction history is not always available and users have to activate it manually. Nonetheless, (Rich and Sidner, 1998) indicate that the segmented interaction history enables new ways for users to directly control the dialogue flow. In their system users can click any segment in the segmented interaction history and ask the system to stop, restart or replay that segment. Similar functionalities can be envisioned for users of the NM: e.g. going back to previous system questions, skipping instruction caused by interpretation errors, pausing instruction, accessing additional tutoring for topics of interest, etc.

Also of interest is the work presented in (Mancini et al., 2006). Their work investigates graphical representations of discourse coherence relations. For example, they represent an ELABORATION relationship between two text segments by overlapping the rectangular representation of the two segments; the rectangle of the elaboration segment is indented and on top of the other rectangle. Animation is used to introduce new text segments graphically. While their study and ours differ on many dimensions, there are many avenues for cross-pollination. Among the important differences, their work looks only at discourse relations while we look at discourse segment purpose information too. Nonetheless, they investigate 8 relations inspired by the Rhetorical Structure Theory (Mann and Thompson, 1987) (e.g. elaboration, contrast, conditionality, etc.) in a monologue setting while we look only at the 2 relations proposed in the Grosz and Sidner theory (dominance and satisfaction-precedence) in a dialogue setting. In terms of evaluation, they are investigating if users exhibit "stereotypical preferences for particular visual renderings of coherence relations" while we look at the impact of the graphical representation on performance of the system (both subjective and objective). The NM can benefit from this work in case more complex relationships need to be represented graphically. This might be particularly important in other complex domains like legal problem-solving where the focus is on learning to identify relationship between entities (Lynch et al., 2007).

Visual improvements for dialogue-based computer tutors have been investigated in the past. For example, (Graesser et al., 2003) investigate various output modalities for the AutoTutor system (a dialogue-based computer tutor for computer literacy topics like hardware, software, etc): text-only, speech-only, speech plus a talking head, speech plus text plus a talking head. Their results indicate that modality affects the performance of the tutor but the differences are not significant (conditions are ordered as follows according to average posttest score: text-only ~ speech-only < speech plus a talking head < speech plus text plus a talking head). However the AutoTutor talking head and the NM differ in terms of the type of information they facilitate for the user: while the AutoTutor talking head is used to signal dialogue moves, turn-taking and feedback through facial expressions, the NM offers different views on the tutoring information through its use of discourse segment purpose and the discourse segment hierarchy.

A different AutoTutor study (Jackson and Graesser, 2007) shows that graphical improvements that provide additional tutoring information (e.g. concept highlighting) result in significant improvements. The NM as we described it in Section 3.4 has an equivalent: the information highlight feature. It highlights not only concepts but also correct answers and other important information that will facilitate integration (e.g. highlighting the items in a list). Although this feature was enabled in our first NM study, it was disabled in our second study. Nonetheless, even without this feature we still have positive results.

Another interesting related work is that of (Lynch et al., 2007). LARGO, their tutoring system for legal domain, supports students' graphical note taking of legal court arguments: students identify spans of text (e.g. facts, hypotheticals, tests, etc), summarize them and draw relationships between these spans (e.g. modification, supports, distinguishing, etc). This process is similar in spirit with annotating the discourse structure: spans of text have to be identified (i.e. discourse segments) and summarized (i.e. intention of a discourse segment) and various types of relationships between these spans have to be identified (i.e. the discourse segment hierarchy). In fact, their approach and our Navigation Map fall under the same framework: graphical representations of entities and of the relations between them. In their case, students have to identify and relate these entities; for the NM, this work has already been done by the annotator and it is used as a graphical support for spoken instruction.

The NM requires an annotation of the discourse segment purpose and discourse segment hierarchy information. For our system, we manually annotated this information using a single annotator (the author). Since the main goal of this proposed work is to see if a graphical representation of this information is of any help, the reliability of our annotation is of secondary importance. Nonetheless, we believe that our annotation is relatively robust as our system follows a carefully designed tutoring plan and previous studies have shown that naïve users can reliably segment discourse (e.g. (Passonneau and Litman, 1997)). Moreover, because ITSPOKE uses system initiative (i.e. the system drives the conversation by asking questions), the discourse structure annotation is simplified as we do not need to recognize the user plan as many user-initiative systems have to do (e.g. (Allen et al., 2000; Blaylock and Allen, 2006)). Additional studies can be run to measure the reliability of the discourse structure annotation and to investigate other NM issues (e.g. the design choices we made: graphic layout, showing correct answer, information highlight).

6.6 SUMMARY & FUTURE WORK

We find that discourse structure information is useful for the graphical output of SDS. We use the discourse structure hierarchy and the discourse segment intention/purpose information that is part of discourse structure. Use of our graphical representation, the Navigation Map (NM), improves over specific baseline versions of our system on a number of subjective and objective metrics. When exposing users to both a NM version and a baseline version with no graphical support during the conversation, we find that users prefer the NM version and rate it better on a number of dimensions (e.g. ability to identify and follow tutoring plan, integrate instruction, concentration, etc.).

When compared to a baseline that provides graphical support during conversation in the form of a dialogue transcript (regular ITSPOKE), we find that the NM version leads to increased system performance (i.e. learning) for high pretesters but decreased performance for low pretesters (the version of the system and user pretest score aptitude have a trend effect on learning). As a result, the overall performance of the NM version is similar to that of the regular ITSPOKE. However, even overall there are a number of positive effects for the NM version: significantly shorter dialogue time and better user ratings (trend differences when we aggregate ratings based on timing). The NM is also associated with a "hand glove" effect (a positive correlation between user's perception and their learning), an effect which we do not observe for the regular version of ITSPOKE. Note that there are two differences between the Navigation Map version of the system and the baseline used in this study. One is that the Navigation Map is present and the other is that *the dialogue history is disabled*. Thus improvements are attributed to two factors: showing the NM *and* hiding the dialogue history. Additional users studies are needed to investigate how much each of the two factors contributes to observed improvements.

Our findings motivate further studies to better understand the effect of presenting the discourse structure information on the graphical output. Our results suggest that the simplest way to improve our system is to use the regular version for low pretesters and the NM version for high pretesters. Based on the average pretest score from all our corpora (Table 5) we believe that a threshold of 12 out of 26 should be used to define low/high pretesters. Validation of this approach should be done by running a user study which compares the regular version of the system with this new mixed version.

Another alternative is to have students choose between the two versions instead of assigning them based on the pretest score. Two observations suggest this approach might work. First, due to the observed NM "hand glove" effect, users that rate the system higher also tend to learn higher. While this correlation does not imply a causal link, we hypothesize that we can trust user's perception as an assignment criterion. Second, our data shows that some of the low pretesters in the *NM* condition do achieve higher learning (4 out of the 16 low *NM* pretesters have an NLG of 0.6 or higher (see Figure 39). Since these users also rate the system higher (ALL rating of 3.6 or higher - 6.4.2.3), the new approach might assign them to the *NM* condition instead of the *R* condition as the pretest assignment will do. To test this new approach, a user study can be run in which users try both versions in the first two ITSPOKE problems and then choose the version to be used for the other three problems.

It is interesting to investigate how to make the NM effective for low pretesters. One way to address this problem is informed by our hypothesis that the NM was not effective for low pretesters because it contains "expert" text (the discourse structure intention) instead of "novice" text (the system turn transcript). We can imagine a version of the NM where under each discourse segment intention we also have the corresponding discourse segment text. However, it is not clear the best way to fit this information on the screen. One solution would be to show the discourse segment text whenever a discourse segment is active or the user clicks it.

In our results on objective metrics, we used a version of the NM that has one feature disabled: the information highlight. We disabled this feature because this information source is not part of discourse structure. It would be interesting to see if enabling this feature will further increase some of the NM positive effects. Also, the NM effect might be even stronger for versions of the system that used synthesized speech. To account for potential negative effects of using a synthesized speech output, in our NM experiments we used human prerecorded prompts. It would be interesting to replicate our experiments with synthesized speech output.

The NM uses two discourse structure elements: the discourse segment hierarchy and the discourse structure intentions. It would be interesting to understand which of the two elements is useful for users. We can imagine an experiment that compares a version of the NM that shows only the discourse structure intentions with a version that shows only the discourse segment hierarchy (though it is not clear how to present the indentation without other textual information).

The NM falls under a general framework: graphical representations of entities and of the relations between them. It would be interesting to understand how to best instantiate this framework for a given domain: what are the best entities and relations to present and what is the best way to present these graphically (e.g. (Mancini et al., 2006)).

The NM enables novel ways of interaction between users and systems. Similarly to (Rich and Sidner, 1998)'s segmented interaction history, users can use the NM to directly control the dialogue flow. For example, by clicking on a discourse segment, users can ask the system to pause, restart or replay that segment. In addition, we can provide additional explanations for certain discourse segments or even allow users to skip certain topics (e.g. if they are familiar with that topic or the topic was activated due to semantic interpretation errors). These new features raise additional design issues as, for example, whenever we hand over part of the control to users, user can game the system (e.g. (Baker et al., 2008)).

Finally, we have investigated the NM for a system that uses system initiative. It would be interesting to study how to construct and use the NM in a system with user initiative. For these systems, the NM requires recognition or inference of user's plan (e.g. (Allen et al., 2000; Blaylock and Allen, 2006)). In the author's unpublished work with Shimei Pan, we looked at such a system: a multimodal system that allows users to browse a large real-estate database (Zhou et al., 2005). Users can browse various entities like houses, cities, school districts, hotels, etc. Our choice of discourse segments for that system was user queries. To build the NM, we attempt to infer dominance relationships between these queries. A NM was built for each type of entity. The resulting NM offers opportunities for a number of important tasks in such domains: query relaxation, result summarization, query refinement and information push.

7 CONCLUSIONS & FUTURE WORK

In this work we establish discourse structure as an important information source for spoken dialogue systems (SDS). Dialogues, either human-human or human-computer, have an inherent structure called the **discourse structure** (3.1). Due to the relatively simple structure of the dialogues that occur in the information-access domains of typical spoken dialogue systems (e.g. travel planning), discourse structure has often seen limited application in such systems. However, we believed that novel applications of discourse structure will be enabled for SDS in more complex domains (e.g. tutoring). We investigate this in the context of a SDS that tutors conceptual physics (ITSPOKE - 2.1).

We use the Grosz & Sidner theory of discourse structure (Grosz and Sidner, 1986) and we show that the discourse structure information is valuable for two system-side applications: performance analysis and characterization of dialogue phenomena. In addition, we show that discourse structure information is also valuable for users of a SDS through a graphical representation, which we call the Navigation Map (NM).

The investigation of the system-side applications was primarily driven by our intuition that the "position in the dialogue" is a critical information source. We show how we can define this concept using the notion of discourse structure transitions (3.3). These transitions capture events that happen in the discourse segment hierarchy: crossing of discourse segment boundaries (Push, PopUp, PopUpAdv), remaining in the same discourse segment (Advance) or lack of new information (SameGoal). We offer an automatic way of annotating the discourse segment hierarchy and consequently of automatically annotating discourse structure transitions.

For performance analysis (Section 4), we use the transition information in combination with other dialogue phenomena to derive a number of factors (i.e. transition–phenomena and transition–transition bigrams/interaction parameters). Our results show that these factors are both predictive and informative. In terms of predictiveness, we find that several of these factors are correlated with performance. The predictiveness of one of these factors (PopUp–Incorrect) generalizes to other corpora. This factor also informs a promising modification of our system: offer additional explanations after PopUp–Incorrect events. We implement this modification and we compare it with the original version of the system by running a user study. Analysis of the data from the user study reveals several interesting findings. We find that the modification breaks down the negative correlation between the PopUp–Incorrect factor and system performance. In addition, users that need the modification the most (i.e. users with a larger number of PopUp–Incorrect events) show significant improvement in performance in the modified version of the system over corresponding users in the old version of the system. However, this improvement is not strong enough to generate significant differences at the population level. Even though the additional explanations that are part of the new modification add extra time to the dialogue, overall we see no increase in dialogue length in terms of dialogue time or number of system turns for the new version of the system.

For characterization of dialogue phenomena (Section 5), we find statistical dependencies between transitions and two dialogue phenomena (uncertainty and speech recognition problems) in a previously collected corpus. Several transitions in the discourse structure are associated with an increase or decrease of uncertainty and have specific interaction patterns with speech recognition problems. The interactions we find allow us to formulate hypotheses behind the presence/absence of the phenomena, which in turn suggest ways to modify our system. For example, some of these dependencies provide additional motivations for our graphical representation of discourse structure (the Navigation Map). At the very least, these dependencies suggest that discourse structure transition information can be used as an informative feature for predicting dialogue phenomena.

We find that discourse structure information is useful for the graphical output of SDS (Section 6). We use the discourse structure hierarchy and the discourse segment intention/purpose information that is part of discourse structure. Use of our graphical representation, the Navigation Map (NM), improves over specific baseline versions of our system on a number of subjective and objective metrics. When exposing users to both a NM version and a baseline version with no graphical support during the conversation, we find that users prefer the NM version and rate it better on a number of dimensions (e.g. ability to identify and follow tutoring plan, integrate instruction, concentration, etc.). When compared to a baseline that provides graphical support during conversation in the form of a dialogue transcript (regular ITSPOKE), we find that the NM version leads to increased system performance (i.e. learning) for high pretesters but decreased performance for low pretesters (the version of the system and user pretest score aptitude have a trend effect on learning). As a result, the overall performance of the NM version is similar to that of the regular ITSPOKE. However, even overall there are a number of positive effects for the NM version: significantly shorter dialogue time and better user ratings (trend differences when we aggregate ratings based on timing). The NM is also associated with a "hand glove" effect (a positive correlation between user's perception and their learning), an effect which we do not observe for the regular version of ITSPOKE.

Besides establishing discourse structure as a useful information source for SDS in complex domains, our work also contributes in other aspects. Through our discourse structure transitions (3.3), we offer a way of defining the notion of "position in the dialogue", a definition which is domain-independent and automatically computable. Defining this concept was crucial for our performance analysis and characterization of dialogue phenomena investigations. Through our transition–phenomena bigrams (4.3), we provide a way to look at dialogue phenomena in their dialogue context. Our results show that looking at dialogue phenomena (correctness and user affect) in their dialogue context offers more insights about system performance compared to looking at dialogue phenomena out of context (4.4). Through our transition–transition bigrams (4.3), we offer a way of quantifying the structure of a dialogue. Our results (4.4) show that the structure of the dialogue, as measured by parameters derived from these bigrams, can discriminate between "good" and "bad" dialogues with our system (where good/bad is defined in terms of learning). Statistical dependencies between discourse structure transitions and dialogue phenomena suggest discourse structure transition as an informative feature for predictive models of dialogue phenomena (5.4 and 5.5). These dependencies also show how discourse structure transitions can be used to better understand the behavior of a system. It is interesting to note that the two applications make different conceptual use of discourse structure transitions. For performance analysis, we use transitions in collaboration with dialogue phenomena to model system performance. For characterization of dialogue phenomena, we use transitions to model dialogue phenomena by looking at statistical dependencies between them.

For performance analysis, we provide one of the few examples of a complete application of this approach to empirical dialogue design. By examining existing dialogues, we discover *new* factors related to performance of a SDS. These factors relate to issues in the design of the SDS that were not anticipated beforehand (e.g. the PopUp–Incorrect behavior - 4.5.1). We do not stop at identifying these issues, but we also show that by addressing them we obtain performance improvements (at least for certain users – 4.5).

We also show that the information presented on the graphical output of a SDS (if available) can have a considerable impact on system performance. We show that by replacing the transcript of the dialogue with a graphical representation of discourse structure (the NM) we obtain a number of subjective and objective improvements. It is interesting to note that this improvement comes without making any changes to actual spoken dialogue between the system and users. We provide one of the first graphical representations of discourse structure (3.4).

We also highlight some of the issues that might arise from similar use of discourse structure in other systems/domains (Section 3.5): the inherent connection between discourse structure transitions and user correctness/remediation subdialogues in our system, issues related to the interpretation of discourse structure transitions, reliability and granularity issues for the NM, and the effects that the initiative type and the task representation have on obtaining the discourse structure information.

Our work opens many avenues for interesting future work. For performance analysis, it would be interesting to replicate our analysis on other systems/domains. This should be a relatively straightforward process because our definition of discourse structure transitions is domain independent and a similar automatic annotation of the discourse structure can be performed in SDS that rely on dialogue managers inspired by the Grosz & Sidner theory of discourse structure. The transition–transition bigrams are already domain independent and can be easily computed once we have the transition information. For the transition–phenomena bigrams, researchers have only to identify relevant dialogue phenomena (e.g. speech recognition problems, corrections, out-of-domain utterances, semantics of user input, etc). In addition, it would be interesting to replicate our analyses of predictiveness for other metrics in other systems: e.g. task completion, user satisfaction, metrics that include business costs, etc.

The predictiveness of the discourse structure-based parameters suggests their potential for PARADISE modeling (Walker et al., 1997). Work in collaboration with other colleagues (Forbes-Riley et al., 2008) has already made use of our transition–user affect and transition–transition parameters. Although the purpose of that study was to show that the inclusion of parameters that use user affect increases the quality and generality of PARADISE models, results show that the resulting PARADISE models make heavy use of the discourse structure-based parameters. It would be interesting to do a similar study that looks at the relative utility of the discourse structure-based parameters for PARADISE modeling.

In terms of design methodology for SDS in tutoring, our performance analysis results suggest the following design principle: “do not give up on students but try other approaches” (i.e. in our case, we do not give up after a PopUp-Incorrect but we provide additional explanations). Note that our experiment design provides no direct support for this design principle. This is because our user study lacks a third condition typically used in adaptation studies in SDS (e.g. (Forbes-Riley et al., 2008; Pon-Barry et al., 2006)). This third condition will activate the additional explanations randomly instead of using the appropriate trigger (i.e. the PopUp–Incorrect event). This condition is not required in this work due to our performance analysis perspective: we are given a system, we perform an offline analysis of the system using the discourse structure information, we propose a modification and then we test if the modification improves the system. However, it would be interesting to run additional user studies to test this design principle.

For characterization of dialogue phenomena, the dependencies observed in our corpus between transitions and two dialogue phenomena offer insights into system behavior. The transition–speech recognition problems interactions suggest that particular attention should be paid to specific locations in the discourse structure. The dependencies between transitions and user affect offer insights into what contexts are associated with increase/decrease in user affect. Knowledge of these contexts is an important step towards context-dependent affect adaptation. This knowledge allows us to formulate hypotheses to explain these dependencies which in turn can be used to design appropriate system responses. This direction is currently being pursued by other colleagues (e.g. (Forbes-Riley et al., 2007)).

Some of the transition-phenomena dependencies we observed in the F03 corpus provide additional motivations for the Navigation Map (discussed in Section 6.1). It will be interesting to rerun these analyses on the Main experiment corpus and to see if these dependencies are affected by the presence of the Navigation Map. Besides suggesting that our hypotheses behind these dependencies are true, these analyses will allow us to further understand the effects of the Navigation Map on the system and user behavior.

Our results also suggest that discourse structure transitions might be an informative feature for prediction of phenomena that occur in dialogues. Previous work by other colleagues (Ai et al., 2006) has already used this information source to predict user affect however the study does not explicitly look at the contribution of the discourse transitions. Empirical studies can be run to test if indeed discourse transition features are informative for predicting dialogue phenomena.

Our positive results for the NM motivate further studies to better understand the effect of presenting the discourse structure information on the graphical output. Our results suggest that the simplest way to improve our system is to use the regular version for low pretesters and the NM version for high pretesters. Validation of this approach should be done by running a user study that compares the regular version of the system with this new mixed version.

Another alternative is to have students choose between the two versions instead of assigning them based on the pretest score. Two observations suggest this approach might work. First, due to the observed NM "hand glove" effect, users that rate the system higher also tend to learn higher. While this correlation does not imply a causal link, we hypothesize that we can trust user's perception as an assignment criterion. Second, our data shows that some of the low pretesters in the *NM* condition do achieve higher learning (4 out of the 16 low *NM* pretesters have an NLG of 0.6 or higher (see Figure 39). Since these users also rate the system higher (ALL rating of 3.6 or higher - 6.4.2.3), the new approach might assign them to the *NM* condition instead of the *R* condition as the pretest assignment will do. To test this new approach, a user study can be run in which users try both versions in the first two ITSPOKE problems and then choose the version to be used for the other three problems.

It is interesting to investigate how to make the NM effective for low pretesters. One way to address this problem is informed by our hypothesis that the NM was not effective for low pretesters because it contains "expert" text (the discourse structure intention) instead of "novice" text (the system turn transcript). We can imagine a version of the NM where under each discourse segment intention we also have the corresponding discourse segment text. However, it is not clear the best way to fit this information on the screen. One solution would be to show the discourse segment text whenever a discourse segment is active or the user clicks it.

For our investigation on the objective utility of the NM, we used a version of the NM that has one feature disabled: the information highlight. We disabled this feature because this information source is not part of discourse structure. It would be interesting to see if enabling this feature will further increase some of the NM positive effects. Also, the NM effect might be even stronger for versions of the system that used synthesized speech. To account for potential negative effects of using a synthesized speech output, in our NM experiments we used human prerecorded prompts. It would be interesting to replicate our experiments with synthesized speech output.

The NM uses two discourse structure elements: the discourse segment hierarchy and the discourse structure intentions. It would be interesting to understand which of the two elements is useful for users. We can imagine an experiment that compares a version of the NM that shows only the discourse structure intentions with a version that shows only the discourse segment hierarchy (though it is not clear how to present the indentation without other textual information).

The NM falls under a general framework: graphical representations of entities and of the relations between them. It would be interesting to understand how to best instantiate this framework for a given domain: what are the best entities and relations to present and what is the best way to present these graphically (e.g. (Mancini et al., 2006)).

The NM enables novel ways of interaction between users and systems. Similarly to (Rich and Sidner, 1998)'s segmented interaction history, users can use the NM to directly control the dialogue flow. For example, by clicking on a discourse segment, users can ask the system to pause, restart or replay that segment. In addition, we can provide additional explanations for certain discourse segments or even allow users to skip certain topics (e.g. if they are familiar with that topic or the topic was activated due to semantic interpretation errors). These new features raise additional design issues as, for example, previous studies (e.g. (Baker et al., 2008)) have shown that whenever we allow users to control certain system features, some users tend to abuse or game the system (e.g. abuse of help hints in (Baker et al., 2008)).

In our work, we have investigated the NM for a system that uses system initiative. It would be interesting to study how to construct and use the NM in a system with user initiative. For these systems, the NM requires recognition or inference of user's plan (e.g. (Allen et al., 2000; Blaylock and Allen, 2006)). In the author's unpublished work with Shimei Pan, we looked at such a system: a multimodal system that allows users to browse a large real-estate database (Zhou et al., 2005). Users can browse various entities like houses, cities, school districts, hotels, etc. Our choice of discourse segments for that system was user queries. To build the NM, we also needed the discourse segment hierarchy: i.e. how the queries/discourse segments relate to other queries/discourse segments. To construct this hierarchy we had to infer at runtime the user's exploration plan. Besides visual support, the resulting NM offers additional opportunities for a number of important tasks in such domains: query relaxation, result summarization, query refinement and information push.

Our positive results validate future work in terms of how we extract and use the discourse structure information. While our definition of discourse structure transitions is domain independent and automatically computable (at least in dialogue managers inspired by the Grosz & Sidner theory), it would be interesting to apply it to other annotations of the discourse segment hierarchy or other systems/domains. In this way we can study issues related to the interpretation of discourse structure transitions. For example, one can investigate issues that arise from using a manual annotation of the discourse segment hierarchy or if the 6 transitions we defined are enough in other domains. For the Navigation Map, we can study the reliability of the discourse structure annotation and investigate the right choice of granularity and its impact on the effectiveness of the Navigation Map.

There are many ways to continue the analyses from the Main experiment (4.5.4 and 6.4.2) that compared regular ITSPOKE with a version with the Navigation Map and a version with a new strategy informed by our performance analysis results. While we have looked at a number of metrics (e.g. learning, ratings, dialogue time), we can get a better understanding of the differences between the three versions by looking at other metrics that measure user correctness, speech recognition problems, lexical choice, material covered, etc. Also, an analysis worth pursuing in the future is to understand the differences between low learners and high learners. One way to do this is to perform a split based on the NLG score (similarly to the pretest split: PRE Split - 2.3.3.1). This will result in a low learners subset and a high learners subset. Similarly to the PRE Split investigations, we can study the effect of the NLG Split on specific metrics (e.g. dialogue time, correctness, etc) and speculate based on the results why high learners achieve better learning. This is particularly important as high learners come from both low and high pretesters (i.e. no correlation between PRE and NLG – $R=0.06$, $p<0.57$).

Finally, both the performance analysis application and the user-side application of discourse structure have produced better versions of our system (at least for some users). While here we have focused on comparing these new versions with our regular version of the system (*NM* vs. *R* and *PI* vs. *R*), our experiment design also allows us to compare the two new versions (*NM* vs. *PI* – see 2.2.4). We leave this comparison for future work. Note that the two versions can be easily combined: a new version of ITSPOKE that displays the Navigation Map and uses the new PopUp–Incorrect strategy. It would be interesting to study the utility of this new version and see if it yields further improvements.

APPENDIX A

ADDITIONAL DETAILS ON ITSPOKE EXPERIMENTS

A.1 READING MATERIAL

Before the pretest step, all users read an introductory physics text which describes the basic physics concepts and laws that will be used in the later steps of the experiment. This text was developed by the authors of the Why2-Atlas system and was used in previous Why2-Atlas experiments. For more details see (VanLehn et al., 2007).

The material is available below:

This material is meant to introduce you to the basic physics concepts that you will use later while working through the set of physics problems. These basic concepts are the building blocks that physicists use to construct explanations for how and why objects move the way they do, and how they interact with each other.

Distance and Displacement: In everyday language, the word "distance" has two meanings. For instance, if someone asks you "how far is it from my house to yours?," you might answer, "It's one mile as the crow flies, but if you drive, it is 2 miles." That is, "distance" can mean either the length of some particular path between two points, or it can mean the length of a straight line between two points. (The American idiom "as the crow flies" assumes crows fly in straight lines!) Physicists use different words for the two concepts. In physics, the *distance* between points A and B is the length of some specified path between points A to B, and you have to say what that path is. On the other hand, the *displacement* between A and B refers to the straight line running from A to B. In particular, displacement refers to both the length of straight line between A and B and also to its direction.

Velocity: In everyday language, we can use the words *speed* and *velocity* interchangeably. In physics, we make a distinction between the two. Very simply, the difference is that velocity is speed in a given direction. We say a car travels at 1.1 m/s (meters per second), we are specifying its speed. But if we say a car moves at 1.1 m/s to the north, we are specifying its velocity. Speed is a description of how fast; velocity is how fast and in what direction.

Vector quantities: Some quantities require both magnitude and direction for a complete description. These are called vector quantities. Displacement and velocity are vector quantities, and there are many others.

Scalar quantities: Many quantities in physics, such as mass, volume and time, can be completely specified by their magnitudes. They do not involve any idea of direction. These are called scalar quantities. They obey the ordinary laws of addition, subtraction, multiplication and division. A 3 kg bag of sand ("kg" is an abbreviation for "kilogram") is added to 1 kg of cement, the resulting mixture has a mass of 4 kg. If 5 liters of water is poured from a pitcher which initially had 8 liters of water in it, the resulting volume is 3 liters. In both of these cases, no direction is involved. We see that 10 kilograms north or 5 liters east have no meaning. Quantities that involve only magnitude and not direction are called scalars.

If you understand the difference between scalars and vectors, then it is easy to understand the difference between speed and velocity: speed is a scalar and velocity is a vector. That is, velocity refers to both how fast an object is moving and to the direction of its movement. Similarly, displacement is a vector and distance is a scalar.

Constant velocity: From the definition of velocity it follows that to have a constant velocity requires both constant speed *and* constant direction. Constant speed means that the motion remains at the same speed—the object does not move faster or more slowly. Constant direction means that the motion is in a straight line—the object's path does not curve at all. Motion at constant velocity is motion in a straight line at constant speed.

Acceleration: We can change the velocity of an object either by changing its speed, by changing its direction of motion, or by changing both. The rate at which velocity is changing is called the *acceleration*. Because acceleration is a rate, it is a measure of how fast the velocity is changing per unit of time. Acceleration = change in velocity / time interval.

In physics, the term "acceleration" is used both for decrease as well as increase in speed. The brakes of a car can produce large retardation; that is, they can produce a large decrease in speed per second. This is often called *deceleration* or negative acceleration.

Acceleration applies to changes in *direction* as well as changes in speed. If you ride around a curve at a constant speed of 1.5 m/s, your velocity is not constant because your direction of motion is changing at every instant. Whenever your state of motion is changing, you are accelerating.

Forces: A *force* is any push or pull. If I push a hockey puck with my stick, I exert a force on the puck. If I pull on it with my hand, I am also exerting a force on it. If I squeeze it or attempt to stretch it, I am also exerting forces on it.

Two objects are always involved whenever force is exerted. When I push a car, I am one object and the car is the other. Physicists say that the force "acts on" one of the objects and is "due to" the other. Thus, when I push a car, the force acts on the car and is due to me.

Contact forces: Forces are put into two categories: *contact forces* and *field forces*. If the two objects are pressing against each other, they are involved in a contact interaction, then the force is called a contact force. For example, suppose I keep a helium balloon from flying away by holding its string. I exert a contact force on the string, which I am holding, and the string exerts a contact force on the balloon, which it is touching. I do not exert a contact force on the balloon because I am not touching it. Contact forces only exist when the two objects are in physical contact.

Field forces: Non-contact forces are called field forces. When a planet pulls on an object near it, the pull is called a *gravitational force*. A gravitational force exists even when the planet and the object are not touching. If you drop a ball, then while it is falling, it is not touching the earth yet the earth is pulling on it. Thus, gravitational force is not a contact force. The other common field forces are *electrical force* and *magnetic force*. For instance, two magnets pull or push each other even when they are not touching. Notice that even though the objects involved in field forces are not touching, there are still always two objects. All forces involve interaction between two objects.

Friction: A few contact forces are so important that they have names. *Friction* is the name given to a contact force that acts between materials that are moving past each other while in contact. Friction arises from the surface properties of the sliding objects. If friction were totally absent, a sliding object would need no force whatsoever to continue sliding on a surface with constant velocity.

Newton's first law: Isaac Newton is famous for his three well known laws of motion, and the first one is: "Every body continues in its state of rest, or of motion in a straight line at constant speed, unless it is compelled to change that state by forces exerted upon it." Simply put, things tend to keep on doing what they're already doing unless something is causing it to change. Dishes on a table top, for example, are in a state of rest. They tend to remain at rest, as is evidenced if you snap out a table cloth from beneath them. If an object is in a state of rest, it tends to remain at rest. A force will have to be applied to change that state.

Now consider an object in motion. If you slide a hockey puck along the surface of a city street, the puck quite soon comes to rest. If you slide it along ice, it slides for a longer distance. This is because the friction force on it on the surface of ice is very small. If friction is absent, it slides with no loss in speed. We thus conclude that in the absence of applied forces, a moving object will move in a straight line with constant speed indefinitely.

Mass: Kick an empty tin can and it moves. Kick a can filled with sand, and it doesn't move as much. Kick a tin can filled with solid lead, and you'll only hurt your foot. The lead-filled can has greater resistance to motion than the empty can because it has larger *mass*. *The mass of* an object is a measure of its resistance to changing its motion.

Mass is not volume: Many people confuse mass with volume. They think that if an object has a large mass, it must have a large volume. But volume is a measure of space and is measured in units such as cubic centimeters or liters. Mass is measured in *kilograms*. For anything that is mostly water, like milk, juice or soda pop — one liter of it has a mass of about one kilogram. However, a kilogram of lead would be much smaller. How many kilograms of matter are in an object and how much space is taken up by the object are two different things. Which has the greater mass—an oversized feather pillow or a common automobile battery? Clearly the battery is more difficult to set in motion. This is evidence of the battery's larger mass. The pillow may be bigger—that is, it may have a larger volume—but it has less mass.

Mass is not weight: Now that you have an understanding of field forces, you are ready to understand the distinction between mass and weight. Mass should not be confused with weight. Mass is an intrinsic property of an object; it is determined by the actual material in the body. It depends only on the number and kind of atoms that compose it. Weight is a measure of the gravitational force that acts on the body, and hence depends on where the object is located.

The amount of material in a particular stone is the same whether the stone is located on the earth, on the moon or in outer space. Hence, its mass is the same in any of these locations. This could be shown by shaking the stone back and forth. The same force would be required to shake the stone with the same rhythm whether the stone was on earth, on the moon, or in the force-free region of outer space. That's because the mass of the stone is solely a property of the stone and not its location.

But the weight of the stone would be very different on the earth, on the moon, and in outer space. On the surface of the moon, the stone would have only one-sixth of the weight it has on the earth. This is because the acceleration due to gravity is only one-sixth as strong on the moon as compared to that on the earth. If the stone were in a gravity-free region of space, its weight would be zero. Its mass, on the other hand, would remain the same everywhere.

Near a planet, the mass of an object is directly proportional to the magnitude of its weight (and weight is gravitational force). The constant of proportionality is called g, so $W = mg$, where W is the magnitude of the object's weight, m is the object's mass and g is a constant that depends on the planet. For Earth, $g = 9.8$. For the moon, $g = 1.6$.

Forces produce acceleration: The key idea behind Newton's first law is that it takes a force acting on an object in order to cause a change its motion. Consider an object at rest, such as a hockey puck on a smooth, nearly frictionless ice. Push it with a stick and it begins to move. Its velocity changed from zero in the beginning to some value at the end of the push or applied force. So, it has accelerated—in other words, a change the state of motion of the puck has occurred due to the applied force. When the stick is no longer in contact with the puck, the puck moves at constant velocity. Apply another force by striking it with the stick again, and the motion changes. Again the puck has accelerated. Force produces acceleration. Thus, we see here that force can cause an object to speed up. We have already discussed how friction, which is a force, causes objects to slow down.

Net force: Often, the force we apply is not the only force that acts on an object. Other forces may act as well. The combination of all the forces that act on an object is called the *net force*. It is the net force that produces the acceleration of an object. For instance, suppose you attach a thread to a puck on smooth, nearly frictionless ice. If you pull on the thread, the puck accelerates. If your friend also attaches a thread to the puck and pulls in the same direction you are pulling, then the puck has greater acceleration. That is, acceleration of an object is proportional to the net force acting on it. In this case, the net force is the combination of the two forces exerted on the puck, one due to your thread and the other due to your friend's thread. Now suppose that your friend pulls away from you. In this case, the force your thread exerts is opposite the force that your friend's thread exerts. If the two forces are equally strong, then they cancel each other, so the net force is zero and the puck has zero acceleration. It remains stationary. Thus, acceleration is due to the net force on an object, which is the sum of all the individual forces acting on the object.

Newton's second law: Push on an empty shopping cart, then push equally hard on a heavily loaded shopping cart, and you'll produce much less acceleration in the second case. This is because acceleration depends on the mass of the object being pushed. For objects of greater mass, we find smaller accelerations for the same force. Isaac Newton was the first to realize that the acceleration we produce when we move something depends not only on how hard we push or pull (the force) but on the mass as well. He came up with one of a most important rules of nature ever proposed, his second law of motion: "The acceleration of a body is directly proportional to the magnitude of the net force acting on it and inversely proportional to its mass, and the acceleration is in the same direction as the net force." That is, acceleration = net force / mass.

From this relationship, we can see that if the net force that acts on an object is doubled, the acceleration will be doubled. Suppose instead that the mass is doubled. Then the acceleration will be halved. If both the net force and the mass are doubled, then the acceleration will be unchanged.

Normal force: How many forces act on your book as it lies motionless on the table? Don't say one, its weight. If that were the only force acting on it, you'd find it accelerating. The fact that it is at rest, and not accelerating, is evidence that net force on it is zero. So, another force must be acting in opposite direction. The other force is called "the normal force due to the table acting on the book." The table actually pushes up on the book with the same amount of force that the book presses down. If the book is to be at rest, the sum of the forces acting on it must balance to zero.

Air resistance: Friction is not restricted to solids sliding over one another. One common form of friction is *air resistance*, the friction that acts on something moving through air. You don't normally notice it when walking or jogging, but you'll notice it at higher speeds as in skiing downhill or skydiving. Air resistance is proportional to both the speed and the size of the side of the object that the air is blowing against. For instance, suppose you hold your hand out the window of a speeding car. If your palm faces the front of the car, you'll feel a strong force due to air resistance. If you turn the edge of your hand to face the front of the car, you'll feel less force. This is because air resistance depends on the size and shape of the side that the air strikes against. In general, when an object is moving slowly through the air, or when the air strikes a small size (like the tip of a spear), then the air resistance is so small that it can be ignored as compared to the other forces acting.

Fall through air: Drop a stone and it falls. Does it accelerate while falling? We know it starts from a rest position and gains speed as it falls. We know this because it would be safe to catch if it fell a meter or two, but not from the top of a tall building. Thus, the stone must gain more speed during the time it drops from a building than during the shorter time it takes to drop one meter. This gain in speed indicates that the stone accelerates as it falls.

Gravitational force causes the stone to fall downward once it is dropped. Air resistance tends to slow it down, but if the stone is still moving slowly, then air resistance can be ignored, and the gravitational force is the only force acting on the stone. Whenever gravitational force is the only force acting on an object, the object is said to be in *free fall*.

Newton's third law: In the simplest sense, a force is a push or a pull. Looking closer, however, we find that a force is not a thing in itself, but is due to the interaction between one thing and another. One force is called the *action force*. The other is called the *reaction force*. It doesn't matter which force we call *action* and which we call *reaction*. The important thing is that neither force exists without the other. The action and reaction forces make up a pair of forces.

In every interaction, forces always occur in pairs. For example, in walking across the floor you push against the floor, and the floor in turn pushes against you. Likewise, the tires of a car push against the road, and the road in turn pushes back on the tires. In swimming you push the water backward, and at the same time the water pushes you forward. There is a pair of forces acting in each instance.

Coordinate axes: When we deal with vectors it is necessary to define your coordinate axes so as to define the directions of the vector quantities. Coordinate axes are two mutually perpendicular axes, which we will refer to as the x- and the y-axes. Often we are trying to analyze the motion of an object that moves both horizontally and vertically, such as a cannon ball shot at an angle. If we choose the coordinate axes so that the x-axis is horizontal and the y-axis is vertical, the analysis is much easier.

Components of vectors: Any single vector can be regarded as the sum of two vectors, each of which acts on the body in some direction other than that of the given vector. These two vectors are known as the *components* of the given vector that they replace.

A man pushing a lawnmower applies a force that pushes the machine forward and also against the ground. In Figure 6-10, vector F represents the force applied by the man. We can separate this force into two components. Vector Y is the vertical component, which is the downward push against the ground. Vector X is the horizontal component, which is the forward force that moves the lawnmower.

The rule for finding the vertical and horizontal components of any vector is relatively simple, and is illustrated in Figure 6-11. A vector V is drawn in the proper direction to represent the force, velocity or whatever vector is in question (Figure 6-11 left). Then vertical and horizontal lines are drawn at the tail of the vector (Figure 6-11 right). A rectangle is drawn that encloses the vector V in such a way that V is the diagonal and the sides of the rectangle are the desired components. We see that the components of the vector V are then represented in the direction and magnitude of the vectors X and Y.

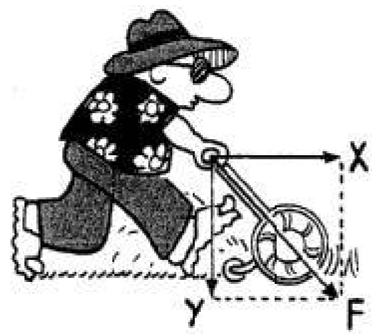


Fig. 6-10 The force F applied to the lawnmower may be resolved into a horizontal component, X, and a vertical component, Y.

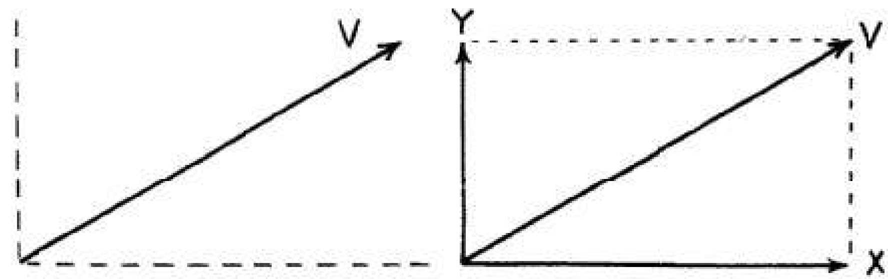


Fig. 6-11 The vector V has component vectors X and Y.

A.2 ITSPOKE PROBLEMS

ITSPOKE discusses 5 problems with users. These 5 problems are part of a larger set of 10 problems used in previous Why2-Atlas experiments with both human and computer tutors. For more details see (VanLehn et al., 2007).

We present the problems below in the order in which they are discussed by ITSPOKE:

1. 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.
2. An airplane flying horizontally drops a packet when it is directly above the center of a swimming pool. Does the packet hit that spot? Explain
3. The sun pulls on the earth with the force of gravity and causes the earth to move in orbit around the sun. Does the earth pull equally on the sun? Defend your answer.
4. Suppose a man is running in a straight line at constant speed. He throws a pumpkin straight up. Where will it land? Explain.
5. Suppose a lightweight car and a massive truck hit a patch of frictionless ice and have a head-on collision. Upon which vehicle is the impact force greater? Which vehicle undergoes the greater change in its motion? Defend your answers.

In general, each problem builds upon previous ones. For example, in the first problem we have an object that starts with a zero horizontal and vertical velocity and is in freefall. The second problem extends this by adding a constant non-zero horizontal velocity. The fourth problem extends even further by discussing a situation where the object has a constant horizontal velocity and it is first accelerated up (i.e. an upward vertical velocity) and then it freefalls. These three problems make heavy use of the 1st and 2nd Newton's law. The third problem, which is the smallest in terms of instruction, introduces the 3rd Newton's law. The last problem, makes use of the 3rd law again but in a different situation and offers a new way of applying the 2nd law compared with problem 1, 2 and 4.

A.3 PRETEST & POSTTEST

All ITSPOKE experiments use a pretest/posttest with 4 essay questions (open-answer) and 26 multiple-choice questions. For each test, the 26 multiple-choice questions are part of a larger set of 40 multiple-choice questions used in previous Why2-Atlas experiments with both human and computer tutors[40]. For more details about these tests see (VanLehn et al., 2007). The 26 multiple-choice questions[41] correspond to the 5 problems discussed by ITSPOKE; the rest correspond to the extra 5 problems not discussed by ITSPOKE (see A.2).

We present below the questions from the pretest (A.3.1) and the posttest (A.3.2) in the order in which the student sees them. But before that, we would like to acknowledge a number of factors that can influence user performance on these tests:

- *Knowledge of physics* – while all users were required to have not taken a college level physics course, some users have more exposure than other to this subject through their high-

[40] Although the pretest and the posttest as we present them here were designed to be isomorphic, earlier Why2-Atlas experiments have alternated which one of the tests was given as the pretest/posttest in the experiment (i.e. half the users got them in the same order as in our experiments while the other half had the tests swapped). This alternation was needed to ensure observed learning was not due to one test being easier than the other. For more details see (VanLehn et al., 2007).

[41] In the S05 experiment (2.2.2) the full set of 40 multiple-choice questions was used. For a fair comparison with other ITSPOKE experiments, the pretest and posttest scores (2.3.1.1) were computed on the 26 question subset. However, current work by other colleagues (Ward and Litman, 2008) makes use of the 14 extra questions by interpreting them as far-transfer questions.

school curricula or personal interests. This factor can have big effect on the PRE score primarily.

- *Academic background* – users in our experiments come from a variety of academic backgrounds ranging from science to humanities studies. Since particular academic backgrounds require and develop different sets of skills, this factor can have an impact on user's ability to learn a technical domain, affecting both PRE and POST scores
- *Motivation* – plays an important role as it determines how involved the user is in the whole experiment process. Although users were paid by hour, there is no other external reward for their performance (e.g. course grade contribution). The factor can have a big impact on the PRE/POST scores
- *Reading effort* – in all experiments, users read an introductory material before the pretest. The amount of effort, in terms of concentration and time, users put in the reading step can have a big impact on the PRE score and as a result on the measured learning performance.
- *Test-taking effort* – the amount of time and concentration users dedicate to the two tests can have a huge impact on their performance. Users that do not concentrate enough can miss important details that can influence the correctness of their answer. As a result some can resort to a "click-through" behavior as described in Section 2.2.4.1.
- *Fatigue* – in all experiments users spend a considerable amount of time (between 2 and 5 hours). As a result, users are naturally more fatigued at the end of the experiment. This in particular can affect the POST score.

Ideally, we would like to have experimental conditions balanced on all these factors. Practically, this is not possible and it is an inherent issue of all smaller size tutoring studies. However, it can be noted that random assignment to conditions produces relatively similar populations in terms of PRE averages and standard deviation (recall Table 5).

A.3.1 Pretest questions

Essay questions

Question: A cat walks to the roof-edge of a building and just drops off. She takes 20 seconds to land on the ground. How does her velocity after the first half period of the fall compare with that on landing (air resistance is negligible)? Explain.

Question: You observe two rocket ships in deep space where they are weightless. They have the same motors, which are generating exactly the same thrust force, but one ship is speeding up quickly while the other is speeding up slowly. Explain how this can be true.

Question: A diver jumps off a high platform. During the fall, he is being pulled down by the force of earth's gravity. Is there a force on earth due to the diver? If so, what is the earth's acceleration? Explain, stating principles of physics involved.

Question: A layer of water on the roof of a high building is frozen so as to provide a smooth icy horizontal surface. Two small closed boxes, one empty and the other filled with sand, are pushed such that they have equal velocity at the instant of falling off the edge. A little later, they land on the flat horizontal ground below. How will the distances of their landing points from the foot of the building be related? Explain, stating principles of physics involved.

Multiple choice questions

Question: A boy tosses a rock off a cliff with an initial velocity, Vi, in the horizontal direction. Assuming air resistance is negligible, what is true of the horizontal component of the velocity of the rock while it is falling?

Option 1: It will increase.
Option 2: It will decrease.
Option 3: It will remain the same.
Option 4: there is not enough information to answer

Question: Suppose that a kangaroo maintains a constant horizontal velocity despite the fact that it runs by bouncing along. Suppose you are driving your LandRover and pull along side a kangaroo that is bouncing along in a straight line. Just then you get a call on your cell phone. You maintain your speed, ignoring the kangaroo (which also ignores you and keeps bouncing along). At the end of your call, you look out the window, what should you see?

Option 1: The kangaroo has pulled ahead of you
Option 2: the kangaroo is bouncing along beside you
Option 3: the kangaroo has fallen behind
Option 4: other

Question: A motorcycle is driving west on a flat road at 10 m/s. A car is driving west down a hill sloped at 45 degrees. The western component of the car's velocity is 10 m/s. How does the horizontal displacement of the car compare to the horizontal displacement of the motorcycle at any time?

Option 1: they are the same
Option 2: the horizontal displacement of the car is greater than that of the motorcycle
Option 3: the horizontal displacement of the motorcycle is greater that that of the car
Option 4: there is not enough information to answer

Question: Suppose a drunken parachutist dons his backpack, which is full of books instead of his parachute, then jumps from the plane. When he realizes his mistake, he throws the backpack away from him as hard as he can. When both backpack and man are still falling, but seperately now, which one(s) is/are in freefall?

Option 1: the man
Option 2: the backpack
Option 3: both
Option 4: neither
Option 5: other

Question: A child rides down a hill on a sled. The sled is moving at a constant speed of 3 m/s. What is the speed of the child?

Option 1: 3 m/s
Option 2: a little faster than 3 m/s
Option 3: a little slower than 3 m/s
Option 4: there is not enough information
Option 5: other

Question: Alice and Bob decide to race each other a short distance. Starting at rest, they accelerate at

the same rate, R, until they reach the finish point a few seconds later. Who wins the race?
Option 1: Alice
Option 2: Bob
Option 3: neither, it's a tie
Option 4: there is not enough information to answer

Question: A ball is launched horizontally from the top floor of a building and moves with only the earth's gravitational force acting on it. In which direction does gravity act on the ball?
Option 1: vertically down, toward the center of the earth
Option 2: initially horizontal, then increasingly vertical and finally completely vertical
Option 3: almost vertically down, but slightly angled due to the rotation of the earth
Option 4: in the direction of the motion of the ball
Option 5: other

Question: Two billiard balls of equal size and weight initially traveling in opposite directions collide with each other. The magnitude of the force the first ball exerts on the second is 3 N. What is the magnitude of the force of the second ball on the first?
Option 1: 3 N
Option 2: less than 3 N
Option 3: More information is necessary to answer this question

Question: Two billiard balls of equal size and weight initially traveling in opposite directions collide with each other. The magnitude of the force the first ball exerts on t6he second is 3 N. If the direction of the force of the first ball on the second ball is to the right, what is the direction of the force of the second ball on the first?
Option 1: Also to the right
Option 2: To the left
Option 3: More information is necessary to answer this question

Question: You are on a train that is not yet moving. Before the train starts up, you set a bowling ball on the floor. Since the train is level, the bowling ball does not roll. Now the train starts up, and steadily gains speed. Do you have to exert a force on the ball to get it to stay still relative to the train?
Option 1: Yes, it will roll toward the front of the train if I don't
Option 2: Yes, it will roll toward the back of the train if I don't
Option 3: No force is necessary. It moves along with the train just like me.
Option 4: Other:

Question: A long-jumper runs at full speed, and jumps from the take-off point and a second later lands in the sand. While he is in the air during his jump, in which direction is his acceleration due to the earth's gravitational force?
Option 1: vertically downward
Option 2: both vertically downward and horizontally in the direction he is moving in
Option 3: only horizontally in the direction he is moving in
Option 4: he is not accelerating due to the earth's gravity while he is in the air

Question: A hot-air balloon moves vertically downward with a constant speed of 3 m/s. What is the magnitude of the acceleration of the hot-air balloon?
Option 1: 0 m/s^2
Option 2: 3 m/s^2
Option 3: 9.8 m/s^2
Option 4: 12.8 m/s^2
Option 5: Other:

Question: A hot-air balloon of mass, mb, moves vertically downward with a constant speed of 3 m/s. What is the magnitude of the net force acting on the hot-air balloon while it is descending at this constant speed?

Option 1: 0 Newtons
Option 2: mb*g (where g is the acceleration due to gravity) Newtons
Option 3: 3*mb Newtons
Option 4: 12.8*mb Newtons
Option 5: Other:

Question: A Navy Seal on a rescue mission steps out of a helicopter and at the exact moment that rescue supplies are dropped from rest out of the helicopter. The Navy Seal is heavier than the rescue supplies. Both are in freefall until they land in the ocean, below. Which of the following is true?
Option 1: the Navy Seal will hit the water first
Option 2: the supply kit will hit the water first
Option 3: the Navy Seal and supply kit will hit the water at the same time
Option 4: there is not enough information to answer this question

Question: You toss a coin straight up into the air by applying a force on the coin with your hand. Does the force from you hand continue to act on the coin immediately after you release the coin into the air?
Option 1: Yes. The coin will still feel the force from your hand.
Option 2: No. The coin will only feel the force from your hand while your hand is pushing on it.
Option 3: Yes, but only for a short while. After a few seconds, the force from the hand has dissipated.

Question: You toss a coin straight up into the air by applying a force on the coin with your hand. When the coin is halfway between where it was first thrown and its maximum height (and still moving upward), which force(s) act on it?
Option 1: Gravity
Option 2: The force of the throw
Option 3: There are no forces acting on it
Option 4: Both gravity and the force of the throw

Question: Two metal balls are the same size but one weights twice as much as the other. The balls are dropped from the roof of a single story building at the same instant. What is the relationship between the accelerations of the heavier and lighter metal balls during the fall?
Option 1: the acceleration of the heavier ball is the same as that of the lighter ball
Option 2: the acceleration of the heavier ball is the same as that of the lighter ball
Option 3: the acceleration of the heavier ball is greater than that of the lighter ball
Option 4: the acceleration of the heavier ball is less than that of the lighter ball

Question: In the figure below, an object is moving in the positive x-direction at 5 m/s when a force, F, begins to act on the object in the positive y-direction.. What is the x-component of the velocity of the object 5 seconds after the force begins to act on the object?
Option 1: greater than 5 m/s
Option 2: less than 5 m/s
Option 3: 5 m/s
Option 4: Insufficient information to tell

Question: As a truck moves along the highway at constant speed, a nut falls from a tree and smashes into the truck's windshield. If the truck exerts a 1,000 N force on the nut, what is the magnitude of the force that the nut exerts on the truck?
Option 1: 1,000 N
Option 2: less than 1,000 N
Option 3: 0 N (the nut does not exert a force on the truck)
Option 4: greater than 1,000 N (because the nut hit the truck, it exerts a greater force on the truck than the truck exerts on the nut)

Question: As a truck moves along the highway at a constant speed, a nut falls from a tree and smashes into the truck's windshield. During the impact, which is true of the relationship between the magnitudes of the acceleration of the nut and the truck?

Option 1: the acceleration of the nut is greater than the acceleration of the truck
Option 2: the acceleration of the truck is greater than the acceleration of the nut
Option 3: the accelerations of the truck and nut are equal
Option 4: there is not enough information

Question: A car driving along a straight and flat (horizontal) road comes upon a pool of oil on the road. While driving through the oil, the frictional force between the car's tires and road is zero and there are no forces acting on the car in the horizontal direction (including the force from the car's engine). What happens to the horizontal speed of the car while it is moving over the oil slick?
Option 1: because there are no forces acting on the car in the horizontal direction, the car's horizontal speed decreases
Option 2: because there are no forces acting on the car in the horizontal direction, the car's horizontal speed remains constant.

Question: A television satellite maintains a fixed height of 5 miles above the earth's surface and remains directly above a small town. Does the earth's gravity act on the satellite?
Option 1: Yes. The earth's gravity acts on everything near it's surface
Option 2: Because the satellite is orbiting the earth, it is accelerating toward the earth; thus the earth does not exert a gravitational force on the satellite
Option 3: Because the satellite is so high above the earth's surface, it does not experience the earth's gravitational force

Question: A distant planet orbits a nearby star. Which of the following statements is true:
Option 1: The star exerts a gravitational force on the planet, but the planet does not exert a gravitational force on the star
Option 2: Both the star and the planet exert a gravitational force on the other, but the force of the star on the planet is greater than the force of the planet on the star
Option 3: Both the star and the planet exert a gravitational force on the other, and the gravitational force of the planet on the star is the same as the force of the star on the planet

Question: An oceanliner traveling due east collides with a much smaller yacht, traveling due west. During the collision, the front end of the yacht is smashed in (causing the yacht to sink and the passengers to evacuate to their lifeboat). The oceanliner merely suffered a dent. Which is true of the relationship between the force of the oceanliner on the yacht and the force of the yacht on the oceanliner?
Option 1: because the yacht's acceleration during the collision was greater then the oceanliner's acceleration, the force of the yacht on the oceanliner is greater than the force of the oceanliner on the yacht
Option 2: the force of the oceanliner on the yacht is greater than the force of the yacht on the oceanliner
Option 3: the force of the oceanliner on the yacht is equal to the force of the yacht on the oceanliner

Question: You are in a seat in a rollercoaster when it accelerates forward, causing you to be pressed against the back of your seat. While the rollercoaster accelerates forward and you are pressed against the back of your chair, which of the following is true:
Option 1: there is a force on you in the forward direction
Option 2: there is a force on you in the backward direction (opposite the direction you are moving in)
Option 3: there are no forces acting on you

Question: You toss a coin straight up into the air by applying a force on the coin with your hand. When the coin reaches it's maximum height, which force(s) are acting on it?
Option 1: gravity
Option 2: the force of the throw
Option 3: there are no forces acting on it
Option 4: both gravity and the force of the throw

A.3.2 Posttest questions

Essay questions

Question: On the surface of the moon, a steel ball and a cotton ball are dropped at the same time from the same height. What would be the relation between the velocity they, respectively, acquire on reaching the moon's surface?

Question: The Olympic rocket sled competition of 3002 is held in deep space, where gravity is negligible, and there is no air resistance. The qualifying race is simply to start at rest, race exactly 100 km in a straight line, turn around, and return. One of the sled riders, Barry, gets sick at the last moment. He needs to choose a replacement rider from among his two friends, Maurice and Little Joe. Maurice is much larger than Little Joe, and on earth Maurice would weigh more. However, they both fit inside the sled easily and they are equally skilled sled riders. Does it matter which rider Barry chooses? Explain.

Question: A hiker claims that she can get out of any difficult situation! As a challenge, she is picked up by a helicopter and put in the middle of a frozen icy pond and asked to reach the edge of the pond. The ice is so smooth and frictionless that when she tries to walk, her feet slide on the ice but her body stays where it is. She does some quick thinking, and then throws her helmet away, horizontally, as hard as she can. Will this help her get to the shore? Explain.

Question: The driver of a speeding truck finds that the truck brakes have failed just as he approaches the edge of a cliff. Rather than fly off the cliff in his truck, he opens the door and jumps out horizontally and perpendicular to the direction of the trucks motion just as the truck reaches the cliff-edge. Is he expected to land on the cliff? Explain.

Multiple choice questions

Question: A girl riding a bike in a straight line at a constant speed drops an ice cream cone she was holding. Immediately after she drops the ice cream cone, what is the relationship between the horizontal speed of the girl and the horizontal speed of the cone? (Assume air resistance is negligible)

Option 1: the horizontal speed of the girl is greater than the horizontal speed of the cone
Option 2: the horizontal speed of the cone is greater than the horizontal speed of the girl
Option 3: the horizontal speed of the girl is the same as the horizontal speed of the cone
Option 4: there is not enough information to answer

Question: A frustrated programmer throws her laptop out of the window of a tall building with an initial velocity, Vi, in the horizontal direction. Assuming air resistance is negligible, what is true of the horizontal component of velocity of the laptop while it is falling?

Option 1: it will increase
Option 2: it will decrease
Option 3: it will remain the same
Option 4: there is not enough information to answer

Question: Suppose that a rollerblader is skating down a city street and maintains a constant horizontal velocity. You pull alongside the rollerblader in your car. Just then you get a call on you cell phone. You maintain you speed, ignoring the rollerblader. At the end of your call, you look out the window. What should you see?

Option 1: The rollerblader has pulled ahead of you
Option 2: The rollerblader is skating along beside you
Option 3: the rollerblader has fallen behind
Option 4: Other:

Question: A woman carrying her groceries home is walking north at 1 m/s. A jogger is moving northeast; the northern component of his velocity is 1 m/s. How does the norther displacement of the jogger compare to the norther displacement of the woman at any time?

Option 1: they are the same

Option 2: the northern displacement of the jogger is greater than that of the woman
Option 3: the northern displacement of the woman is greater than that of the jogger
Option 4: there is not enough information to answer

Question: A model rocket is launched vertically. When it is 1,000 meters above the ground it loses all engine power and breaks into two pieces, the front end and the rear end, which has the failed engine. After the engine power is lost which of the two pieces of the rocket is/are in free fall?
Option 1: the front end
Option 2: the rear end
Option 3: both ends
Option 4: neither end
Option 5: other:

Question: A stuntwoman drives a motorcycle up a ramp and jumps over a row of cars and lands safely, remaining on her bike. When it lands, the motorcycle is moving at 20 m/s. What is the speed of the stuntwoman when she lands?
Option 1: 20 m/s
Option 2: a little faster than 20m/s
Option 3: a little slower than 20 m/s
Option 4: there is not enough information
Option 5: Other:

Question: An old Volkswagon has a maximum acceleration of 2 m/s^2 as it starts from a stop sign; a cyclist can also accelerate at a maximum rate of 2 m/s^2. Starting at rest, the Volkswagon and cyclist accelerate at their maximum rates for the same time. Which has traveled farther in that time?
Option 1: the Volkswagon
Option 2: the cyclist
Option 3: they travel the same distance
Option 4: there is not enough information to answer

Question: An airplane is taking off from the ground. In which direction does gravity act on the airplane while it is taking off?
Option 1: vertically down, or toward the center of the earth
Option 2: initially horizontally, then increasingly vertical
Option 3: almost vertically down, but slightly angled due to the rotation of the earth
Option 4: in the direction of the motion of the airplane
Option 5: Other:

Question: Two rugby players, Collin and Ewan, of the same mass, are running toward each other and collide with each other. The magnitude of the force of Collin on Ewan is 500 N. What is the magnitude of the force of Ewan on Collin?
Option 1: 500 N
Option 2: less then 500 N
Option 3: more information is necessary to answer this question

Question: Two rugby players, Collin and Ewan, of the same mass, are running toward each other and collide with each other. The magnitude of the force of Collin on Ewan is 500 N. If the direction of the force Collin exerts on Ewan is to the east, what is the direction of the force that Ewan exerts on Collin?
Option 1: Also to the east
Option 2: to the west
Option 3: more information is necessary to answer this question

Question: A wild horse runs at full speed, and jumps over the fence that had been trapping him. While he is in the air during his jump, in which direction is his acceleration due to the earth's gravitation force?
Option 1: vertically downward
Option 2: both vertically downward and horizontally in the direction the horse is moving

Option 3: only horizontally in the direction the horse is moving
Option 4: he is not accelerating due to the earth's gravity while the horse is in the air

Question: An empty chair lift of mass, mcl, is moving up a hill at a 45-degree angle with a constant speed of 2 m/s. What is the magnitude of the acceleration of the chair lift?
Option 1: 0 m/s^2
Option 2: 2 m/s^2
Option 3: 9.8 m/s^2
Option 4: 11.8 m/s^2
Option 5: Other:

Question: An empty chair lift of mass, mcl, is moving up a hill at a 45-degree angle with a constant speed of 2 m/s. What is the magnitude of the net force acting on the chair-lift while it is moving up the hill?
Option 1: 0 Newtons
Option 2: mccl * g (where g is the acceleration due to gravity) Newtons
Option 3: 2 * mb Newtons
Option 4: 11.8 * mb Newtons
Option 5: Other:

Question: A stuntman jumps horizontally off the edge of a tall building and freefalls until he lands safely on the foam mats below. At the moment the stuntman stps off the building, a stunt dog also jumps from the edge of the same building (and, being a professional, also lands safely). Which of the following is true?
Option 1: the stuntman will land first
Option 2: the stunt dog will land first
Option 3: the man and the dog will land at the same time
Option 4: there is not enough information to answer the question

Question: Using a slingshot, you shoot a stone straight up into the air. Does the force from the slingshot continue to act on the stone after the stone leaves the slingshot?
Option 1: Yes. The stone will still feel the force from the slingshot.
Option 2: No. The stone will only feel the force from the slingshot while the slingshot is pushing on it.
Option 3: Yes, but only for a short while. After a few seconds the force from the slingshot has dissipated.

Question: Using a slingshot, you shoot a stone straight up into the air. When the stone is halfway between where it was first shot and it's maximum height (and is still moving upward), which force(s) act on it?
Option 1: gravity
Option 2: the force of the slingshot
Option 3: there are no forces acting on it
Option 4: both gravity and the force of the throw

Question: Using a slingshot, you shoot a stone straight up into the air. When the stone shot from the slingshot reaches it's maximum height, which force(s) are acting on it?
Option 1: gravity
Option 2: the force of the slingshot
Option 3: there are no forces acting on it at that instant
Option 4: both gravity and the force of the slingshot

Question: Assume you have two balloons: one is filled with water and the other is filled with oil. The balloon filled with oil weighs less than the balloon filled with water, but they are identically shaped. You drop both balloons from a bridge at the same time. What is the relationship between the accelerations of the heavier (water-filled) and the lighter (oil-filled) balloons as they fall to the river below?
Option 1: the acceleration of the heavier balloon is the same as that of the lighter balloon
Option 2: the acceleration of the heavier balloon is greater than that of the lighter balloon

Option 3: the acceleration of the heavier balloon is less than that of the lighter balloon
Option 4: the acceleration of the heavier balloon is the same as that of the lighter balloon
Option 5: the acceleration of the heavier balloon is greater than that of the lighter balloon

Question: A block slides in a straight line across a flat frictionless surface at a speed of 2 m/s. When the block is moving due east, a strong force begins to push on the block to the north. Which is true of the speed of the block in the eastern direction several seconds after the force begins to act on it?
Option 1: it is greater than 2 m/s
Option 2: it is less than 2 m/s
Option 3: it is 2 m/s
Option 4: insufficient information to tell

Question: a 100 kg man is skating clockwise around a skating rink at 3 m/s. A small 30 kg child is skating counter-clockwise (in the opposite direction from everyone else) around the skating rink at a speed of 2 m/s. Neither are paying attention and the child runs into the man. The magnitude of the force the man exerts on the child is 30 N. What is the magnitude of the force the child exerts on the man?
Option 1: 30 N
Option 2: less than 30 N
Option 3: 0 N (the child does not exert a force on the man)
Option 4: greater than 30 N (because the child hit hte man, he exerts a greater force on the man than the man exerts on him)

Question: a 100 kg man is skating clockwise around a skating rink at 3 m/s. A small 30 kg child is skating counter-clockwise (in the opposite direction from everyone else) around the skating rink at a speed of 2 m/s. Neither are paying attention and the child runs into the man. The magnitude of the force the man exerts on the child is 30 N. During the collision between the man and child, which is true of the relationship between the magnitudes of the acceleration of the man and the acceleration of the child?
Option 1: the acceleration of the man is greater than the acceleration of the child
Option 2: the acceleration of the child is greater than the acceleration of the man
Option 3: the accelerations of the man and child are equal
Option 4: there is not enough information

Question: A cross-country skier is skiing horizontally across the snow and before realizing it, has skied onto a large patch of ice. The ice is very slippery and her skis are waxed, so that there is no frictional force between the skis and ice, and there are no other forces acting on the skier in the horizontal direction. What happens to the horizontal speed of the skier while she is on the ice patch?
Option 1: because there are no forces acting on the skier in the horizontal direction, the skier's horizontal speed decreases
Option 2: because there are no forces acting on the skier in the horizontal direction, the skier's horizontal speed remains constant

Question: A research satellite orbits the earth at a fixed distance of 100 miles from the earth's surface. Does the earth's gravity act on the satellite?
Option 1: Yes. The earth's gravity acts on everything near it's surface
Option 2: Because the satellite is orbiting the earth, it is accelerating toward the earth; thus the earth does not exert a gravitational force on the satellite
Option 3: Because the satellite is so high above the earth's surface, it does not experience the earth's gravitational force

Question: A distant planet has a moon that orbits around it. Which of the following statements is true:
Option 1: The planet exerts a gravitational force on the moon, but the moon does not exert a gravittational force on the planet
Option 2: Both the planet and the moon exert a gravitational force on the other, but the force of the planet on the moon is greater than the force of the moon on the planet
Option 3: Both the planet and the moon exert a gravitational force on the other, and the gravitational force of the moon on the planet is the same as the force of the planet on the moon

Question: In the filming of Mission Impossible 4, Tom Cruise insists on performing his own stunt, which involves riding a 200-kg motorcycle straight into a 100,000-kg tractor-trailer traveling in the opposite direction, then jumping off a moment before the motorcycle and tractor-trailer collide. The motorcycle is smashed to half it's length during the collision, whereas the tractor-trailer's front end is merely dented in a few inches. Which of the following is true of the relationship between the force of the motorcycle and the tractor-trailer and the force of the tractor-trailer on the motorcycle?
Option 1: Because the motorcycles's acceleration during the collision was greater than the tractor-trailer's acceleration, the force of the motorcycle on the tractor-trailer is greater than the force of the tractor-trailer on the motorcycle
Option 2: the force of the tractor-trailer on the motorcycle is greater than the force of the motorcycle on the tractor-trailer
Option 3: the force of the tractor-trailer on the motorcycle is equal to the force of the motorcycle on the tractor-trailer

Question: You are sitting on a subway train facing the wrong way (facing the back of the train, which is west) when the train accelerates east and the upper half of your body lurches west (in the direction you are facing) while you remain on your seat. While you lurch west, which of the following is true:
Option 1: there is a westward force on you, in the same direction you lurch
Option 2: there is an eastward force on you (opposite the direction you lurch)
Option 3: there are no forces acting on you

A.4 MAIN EXPERIMENT DETAILS

This section contains additional materials for the Main experiment (2.2.4).

A.4.1 Instruction transcript

Below are the approximate verbal instructions given to users during the Main experiment by the experimenter (the author). The action performed during some of these instructions is shown between angular parentheses ("<>")

Experiment Overview instructions

<Given while showing a printout of slides designed for this instruction >

"There will be 5 steps today. Shortly you will take a memory test. Next you will do some reading. After that you will have a first physics test, then you will work with our computer tutor and at the end there will be another physics test. The important thing is that you need to call me after each of these steps so I can advance you to the next step. In the reading part you will be reading a webpage that explains basic physics concepts like force acceleration, physics laws and some examples. The first test is made out of 26 questions, most of them multiple choice. You will have 4 open answer questions at the beginning of the test. You will work with our computer tutor, its name is ITSPOKE, through 5 problems and it's going to be a speech-based conversation: you will wear these headphones and you will hear the system talking and asking you questions and you will have to answer back. At the end, you are going to have the second physics test which is similar to the first one: again 26 questions most of them multiple choice. And actually right at the end there will be a very short, informal discussion. We want to know what was your experience with the system: things you liked, things you did not like and if you have any suggestions for improvement.

If there are any problems during the experiment, please let me know."

Memory Test instructions

"As I said, the first step is a memory test. Let me explain you the procedure for this test. You will see sentences on the screen and I am going to ask you to read them aloud and try to remember the last word in the sentence. We'll do a practice shortly. After you read a couple of sentences, I am going to ask you to recall the last word in each of those sentences. You should try to remember the words in order, however if you can not it is fine, there is no penalty but there is one restriction: please don't start with the word from the last sentence. Let's do a practice:"

<Go through the first example with the user>

"Note that you have to start reading as soon as the sentence appears on the screen. You noticed that the last part is not there initially, but that is intentional and is going to be there by the time you get to read it."

<Go through the second example with the user>

"Now lets start the actual test. We will start with sets of two sentences just as in the examples, we're gonna do 5 of those and then I am going to add more sentences. Now, this test sounds simple, but it will get tricky very fast. Please do not be stressed about it, you are not going to be paid based on your performance. Just try to do as good as you can. Let's start!"

<Start test>

ITSPOKE instructions

<given while showing the slides for each version of the system>

Instructions common to all versions of the system

"In this step you will work with our computer tutor. This is how the information is going to look on the screen while you work with the system. I will explain you what each of this boxes does and how you will work with the system. The system is going to give you verbal instructions in terms of what you need to do next but just in case you forget, read the content of this small box <point to the small bottom-left box>. The system will give you a problem, you can read the text in this box <point to the top-right box>. You are asked to type an answer to this problem in this box <point to the essay box>. We call this answer an essay. After you are done with your essay, please press the submit button. This will send the essay to the system and the system is gonna start a discussion with you.

You will hear the system talking and asking you questions and you have to answer back. There are two limitations here. First, you can not ask questions back. And second, you can not interrupt the tutor. That is, even if you know the answer for a question, you have to wait for the tutor to finish speaking otherwise is not going to register you."

Further instructions for the `R` and `PI` versions

"This box here is something we call the dialogue history and it shows you the information so far. Whenever the system says something you will be able to see the text in the box under "Tutor said:". So it is like the tutor reading this text for you. You will also be able to see what you said. Note that we have a program that tries to register/recognize the words you are saying. As you might expect it might not work perfectly always. But anyway, you can see what the system recognized from what you said under "You said:". Now, during the conversation, this box will fill up with text and you will be able to access earlier text by using the regular scroll bar."

Further instructions for the `PI` version

"If the system wants to give you additional information, this interface will change with this one <show a sample additional explanations webpage> and here you will have new material about the question. You will be able to see the question, the correct answer to the question, a graphical outline of what you discussed so far and how you tried to find the correct answer to the question. Below there is a text version of that under "What did we learn so far?" and a restatement of the answer under "How did we try to find the correct answer to this question?". You can also find additional information. For example you can find what is the intuition behind the correct answer, what is the relevance of this question for the solution, in other words why are you being asked this question. Sometimes you might find comparisons with previous problems, that is, if a similar question was present in a previous problem what do they have in common and if the answers are the same and why. You might also find possible pitfalls: why you might give an incorrect answer. Now, some of this information might be redundant and you might know it. So use your judgment to decide which part you should read.

When you are done, if you found something useful please press the Yes button, otherwise press the No button. You are basically answering the question "Was this information useful?". Pressing any of the buttons will return you to the conversation."

Further instructions for the NM version

“This box is something we call the Navigation Map. It is an outline of the conversation. The same way you have an outline for a book, this will be an outline for the conversation. Basically whenever the system talks about a topic, you will see a summary of that topic here. The Navigation Map will follow the conversation, so whenever the system moves to a new topic, you will see the change on the screen too. You can see older topics in the NM above this line and a brief description of the topics that you will discuss in the future, under the line. Below the Navigation map you will find what you said. Note that we have a program that tries to register/recognize the words you are saying. As you might expect it might not work perfectly. But anyway, you can see what the system recognized from what you said under the NM here. Whenever a system is done you will be able to see the correct answer in the NM (you might know it already by that time, but just for your reference). Also, you will see in the demo, older topics are automatically closed but you can access that text by clicking on this small lever icon.”

Further instructions common to all versions of the system

“At the end of the conversation the system is gonna ask you to update your essay. So based on the conversation if you thing you want to make any changes, add something to it, please do that. Then you will send it back to the system and the system will move you to the next problem. In this way you are going to solve a total of 5 problems.

At the end a new windows will pop up and you will have a survey. We basically want you to rate your experience with our system. You will see a number of statements like “It was easy to learn from the tutor” and you need to indicate your agreement with that statement.”

<Demo of the appropriate system version>

“There are two more things I have to tell you. First, it might take some time for the system to analyze your essay. But just in case it takes longer that 3-4 minutes, please call me as the system might have crashed and I need to restart everything for you. And second, as I said you are encouraged to take breaks however you can not take them during the conversation as the system expects you to talk back. But during the essay part you are more than welcomed.”

Posttest instructions

“You have one last step in this experiment, the second test. I know you might by tired, it’s been several hours since you started. This is also a very important test so try to do as good as you can. Please take your time and take breaks if needed.”

Discussion at the end

In the open-answer interview of the end of the experiment, we first asked users if they have any comments, critiques or suggestions. After that, a number of questions were asked depending on the version of the system:

- `R` version: We asked about user's usage of the dialogue history: if it was useful in general, usage during conversation, usage during essay update, if they needed to scroll up and down to access information.
- `PI` version: We asked the questions for `R` version and in addition we asked if they got any additional explanations (i.e. activations of the new PopUp–Incorrect strategy) and if yes if it was useful or not. We also asked which parts of the additional explanations webpage were useful.
- `NM` version: We asked about user's usage of the NM: if it was useful overall, usage during conversation, usage during essay update, if they had problems concentrating on the speech and the NM in the same time, if there was enough information in the NM.

A.4.2 PRE score distribution per condition

Figure 34 shows the distribution of PRE scores for each condition in the Main Experiment. The majority of users have a PRE score between 8 and 14 with a relatively long distribution tail for higher PRE values.

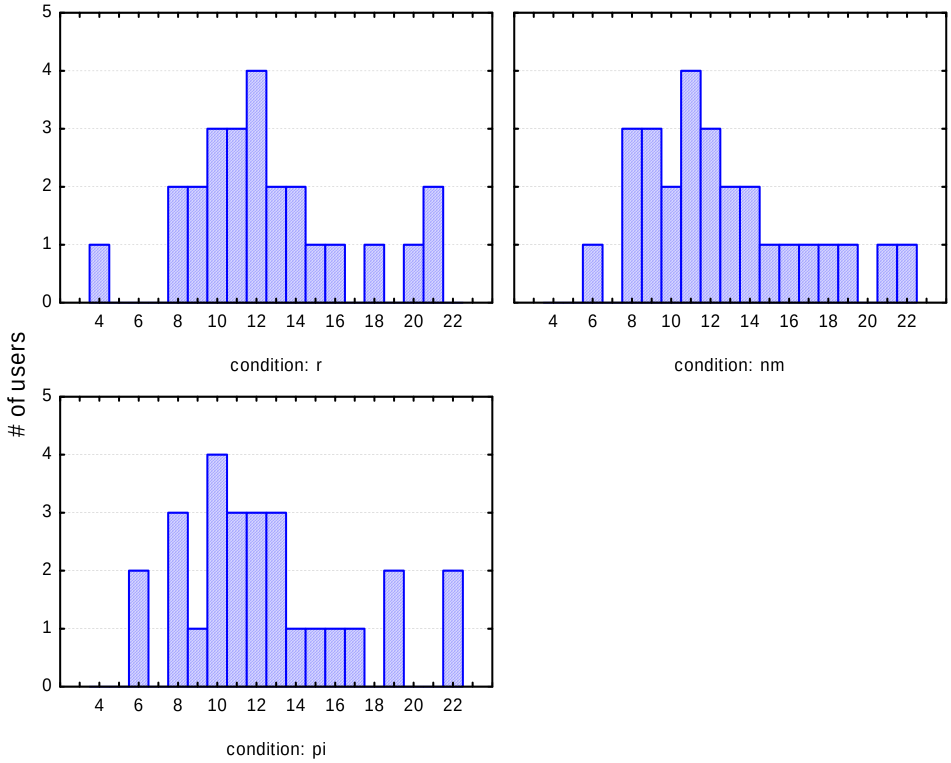


Figure 34. Histograms for PRE scores

A.4.3 Memory test

One of the exploratory hypotheses we wanted to investigate in the Main experiment was if there are any connections between the effectiveness of the NM and user's working memory span. We hypothesized the NM will help users with a shorter working memory span because of the information displayed in the NM and because the NM might act as a external repository of information that users can use while listening to the instruction.

To measure user's working memory span we used the reading memory span test and scoring procedure used in (Daneman and Carpenter, 1980). The test requires users to simultaneously read aloud sets of sentences and attempt to remember the last word of each sentence in the set. After reading a set, users were asked to recall the sentence final-words. The test starts with 5 sets of 2 sentences. A set is marked as completed if the user recalled correctly the final word for each sentence. If at least 3 sets were completed, the number of sentences in the set was increased by 1. For example, after successfully completing at least 3 out of the 5 sets of 2 sentences, the user moved to 5 sets of 3 sentences. The maximum set size was 6. The sets had no common or similar sentences.

The working memory span score (**MT score**) was defined as the largest set size for which the user managed to complete at least 3 out of the 5 sets. In addition, if the user completed exactly 2 out of the 5 sets, a credit of 0.5 was added to the MT score. For example, if a user got to sets of 4 sentences but was able to complete only 2 out of the 5 sets of 4 sentences, his/her MT score will be 3.5 (3 plus the 0.5 credit).

The test was administered on the computer. Depending on user performance, the memory test took between 10 and 35 minutes.

Effect of the assignment procedure on MT score distributions

Although we did not aim to balance for the memory test score, we also wanted to see if the pseudo-random assignment procedure (2.2.4.2) affected the balancing on this metric.

Table 37 shows the MT score average and standard deviation for all three conditions. We find that the averages and standard deviations are very similar. Indeed, a one-way ANOVA with MT score as dependent variable and Condition as a categorical factor (*R* vs. *NM* vs. *PI*) finds no significant differences between the three conditions ($F(2,76)=0.29$, $p<0.75$).

Table 37. Average and standard deviation for MT scores

Condition	MT Score
R	3.44 (0.82)
NM	3.46 (0.81)
PI	3.31 (0.67)

A distribution of the MT scores per condition is available in Figure 35. We observe that the majority of users scored 3 or 3.5 on the MT score. There are somewhat more *R* users with a score of 3 than *NM* and *PI* users and fewer with a score of 3.5. The distribution has a relatively long tail for higher values especially for *NM* and *PI*.

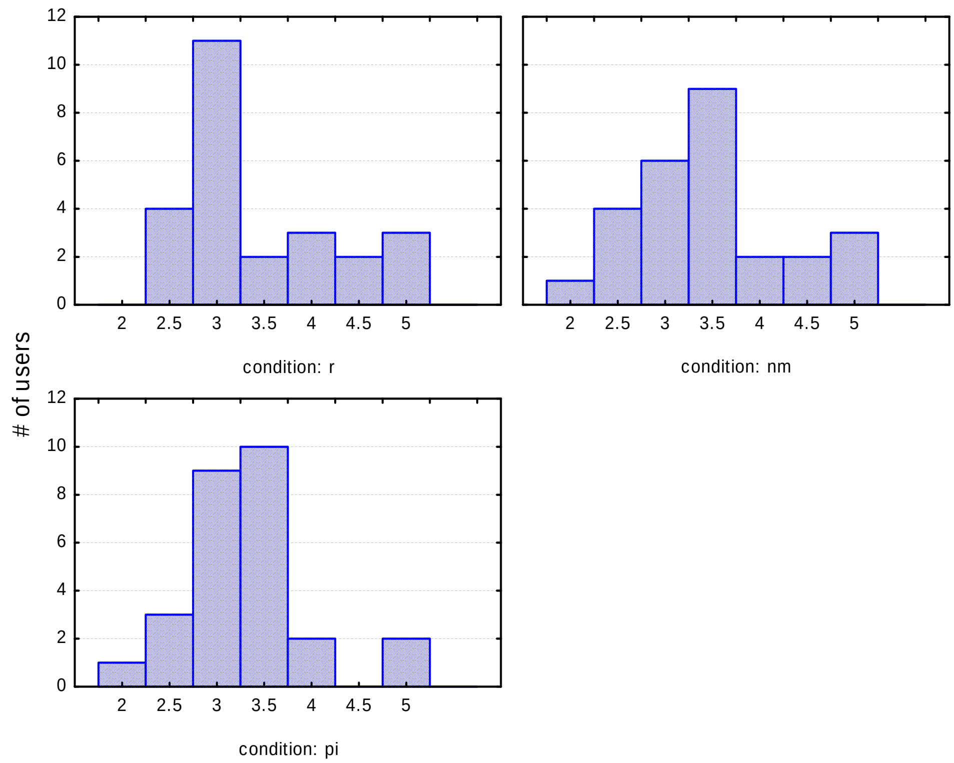


Figure 35. Histograms for MT scores

High/low split based on MT score

Since the MT score is another user aptitude metric, we wanted to investigate if this aptitude interacts with treatment when it comes to learning. Similarly to PRE score, we split the population using the average MT score (3.4) in two subsets: low MT and high MT (**MT Split**). There were 15, 13 and 11 users in the low subset and 10, 14 and 16 users in the high subset for *R*, *PI* and *NM* respectively.

The effect of MT Split on learning will be discussed in Sections B.1.3 and B.2.2.

A.4.4 A complete ITSPOKE-user dialogue

In Table 38 below, we present a transcript of the dialogue between a user from the Main experiment (2.2.4) and the Navigation Map version of ITSPOKE (3.4) on the second problem (A.2). We show the initial user essay at the beginning and the updated user essay at the end (note that only one essay revision was done in this experiment). For each system turn (TUTOR) we also show the depth in the discourse structure hierarchy, the transition (3.3) and the Why2-Atlas goals that form that turn. For each student turn we show the human transcript (TRANS), the system transcript (ASR) and the system and transcript correctness (2.3.1.4). The timing of the user turn is available in the user turn identifier (format:

minutes_seconds at the end of user turn identifier). User turn turn-m08-55-min-07_44 is a good example of a PopUp–Incorrect event (4.5.1).

Table 38. Complete transcript of a user-system dialogue

INITIAL ESSAY	" No. The packet is affected by a horizontal velocity. Therefore, the package will continue to move in teh planes's direction of motion while also being affected by the gravitational force as it it pulled downward."	
TUTOR	" You need more details in your essay. I'll try to help you remember some details to include in the explanation. As usual with problems where an object is launched, dropped, thrown, etc., we need to analyze both its flight and what happened just before it was launched. So let's begin by describing the velocity of the packet before it is dropped. How does the horizontal component of the packet's velocity compare to that of the plane's before it is dropped?"	
	PROPERTIES:	Depth: **1** ; Transition : **NewTopLevel** Goals : FEEDBACK_VAGUE-ESSAY_2, TRANSITION_FROM-ESSAY-TO-ELICITATION_4, PLANE-DIALOGUE10, PLANE-DIALOGUE12
STUDENT	TRANS:	it is the same
	ASR:	it is the same
	PROPERTIES:	Correctness: **C**(asem), **C**(tsem)
TUTOR	" Good. So we have established that while the plane is carrying the packet its velocity is the same as the plane's. Now we will analyze the motion of the packet from the time it is dropped until it hits the ground. To do this we will first find the individual forces acting on the packet, then the net force, the packet's acceleration, and finally we will describe its velocity. Try to name the forces acting on the packet after it is released. Please, specify their directions (e.g., vertically up) and keep in mind that air resistance is negligible."	
	PROPERTIES:	Depth: **1** ; Transition : **Advance** Goals : FEEDBACK_POSITIVE_5, PLANE-DIALOGUE15, PLANE-DIALOGUE16, PLANE-DIALOGUE18
STUDENT	**turn-m08-55-min-03_38**	
	TRANS:	gravity down and horizontal velocity
	ASR:	gravity told they're in and the horizontal velocity
	PROPERTIES:	Correctness: **I**(asem), **C**(tsem)
TUTOR	" Yes. The direction of the force of gravity is always straight down. Because gravity is the only force, it is trivial to determine the NET force (i.e., the vector sum of all forces). So what direction does the NET force have?"	
	PROPERTIES:	Depth: **1** ; Transition : **Advance** Goals : FEEDBACK_POSITIVE_0, PLANE-REM-1811, PLANE-DIALOGUE20
STUDENT	**turn-m08-55-min-04_12**	
	TRANS:	down and over
	ASR:	down dunno for
	PROPERTIES:	Correctness: **C**(asem), **C**(tsem)

TUTOR	" Good. Does the packet have an acceleration? If yes, please specify its direction."	
	PROPERTIES:	Depth: **1** ; Transition : **Advance** Goals : FEEDBACK_POSITIVE_5, PLANE-DIALOGUE22
STUDENT	**turn-m08-55-min-04_26**	
	TRANS:	i don't know
	ASR:	it don't know
	PROPERTIES:	Correctness: **UA**(asem), **UA**(tsem)
TUTOR	" Well... To find the packet's acceleration we will use Newton's second law. What is the definition of this law?"	
	PROPERTIES:	Depth: **2** ; Transition : **Push** Goals : FEEDBACK_NEUTRAL_0, PLANE-REM-2211
STUDENT	**turn-m08-55-min-04_38**	
	TRANS:	f equals m a
	ASR:	f equals m a
	PROPERTIES:	Correctness: **I**(asem), **I**(tsem)
TUTOR	" Well... Newton's second law says that the net force acting on an object equals the mass times the acceleration of the object. We can define Newton's second law as F=m*a. Hence, what can we determine about the acceleration of the packet given that we know the net force acting on it is not zero and points straight down?"	
	PROPERTIES:	Depth: **2** ; Transition : **Advance** Goals : FEEDBACK_NEUTRAL_0, PLANE-REM-22111, PLANE-REM-2212
STUDENT	**turn-m08-55-min-05_16**	
	TRANS:	down
	ASR:	i don't know
	PROPERTIES:	Correctness: **UA**(asem), **PC**(tsem)
TUTOR	" Well... Since the net force 'F' and the acceleration 'a' are vectors, for the equation F=m*a to hold they both need to have the same direction. Since the net force points straight down, then the acceleration points straight down as well. From Newton's second law F=m*a we can deduce that if the magnitude of the net force is not zero then the magnitude of the acceleration is not zero. Otherwise the relationship would not hold. So, once again on the original question. Does the packet have an acceleration? If yes, please specify its direction."	
	PROPERTIES:	Depth: **1** ; Transition : **PopUp** Goals : FEEDBACK_NEUTRAL_0, PLANE-REM-22122, PLANE-REM-22121, TRANSITION_REASK_2, PLANE-DIALOGUE22
STUDENT	**turn-m08-55-min-06_02**	
	TRANS:	yes down
	ASR:	yes down
	PROPERTIES:	Correctness: **C**(asem), **C**(tsem)
TUTOR	" Fine. Let's look now at the velocity of the packet. In general, for two dimensional motion it is useful to analyze	

	the velocity of an object in terms of its horizontal component and its vertical component separately. So let's begin by analyzing the vertical component of the packet's velocity after the release. How would you describe the vertical component of the packet's velocity? (e.g., decreasing, zero, etc.)"	
	PROPERTIES:	Depth: **1** ; Transition : **Advance** Goals : FEEDBACK_POSITIVE_4, PLANE-DIALOGUE24, PLANE-DIALOGUE26
STUDENT	**turn-m08-55-min-06_35**	
	TRANS:	increasing
	ASR:	increasing
	PROPERTIES:	Correctness: **C**(asem), **C**(tsem)
TUTOR	" Good. How would you describe the horizontal component of the packet's velocity between the time it is released until it hits the ground?"	
	PROPERTIES:	Depth: **1** ; Transition : **Advance** Goals : FEEDBACK_POSITIVE_5, PLANE-DIALOGUE28
STUDENT	**turn-m08-55-min-06_48**	
	TRANS:	decreasing
	ASR:	decreasing
	PROPERTIES:	Correctness: **I**(asem), **I**(tsem)
TUTOR	" Well... We have just found that there is no force acting on the packet in the horizontal direction after it leaves the plane. So we will use Newton's first law to analyze the packet's horizontal velocity from the time it is released until it hits the ground. Do you recall what this law says?"	
	PROPERTIES:	Depth: **2** ; Transition : **Push** Goals : FEEDBACK_NEUTRAL_0, PLANE-REM-2811
STUDENT	**turn-m08-55-min-07_12**	
	TRANS:	Yes
	ASR:	Yes
	PROPERTIES:	Correctness: **C**(asem), **C**(tsem)
TUTOR	" Okay. Newton's first law says that an object that has no forces acting on it will move with constant velocity. Hence, we can deduce that the horizontal component of the packet's velocity will not vary at all. That is, it will be constant until it hits the ground. Once again on the original question. How would you describe the horizontal component of the packet's velocity between the time it is released until it hits the ground?"	
	PROPERTIES:	Depth: **1** ; Transition : **PopUp** Goals : FEEDBACK_ACK_0, PLANE-REM-2812, TRANSITION_REASK_3, PLANE-DIALOGUE28
STUDENT	**turn-m08-55-min-07_44**	
	TRANS:	No
	ASR:	No
	PROPERTIES:	Correctness: **I**(asem), **I**(tsem)

TUTOR	" Well... Constant. Will the packet have a horizontal displacement after it is released?"	
	PROPERTIES:	Depth: **1** ; Transition : **Advance** Goals : FEEDBACK_NEUTRAL_0, ANSWER_0061, PLANE-DIALOGUE30
STUDENT	**turn-m08-55-min-07_54**	
	TRANS:	Yes
	ASR:	Yes
	PROPERTIES:	Correctness: **C**(asem), **C**(tsem)
TUTOR	" Right. Will the packet hit the target?"	
	PROPERTIES:	Depth: **1** ; Transition : **Advance** Goals : FEEDBACK_POSITIVE-EXCLUDE-Y/N_2, PLANE-DIALOGUE32
STUDENT	**turn-m08-55-min-08_02**	
	TRANS:	no
	ASR:	no
	PROPERTIES:	Correctness: **C**(asem), **C**(tsem)
TUTOR	" Fine. You should notice that to analyze this problem we began by looking at the motion of the packet right before it was launched and then during its flight. In analyzing the flight of the packet we found the forces applied on it, the net force, its acceleration, velocity, and finally its displacement. This allowed us to conclude that it would not hit the target. What we've talked about should help you. Let's see how you do now with writing the essay. Please remember to press submit."	
	PROPERTIES:	Depth: **0** ; Transition : **PopUpAdv** Goals : FEEDBACK_POSITIVE-EXCLUDE-Y/N_0, PLANE-DIALOGUE34, TRANSITION_FROM-WALKTHRU-TO-ESSAY_1, ESSAY-DIRECTIVE_TRY-TO-WRITE-ESSAY_3, INTERFACE-DIRECTIVE_REMIND-ENTER-SUBMIT_5
UPDATED ESSAY	" No. The hoizontal component of velocity will maintain and the packet will not hit the spot."	

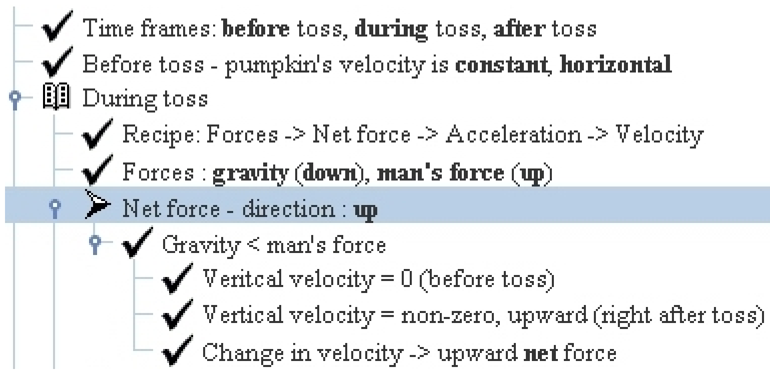
A.4.5 A sample PopUp–Incorrect webpage

We present below in Figure 36 the additional explanation webpage (4.5.2) that is activated for the *PI* version of the system in case of a PopUp–Incorrect event for the system question DIALOGUE20 (4th problem - A.2).

Tutor question: What is the direction of the NET force?

Correct answer: **Vertically up**

Dialogue summary:



What did we learn so far?

We learned that before the toss, the pumpkin's velocity is constant in the horizontal direction. We are now looking what happens while the man is tossing the pumpkin. Note that the man is still holding the pumpkin during the toss. There are two vertical forces acting on the packet: gravity (down) and man's force (up).

How did we try to find the correct answer to this question?

Recall the example with a hockey puck from your reading material:

"Suppose you attach a thread to a puck on smooth, nearly frictionless ice. If you pull on the thread, the puck accelerates. If your friend also attaches a thread to the puck and pulls in the same direction you are pulling, then the puck has greater acceleration. That is, acceleration of an object is proportional to the net force acting on it. In this case, the net force is the combination of the two forces exerted on the puck, one due to your thread and the other due to your friend's thread. Now suppose that your friend pulls away from you. In this case, the force your thread exerts is opposite the force that your friend's thread exerts. If the two forces are equally strong, then they cancel each other, so the net force is zero and the puck has zero acceleration. It remains stationary. Thus, acceleration is due to the net force on an object, which is the sum of all the individual forces acting on the object."

In our case, we know that the pumpkin is accelerating up. This is because before the toss it has a zero vertical velocity (remember the man is running in a straight line at constant speed, thus there is no movement in the vertical dimension). Right after the toss, the pumpkin will have a non-zero upward velocity that will allow it to fly up in the air.

In order for the pumpkin to accelerate up, the **net force needs to be upwards**. Since we have two opposite forces acting on the pumpkin, in order for the pumpkin to have a upwards net force, the force acting upwards needs to be bigger than the force acting downwards. In other words the man's force is bigger than that of gravity. Going back to the puck example, if you want the puck to move towards you, you will need to pull harder than your friend: the force you exert on the puck will be bigger than the force exerted by your friend.

Additional information

If you still feel unsure about the answer to this question, the material below provides additional information

Intuition

Whenever you are holding an object, you need to exert some effort to compensate for the objects weight: you are acting with a force on the object that is canceling the force of gravity. However, you will need to exert a larger effort to throw it in the air than to hold it. In other words, the force you need to exert on the object needs to be bigger than force required to hold it which was equal with the force of gravity.

Here is one other example why the force exerted by the man is bigger than the force of gravity. Imagine somebody is handing you a very heavy object. You attempt to hold it but it is too heavy for you. What this means is that the force you exert on the object to hold it is smaller than the object's weight (i.e. the gravity force). So in our case, the force exerted by the man can not be smaller than the force of gravity as the pumpkin will fall to the ground instead of being tossed.

What is the purpose of this question?

We already know how the pumpkin moves before being tossed. Knowing the forces that act on the pumpkin will allow us to describe its acceleration and velocity. Note that we do not need to know for this problem the magnitude of the net force but only its direction. If we wanted to know how high the pumpkin will fly then we would have to look at the magnitude and the problem text should contain more details on how much force the person puts into tossing the pumpkin.

How does this question relate to the other problems we discussed?

This is the first time we have discussed about a situation in which two forces are acting on the same object. In the previous problems, there was only one force acting on the objects, thus it was easy to find the net force.

This problem is similar to the plane-packet problem we discussed earlier but there is an important difference. In both problems, a carrier (the plane and the man) is moving at constant horizontal speed while holding another object (the plane holds the packet; the man holds the pumpkin). In the end, each object separates from its carrier but in a different way. In the previous problem, the packet was simply dropped: nobody was pushing it down or up. In contrast, in this problem the *pumpkin is being pushed by the man*. For this reason, we need to look at an extra time frame: while the entity is separating from the object (i.e. while the man is tossing the pumpkin). In this time frame, two forces are acting on the pumpkin with one being greater than the other.

Possible pitfalls

If you watch closely somebody running you will see that their body is moving up and down slightly with each step. Thus they also have a small movement on the vertical direction. However in this problem we are ignoring this. You can imagine a robot on wheels instead of the man.

Figure 36. A sample additional explanations webpage

APPENDIX B

MAIN EXPERIMENT: ADDITIONAL ANALYSES & RESULTS

Below we present a number of additional analyses and results. Part of them (B.1.3 and B.2.2) explore possible connections between the effectiveness of the NM and user's working memory span (recall our exploratory hypothesis from A.4.3). The rest offer a more complete picture on some of the results we presented earlier in the text.

B.1 THE POPUP–INCORRECT INVESTIGATION

B.1.1 Complete parameter-learning correlation tables

In our investigation of the predictiveness and informativeness of the discourse structure based parameters (Sections 4.4 and 4.5), whenever we looked at correlations between parameters and learning we only reported the best Pearson's Correlation Coefficient (R) associated with the parameters derived from specific unigrams/bigrams and the statistical significance of that coefficient R (p). Here we show the correlation information for each parameter. We use the naming convention from Section 4.3: for each bigram we have a *total* parameter (identified as bigram followed by "#"), a *percentage* parameter (identified as bigram followed by "%") and a *relative percentage* parameter (identified as bigram followed by "rel%"). For each unigram, we only have a total and a percentage parameter.

Table 39 shows the correlations with learning for all unigram and bigram parameters on the F03 corpus. It is an extension of Table 8, Table 9, Table 10, and Table 11 from Section 4.4.1.

Table 39. Parameters correlated with learning in the F03 corpus

Bigram	Avg (StdDev)	R	p
Neutral %	37% (8%)	**-0.47**	**0.04**
PopUp-Correct #	7 (3.3)	**0.45**	**0.05**
PopUp-Incorrect #	2 (1.8)	-0.42	*0.07*
PopUp-Incorrect %	1.6% (1.2%)	**-0.46**	**0.05**
PopUp-Incorrect %rel	17% (13%)	*-0.39*	*0.10*
PopUpAdv-Correct #	2.5 (2)	*0.43*	*0.06*
PopUpAdv-Correct %	2% (1.3%)	**0.52**	**0.02**
NewTopLevel-Incorrect #	2.3 (1.8)	**0.56**	**0.01**
NewTopLevel-Incorrect %	1.9% (1.4%)	**0.49**	**0.03**
NewTopLevel-Incorrect %rel	15% (12%)	**0.51**	**0.02**
Advance-Correct #	40.5 (9.8)	**0.45**	**0.05**
Advance-Neutral #	27 (8.3)	*-0.40*	*0.08*
Advance-Neutral %	21% (6%)	**-0.62**	**0.00**
Advance-Neutral %rel	38% (10%)	**-0.73**	**0.00**
SameGoal-Neutral %rel	35% (31%)	**0.46**	**0.05**
Advance-Advance #	35 (9.1)	**0.47**	**0.04**
Push-Push #	2.2 (1.7)	**0.50**	**0.03**
Push-Push %	1.8% (1.3%)	**0.52**	**0.02**
Push-Push %rel	11% (7%)	**0.52**	**0.02**
SameGoal-Push %rel	18% (23%)	**0.49**	**0.03**

Table 40 shows the correlation with learning for all PopUp–Incorrect and NewTopLevel–Incorrect parameters on the F03, S05PR and S05SYN corpora. It is an extension of Table 12 from Section 4.4.2.

Table 40. Predictiveness generalization to other corpora

Corpus	Param	PopUp-Incorrect Avg (StdDev)	R	p	NewTopLevel-Incorrect Avg (StdDev)	R	p
F03	#	3.3 (2.0)	-0.27	0.27	2.5 (1.9)	**0.59**	**0.01**
	%	2.8% (1.4%)	*-0.43*	*0.07*	2.1% (1.5%)	**0.46**	**0.05**
	rel %	30% (15%)	-0.31	0.20	16% (12%)	**0.52**	**0.02**
S05PR	#	3.6 (2.5)	*-0.32*	*0.09*	3.8 (1.9)	-0.20	0.30
	%	2.4% (1.3%)	*-0.32*	*0.08*	2.6% (1.1%)	-0.19	0.32
	rel %	24% (13%)	-0.14	0.45	18% (9%)	-0.20	0.30
S05SYN	#	4.5 (2.5)	**-0.50**	**0.00**	4.0 (1.9)	-0.11	0.56
	%	3.1% (1.4%)	**-0.44**	**0.01**	2.8% (1.4%)	-0.13	0.49
	rel %	34% (14%)	**-0.36**	**0.04**	18% (9%)	-0.10	0.58

Table 41 shows the correlations with learning for all PopUp–Correct and PopUp–Incorrect parameters[42] in the `R` and `PI` conditions. It is an extension of Table 15 from Section 4.5.4.4.

Table 41. Partial correlations for PopUp–Correctness bigram parameters

	Param	PopUp-Correct Avg (StdDev)	R	p	PopUp-Incorrect Avg (StdDev)	R	p
R	#	6.7 (2.6)	0.25	0.23	2.6 (1.1)	-0.33	0.12
	%	8.5% (2.5%)	0.28	0.19	3.4% (1.4%)	*-0.34*	*0.10*
	rel %	65% (12%)	0.30	0.15	27% (12%)	-0.33	0.11
PI	#	5.1 (2.3)	0.11	0.60	2.5 (1.6)	-0.11	0.58
	%	7.0% (2.8%)	0.15	0.45	3.5% (2.0%)	-0.15	0.46
	rel %	59% (18%)	0.24	0.23	31% (18%)	-0.23	0.27

Table 42 shows the correlations with learning for all PopUp–Correct and PopUp–Incorrect parameters in the `R` and `PI` conditions *when using last 2 problems only*. It is an extension of Table 16 from Section 4.5.4.4.

[42] The average for the PopUp–Correct rel% parameter and the PopUp–Incorrect rel% parameter do not add up to 100% because after a PopUp we can also have a "Partially Correct" or "Unable to Answer" turn (recall that correctness annotation has 4 labels - 2.3.1.4)

Table 42. Partial correlations for PopUp–Correctness bigram parameters (last 2 problems only)

	Param	PopUp-Correct Avg (StdDev)	R	p	PopUp-Incorrect Avg (StdDev)	R	p
R	#	2.9 (1.1)	**0.44**	**0.03**	1.5 (0.8)	**-0.49**	**0.02**
	%	8.3% (2.6%)	**0.52**	**0.01**	4.4% (2.2%)	**-0.49**	**0.01**
	rel %	55% (15%)	**0.52**	**0.01**	30% (15%)	**-0.52**	**0.01**
PI	#	2.4 (1.3)	-0.05	0.80	1.5 (0.8)	-0.19	0.35
	%	7.2% (3.4%)	0.03	0.89	4.5% (2.4%)	-0.15	0.46
	rel %	51% (21%)	0.10	0.63	34% (20%)	-0.12	0.55

B.1.2 Condition effect on PRE-POST covariance

Figure 37 shows graphically the results of the ANCOVA test that looks at the effect of the Condition (*PI* vs. *R*) on PRE-POST covariance. Each small point in the graph represent a user with larger points indicating two users. The regression lines between PRE and POST is shown for each condition. Since Condition has no effect on the PRE-POST covariance according to the ANCOVA test, this means that there no significant differences between the two regression lines.

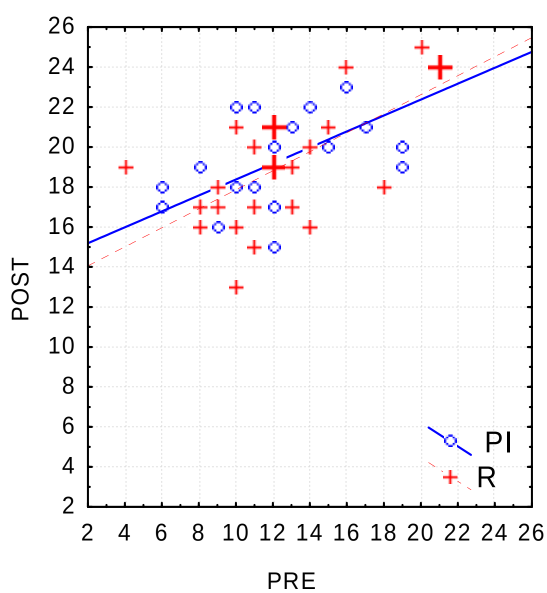


Figure 37. Condition effect on PRE-POST covariance

B.1.3 Memory test – Treatment interaction on Performance

Since the MT score is another user aptitude metric, we wanted to investigate if this aptitude interacts with treatment when it comes to learning. Similarly to PRE Split (4.5.4.3), we run a factorial ANOVA with the learning metric as dependent variable (POST or NLG) and the Condition (*PI* vs. *R*) and the MT Split (A.4.3) as categorical factors. We find that both MT Split and the MT Split × Condition combination have no significant/trend effect on our learning metrics. Figure 38 shows the average and confidence intervals for each group of users (low MT vs. high MT) on each condition for each learning metric.

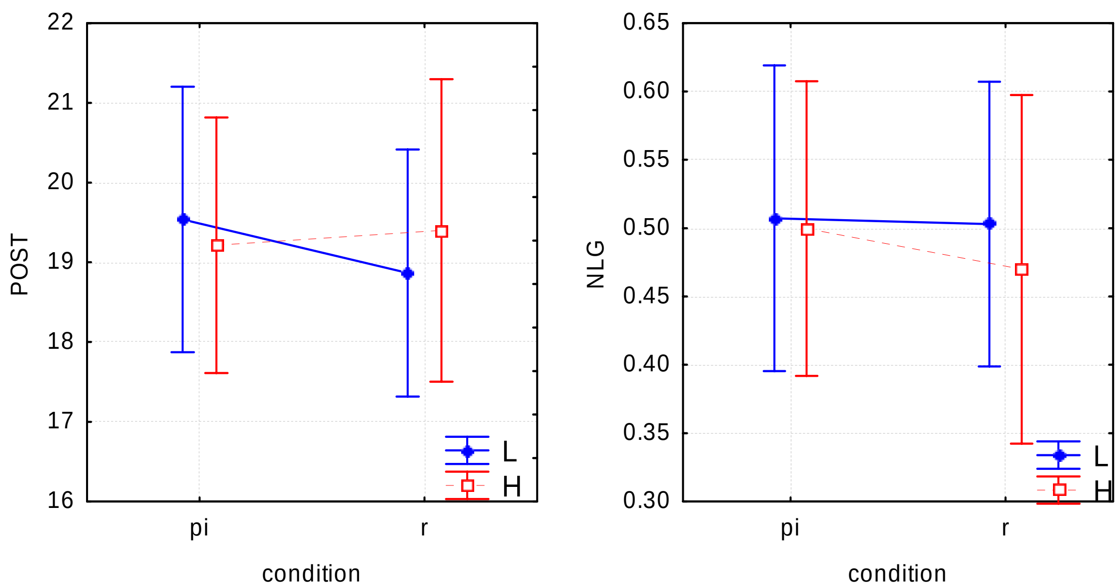


Figure 38. MT Split × Condition effect on learning
(Average and 95% confidence intervals are shown)

B.2 THE NAVIGATION MAP INVESTIGATION

B.2.1 Condition effect on PRE-POST covariance

Figure 39 shows graphically the results of the ANCOVA test that looks at the effect of the Condition (*NM* vs. *R*) on PRE-POST covariance. Each small point in the graph represent a user with larger points indicating two users. The regression lines between PRE and POST is shown for each condition. Since

Condition has no effect on the PRE-POST covariance according to the ANCOVA test, this means that there no significant differences between the two regression lines.

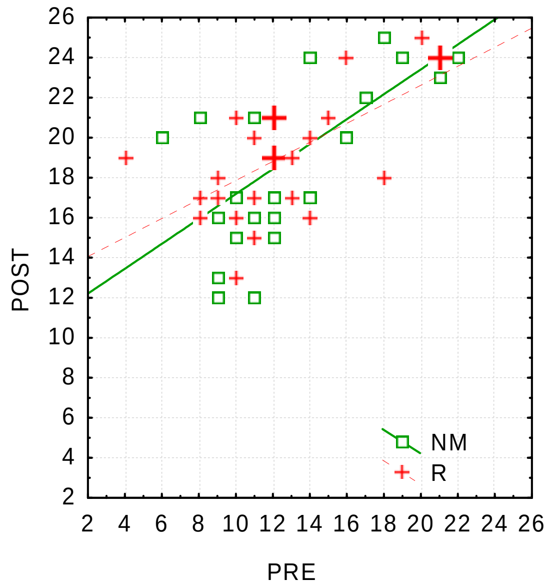


Figure 39. Condition effect on PRE-POST covariance

B.2.2 Memory test – Treatment interaction on Performance

Since the MT score is another user aptitude metric, we wanted to investigate if this aptitude interacts with treatment when it comes to learning. Similarly to PRE Split (4.5.4.36.4.2.2), we run a factorial ANOVA with the learning metric as dependent variable (POST or NLG) and the Condition (*NM* vs. *R*) and the MT Split (A.4.3) as categorical factors. Unlike PRE Split, we find that both MT Split and the MT Split × Condition combination have no significant/trend effect on our learning metrics. Figure 40 shows the average and confidence intervals for each group of users (low MT vs. high MT) on each condition for each learning metric.

This result invalidates our exploratory hypothesis that the NM effect depends on user's working memory span (A.4.3).

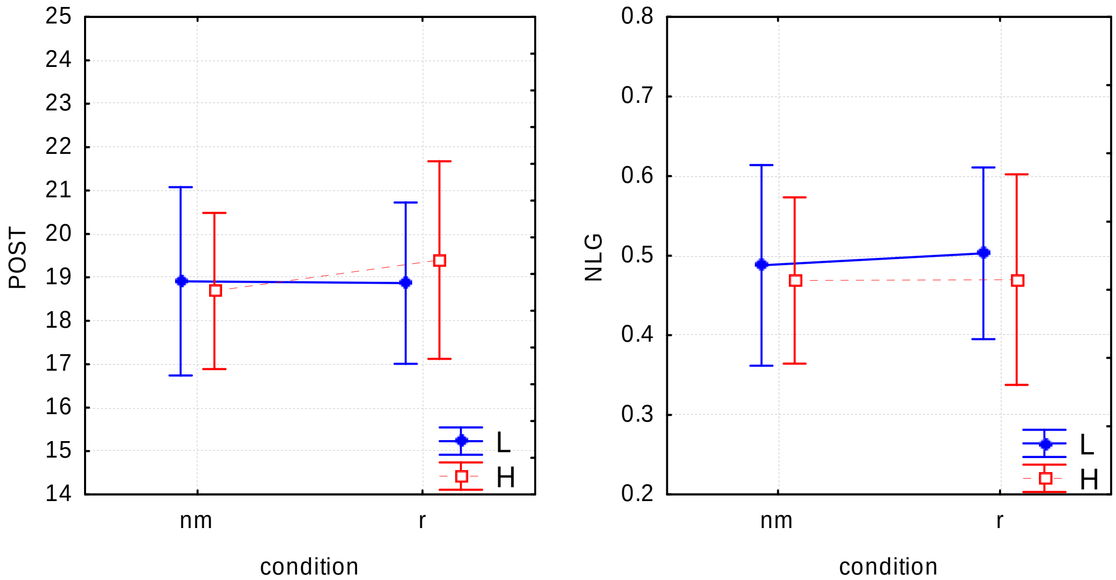


Figure 40. MT split × Condition effect on learning
(Average and 95% confidence intervals are shown)

B.2.3 Subjective metric – performance relationship

Figure 41 shows the ALL rating and the corresponding NLG for all *R* users. The regression line is also shown. There is no significant correlations for *R* users between their ALL rating and NLG (R=-0.04, p<0.86).

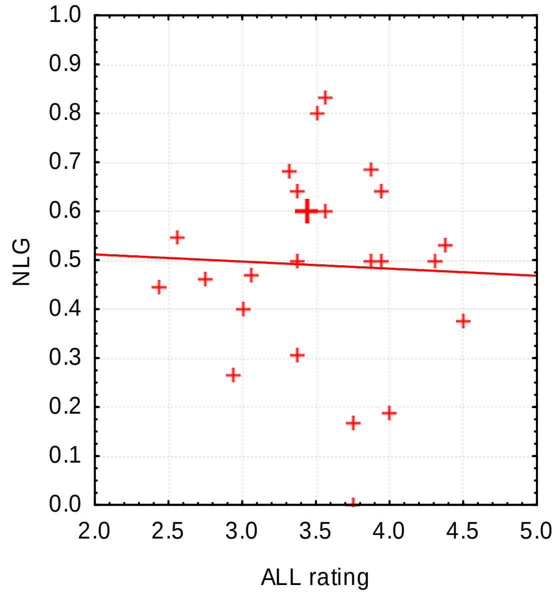


Figure 41. Correlation between the ALL rating and NLG for *R* users

APPENDIX C

THE NAVIGATION MAP ANNOTATION MANUAL

This document provides instructions for annotating the Navigation Map (**NM**) information as well as procedures to follow during annotation. The annotation manual is instantiated to the Why2-Atlas encoding of the tutoring information. Nonetheless, it can be easily extended to similar systems.

The manual was developed by the author and was used by only one annotator so far (the author). Since we are the first to explore the utility of the Navigation Map, the reliability of the annotation was of secondary importance. Nonetheless, this manual can be used a starting point to investigate the reliability of the discourse structure annotation, the right choice of granularity and its impact on the effectiveness of the Navigation Map.

C.1 THE NAVIGATION MAP INFORMATION

The NM requires annotation of the following four elements:

1. *Discourse segmentation* – the breakdown of the dialogue in discourse segments.
2. *Discourse segment intention* – for each discourse segment an annotation of the intention behind that segment on 4 dimensions which will be described later.
3. *Discourse segment hierarchy* – a hierarchical structure that implicitly encodes how discourse segments are related to each other.
4. *Information highlight* – identifies which words in the intention should be highlighted using bold or italics.

Information highlight is the only optional element. The other three elements are mandatory for the NM.

C.2 BRIEF INTRODUCTION TO THE WHY2-ATLAS REPRESENTATION

Before you can annotate the NM elements you need to be familiar with the way the tutoring information is encoded in the Why2-Atlas system. Why2-Atlas uses an system initiative strategy that follows a carefully designed *tutoring plan*. The student is asked each question in the tutoring plan and, depending on the answer, Why2-Atlas advances to the next question if the answer is correct or tries to remediate the incorrect answer. Remediation of incorrect answers is achieved either by a simple statement or by launching into a remediation subdialogue specifically designed for this situation.

Figure 42 shows part of a Why2-Altas tutoring plan. The complete tutoring plan including all goals referred in this manual is available online[43]. Why2-Atlas tutoring plans are organized by problem. For each problem, there is an HTML file with a graphical representation of the plans and an equivalent XML file which can be used for annotation.

The tutoring plan is organized around the notion of **goal**. Each goal has an identifier (e.g. Figure 42, ELEVATOR-KEYS-PROBLEM-WALK-THRU1). There are three types of goals:

1. *Statement* goals – for these goals the system will simply utter the statement(s) for the student and move on to the next goal. These goals have various applications. They can be used to convey strategies for approaching the problem (e.g. Figure 42, ELEV-DIALOGUE10 will specify that there will be two time frames in the analysis: before keys' release and after keys' release), remediate an incorrect answer to a question (e.g. Figure 42, ELEV-REM-121 explains why the velocities are the same) or state conclusions and facts (e.g. ELEV-DIALOGUE40 states the correct answer to the problem). Statement goals are marked with a green vertical bar in the HTML representation.
2. *Question goals* – these are goals that contain a question and for which the system expects an answer from the student. They might also include statements before or after the actual question (e.g. Figure 42, ELEV-DIALOGUE16 asks students to include the direction of the force in their answer to the question about forces on the man). If the student answer is correct, the goal is satisfied and the system moves on. If the student answer is incorrect, the system will try to remediate the incorrect answer with a *remediation goal*. Remediation of incorrect answers is achieved either by a statement goal or by launching into a remediation subdialogue specifically designed for this situation. Question goals are marked with a yellow vertical bar in the HTML representation.

---

[43] http://www.cs.pitt.edu/~mrotaru/thesis

3. *Group goals* – these are goals that are made of an ordered list of other goals. The system will go through each sub-goal in the list in their list order. Optionally, they can have a statement similar to a statement goal which will be uttered before going through the list of goals. They are primarily used to refer to a group of goals that talk about the same issues (e.g. Figure 42 ELEVATOR-KEYS-PROBLEM-WALK-THRU1 contains all the goals that discuss the solution to a problem) or to identify a remediation dialogue (e.g. ELEV-REM-201 will discuss about relationship about forces and acceleration in order to arrive to the correct answer for the question ELEV-DIALOGUE20). Group goals are marked with a green vertical bar in the HTML representation.

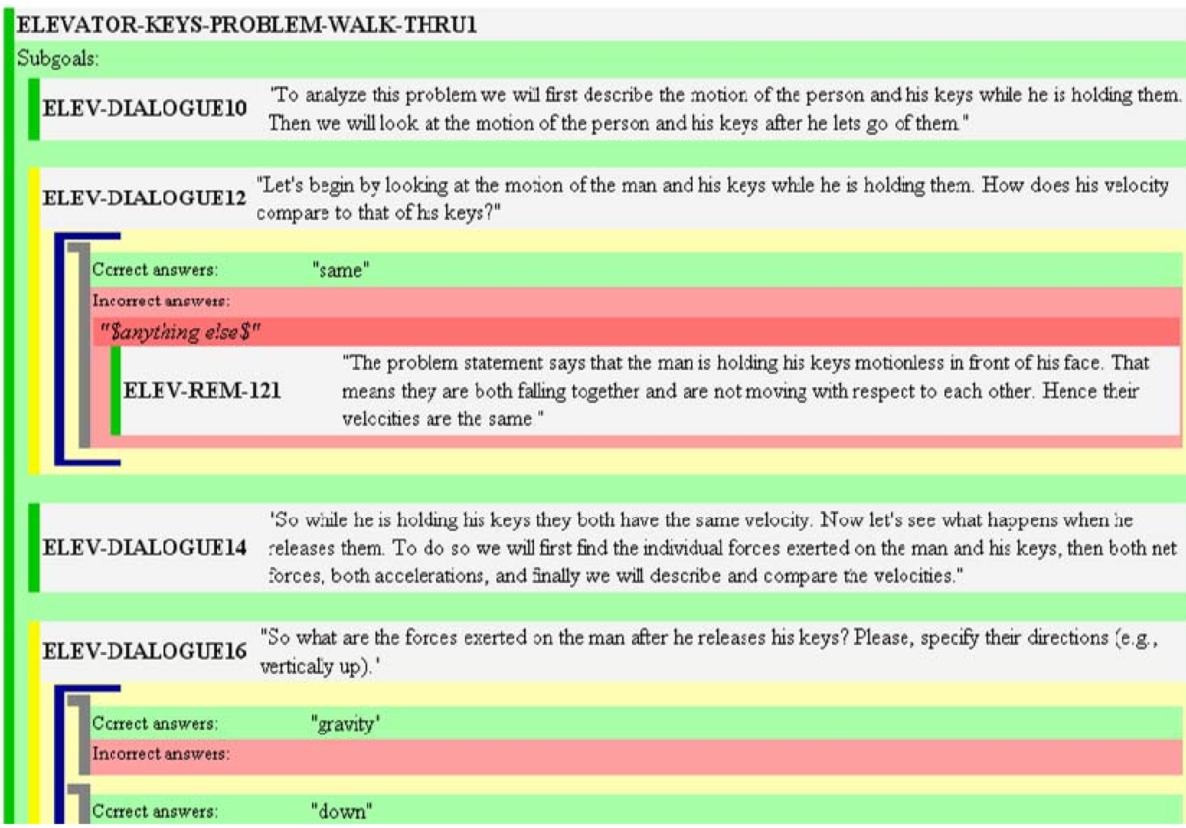


Figure 42. Part of a graphical representation (HTML) of a tutoring plan

When a student interacts with Why2-Atlas, the system first presents a problem and asks the student to type an essay describing the solution to the problem. The essay is analyzed and if it is incomplete or incorrect, the system selects a predefined goal that will address the issues. In most cases, the goal is a group goal and will be called *dialogue-level goal*. Next, the system engages in a dialogue by going through each of the subgoals of the group goal. After the dialogue, the system asks the student to update the essay and the essay revision/dialogue cycle is repeated as needed.

C.3 THE NM ANNOTATION

Your task is to annotate the NM elements for all possible dialogues between the student and the system. As mentioned in the previous section, each dialogue follows a tutoring plan specified in a goal (e.g. Figure 42 ELEVATOR-KEYS-PROBLEM-WALK-THRU1 encodes the solution walkthrough dialogue). You will have to identify the tutor intentions in this tutoring plan, segment it in discourse segments based on these intentions and organize these discourse segments in a hierarchical structure.

We recommend the following procedure for the NM annotation:

1. *Familiarize yourself with the problem* – Take some time to familiarize yourself with the problem. Identify in the problem statement the important objects, concepts, events and context. Also, you should have a clear understanding of what the user is being asked in the problem. For example, in the problem from Figure 43 there are two objects (the man and the keys), one event (the man let's go of the keys), the concept of position. Everything happens in the context of a free-falling elevator. The problem question is to identify the relative position of the keys to the man's face as time passes by.
2. *Annotate all dialogue-level goals* – Annotate all NM elements for each dialogue-level goal (i.e. the goals activated after an essay analysis). You should start with the walkthrough goal. This is the only goal per problem which contains the complete solution to the problem. Annotating the walkthrough goal first will allow you to familiarize yourself with the solution to the problem and help in your annotation of other dialogue-level goals. In addition, some dialogue-level goals might share subgoals with the walkthrough goal. To annotate a dialogue-level goal follow these steps
    2.1. *Familiarize yourself with the goal* – Take some time to go through all the subgoals of this goal. Read carefully the tutor statements, the tutor questions and identify the correct answer to the questions. To reduce this initial effort, follow the remediation goal for questions only if the correct answer does not satisfy your understanding of the discussion. Try to identify and write down the high-level discussion steps and the recipe behind the tutoring plan. Figure 43 shows the result of this step.

2.2. *Annotate intentions* – For each subgoal, read the tutor text and identify the intentions behind the tutor text. More details on this process can be found in Appendix C.4. Discourse segmentation is tightly connected to intentions. If the subgoal has only one relevant intention, the discourse segment is the subgoal itself. If the subgoal has more than one relevant intention, these intentions will naturally segregate the subgoal in its discourse segments.

2.3. *Annotate the hierarchy* – At this step you will need to arrange the intentions in a hierarchical structure. More details are available in Appendix C.5. Note that sometimes it might be easier to do step 2.2 and this step in a single pass.

2.4. *Annotate remediation goals* – Recursively go through each goal that remediates one of the subgoals already annotated and repeat steps 2.1-2.3 for each remediation subgoal. Note that it might be easier to annotate remediation statement goals while annotating the goal being remediated.

We recommend multiple passes through the NM annotation as you will become more familiar with the tutoring plan during the annotation. You can get a different perspective on some of the earlier subgoals after you have finished annotating all the subgoals. Doing multiple passes will also increase the conciseness of the intention annotation, the uniformity of your lexical choices and can help you smooth out transition between certain subgoals.

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.*

Familiarizing with the walkthrough goal:

- Discussion split over two events: before man releases key and after man releases keys
- Short discussion before release, pretty big after release (described below)
- Recipe: force, acceleration, velocity, displacement
- Looks first at force/acceleration for the man
- Repeats for the keys
- Compares their velocities based on accelerations being the same
- Compares their displacement based on the velocities being the same
- Conclude the keys stay in front of man’s face during the fall

Figure 43. A sample Why2-Atlas problem and a sample outcome of familiarizing with the walkthrough goal (problem 58, ELEVATOR-KEYS-PROBLEM-WALK-THRU1 goal)

C.4 INTENTION ANNOTATION

Your task is to identify and label the tutor intentions behind a goal. You should use the intuitive notion of intention and ask yourself the following questions:

- What is the tutor trying to communicate here?
- What is the tutor main point?
- Are there any other secondary points? Which of these points are worth mentioning in the NM and which can be ignored?
- How does the goal relate to what has been discussed so far and what will be discussed in the future?
- Are there any intentions/statements that are important for the future goals?

You will need to annotate the intention on 4 dimensions:

1. FOCUS – This the intention of the tutor when uttering the text of a goal and is the first intention you will annotate.
2. EXIT – This the intention of the tutor after finishing with a goal. In most cases, for statement and group goals, this intention will be the same as the FOCUS intention. However, for question goals, the tutor's intention when finishing with a goal is for the student to know the answer to that question. That is why the EXIT intention should include the correct answer. For example, if the tutor asks what is the value of the acceleration, then the FOCUS intention will be "Acceleration = ?" while the EXIT intention will be "Acceleration = 9.8 m/s^2".
3. LOOKAHEAD – One of the NM features is the limited horizon. This feature enables student to see the intention of the next goal. In a small number of cases, the FOCUS intention of the next goal can give away the correct answer for the current goal. For this reason, every goal requires a LOOKAHEAD intention annotation to be used while the goal is part of the limited horizon.
4. ENTER – This intention is activated only while the tutor is remediating a question goal or while going through the subgoals of a group goal.

Note that a node without an intention annotation will not appear in the NM. All nodes under this node will be promoted at level corresponding to the node without an intention.

The following case studies will give you a good coverage of the situations you will encounter during annotation:

C.4.1 Case study: single intention question goal

`Goal name: ELEV-DIALOGUE20`

`Goal text: Considering the net force, does the person have an acceleration? If yes, please specify its direction.`

`FOCUS Intention: Acceleration`

`EXIT Intention: Acceleration is` **`non-zero`**`,` **`down`**

In many cases, it is straightforward to identify the tutor's intention. For this goal the intention is to learn from student about the acceleration. The direction was not included in the FOCUS intention as it would produce a confusing text for the intention (e.g. "Acceleration, direction?"). However, the answer for the direction is included in the EXIT intention.

C.4.2 Case study: single intention statement goal

`Goal name: ELEV-DIALOGUE10`

`Goal text: 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.`

`FOCUS, EXIT Intention: Time frames:` **`before`** `release,` **`after`** `release`

This goal has two sentences, however there is only one intention behind this goal: to establish the two time frames for analysis. Note that the two time frames will be recalled explicitly in the next goals (i.e. ELEV-DIALOGUE12 and ELEV-DIALOGUE14). Thus, it is a good practice to reuse the formulation in their intention annotation (i.e. "before release" and "after release").

C.4.3 Case study: double-intention question goal

`Goal name: ELEV-DIALOGUE12`

`Goal text: [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?]`

`FOCUS, EXIT Intention 1: Before release`

`FOCUS Intention 2: Man's velocity` **`?`** `key's velocity`

`EXIT Intention 2: Man's velocity` **`same`** `key's velocity`

There are two intentions behind this goal. First, the tutor wants to establish that it will analyze what happens before the release of the keys. Next, the tutor wants to compare the velocities. While the second intention is the most important (the actual question), the first intention is very important for highlighting the structure of the tutor plan (recall the notes from Figure 43). When we will talk about structure annotation, you will see that the first intention dominates the second one (i.e. it is the parent of the second intention in the hierarchy). In a different context, this first intention could have been ignored.

C.4.4 Case study: redundant correct answer

```
Goal name: ELEV-DIALOGUE14
Goal text: So while he is holding his keys they both have the same
      velocity. [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.]
FOCUS, EXIT Intention 1: After release
FOCUS, EXIT Intention 2: Recipe: Forces -> Net force -> Acceleration ->
      Velocity -> Displacement
```

This statement goal has two intentions. However, notice that at the beginning of the goal the tutor is restating the correct answer to the previous question goal (i.e. ELEV-DIALOGUE12). This technique was used in many cases by the tutoring plan designers to make sure the correct answer is clearly stated for the student (though this leads to redundant information when the previous question goal was remediated). Note that the intention of stating the correct answer is ignored in the annotation as the previous goal intention has already displayed this information in its EXIT intention. Also, note that displacement is not included in the tutor text, but a quick look at the solution shows that the tutor will also talk about displacement (in fact ELEV-DIALOGUE41 adds displacement to the recipe).

C.4.5 Case study: redundant remediation goal

```
Goal name: ELEV-REM-28121
Goal text: According to Newton's second law F=m*a. Hence, if the net force
      F is not zero, then the acceleration 'a' cannot be zero.
Intention: NULL
```

Some of statement remediation goals will only state the correct answer without further explanation or if they provide explanation, it can be deduced from the intention of the previous goals. In these cases, you can choose not to provide any intention for the goal to reduce the graphical overload on the student. For this goal, Figure 44 shows the state of the NM while the tutor is going through this goal in an attempt to remediate the goal ELEV-REM-2812 (NM intention "*F* non-zero -> *a zero* or *non-zero*"):

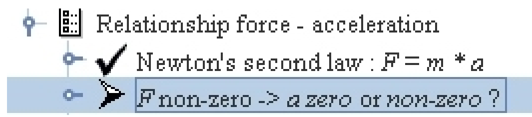


Figure 44. Navigation Map state for a sample redundant remediation goal

You can see that the entire explanation in the goal ELEV-REM-28121 is already present in the NM (i.e. the definition of the second law is already present above in the NM).

Recommendations:

- Try to be as concise as possible when designing the text of an intention. Do not include information that is not crucial as you have very limited space. Make use of mathematical symbols if needed. However, try not to be very cryptic.
- Try to use tutor terminology unless confusing.
- Try to use the same terminology for a concept, law, idea, object across a dialogue and between problems even if the tutor terminology varies slightly.
- Annotate first the FOCUS and EXIT intentions for all the goals in a dialogue-level goal. Then do a second pass to annotate the LOOKAHEAD intention.
- Remediation goals of a question can give you insight in terms of the best way to write the EXIT intention.

C.5 STRUCTURE ANNOTATION

Your task is to organize all the intentions in a hierarchical structure. This structure should explicitly show the dominance relationship between intentions. Figure 45 shows an example hierarchy for the problem from Figure 43.

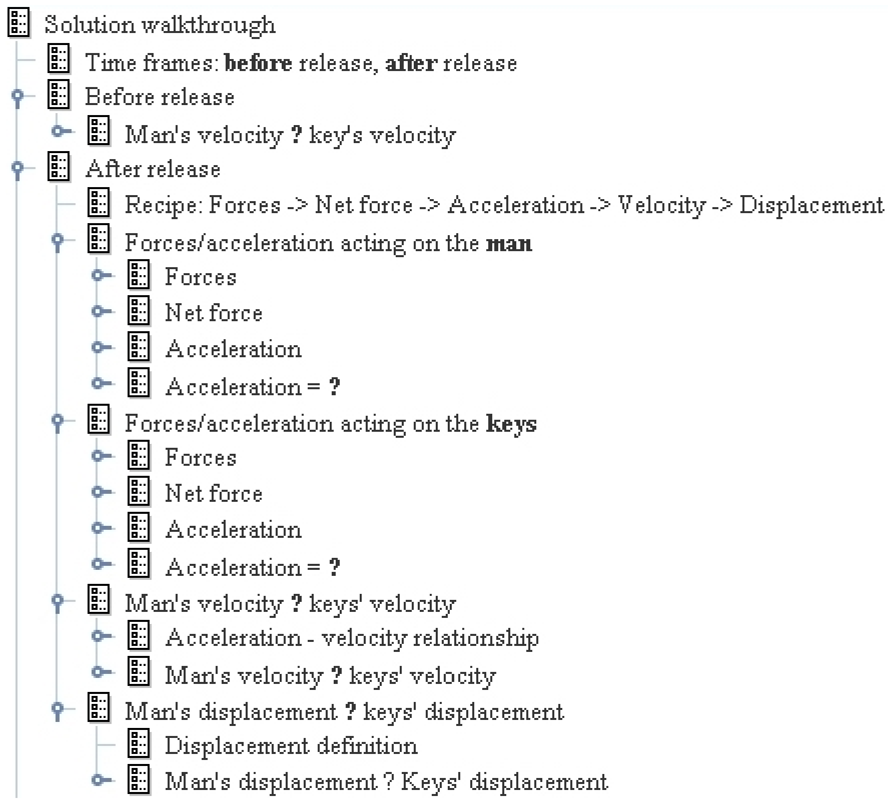


Figure 45. LOOKAHEAD intentions and hierarchy for a Why2-Atlas walkthrough goal (problem and goal from Figure 43)

You can use the structuring created by the group goals as your starting point. However in many cases this structuring is very coarse (e.g. all goals annotated in Figure 45 are part of a single group goal – i.e. they are at the same level in the authoring structure).

In many cases, the tutor text will include a description of the plan which can help you with arranging the intention in a hierarchy. For example, goal ELEV-DIALOGUE10 clearly identifies the two time frames ("before release" and "after release") and refers to them in the next intention. Thus, the "before release" intention will dominate (i.e. be the parent of) all intentions related to that time frame (only 1 in Figure 45); similarly the "after release" intention will dominate all intentions discussing that time frame (most of dialogue in Figure 45). As another example, goal ELEV-DIALOGUE14 identifies the recipe and the discussion will follow relatively closely the recipe. Thus, we choose to group intentions so that the recipe is highlighted (in Figure 45 you can see at the same level with the recipe the intentions that follow the recipe: "Forces/acceleration acting on the man", "Forces/acceleration acting on the keys", "Man's velocity ? keys' velocity", and "Man's displacement ? keys' displacement").

In other cases, the tutor text will not clearly identify the structure of the plan and you will have to manually identify the structure based on the intentions. For example, the tutor first talks about the forces and acceleration acting on the man (goals ELEV-DIALOGUE16 to ELEV-DIALOGUE22) and then engages in a similar discussion about the keys (goals ELEV-DIALOGUE24 to ELEV-DIALOGUE30). Given the recipe identified in ELEV-DIALOGUE14, it is natural to group the discussion about the man in one higher intention and to do the same for keys (see the intentions "Forces/acceleration acting on the man" and "Forces/acceleration acting on the keys" in Figure 45). Note that in most of these cases, you will need to create an extra goal in your annotation as no tutor goal explicitly identifies this intention. As another example, the intentions "Man's velocity ? keys' velocity", and "Man's displacement ? keys' displacement" in Figure 45 were also created to conform with the tutor recipe as the discussion of this points will require several goals.

In some cases, a tutor goal will have two intentions which can be on different levels (goal ELEV-DIALOGUE12 has two intentions in Figure 45 "Before release" and "Man's velocity ? key's velocity" but the first one dominates the second one).

C.6 ANNOTATION TOOLS

To annotate the intentions for each goal, we developed a tool called NMAnnotator. Figure 46 shows a screenshot of NMAnnotator. The tool can be used to preview how the NM will look at runtime and to synchronize the prerecorded system prompts with the discourse segments. Changing the structure has to be done manually by editing the appropriate files after which the information is recompiled and the NMAnnotator is updated.

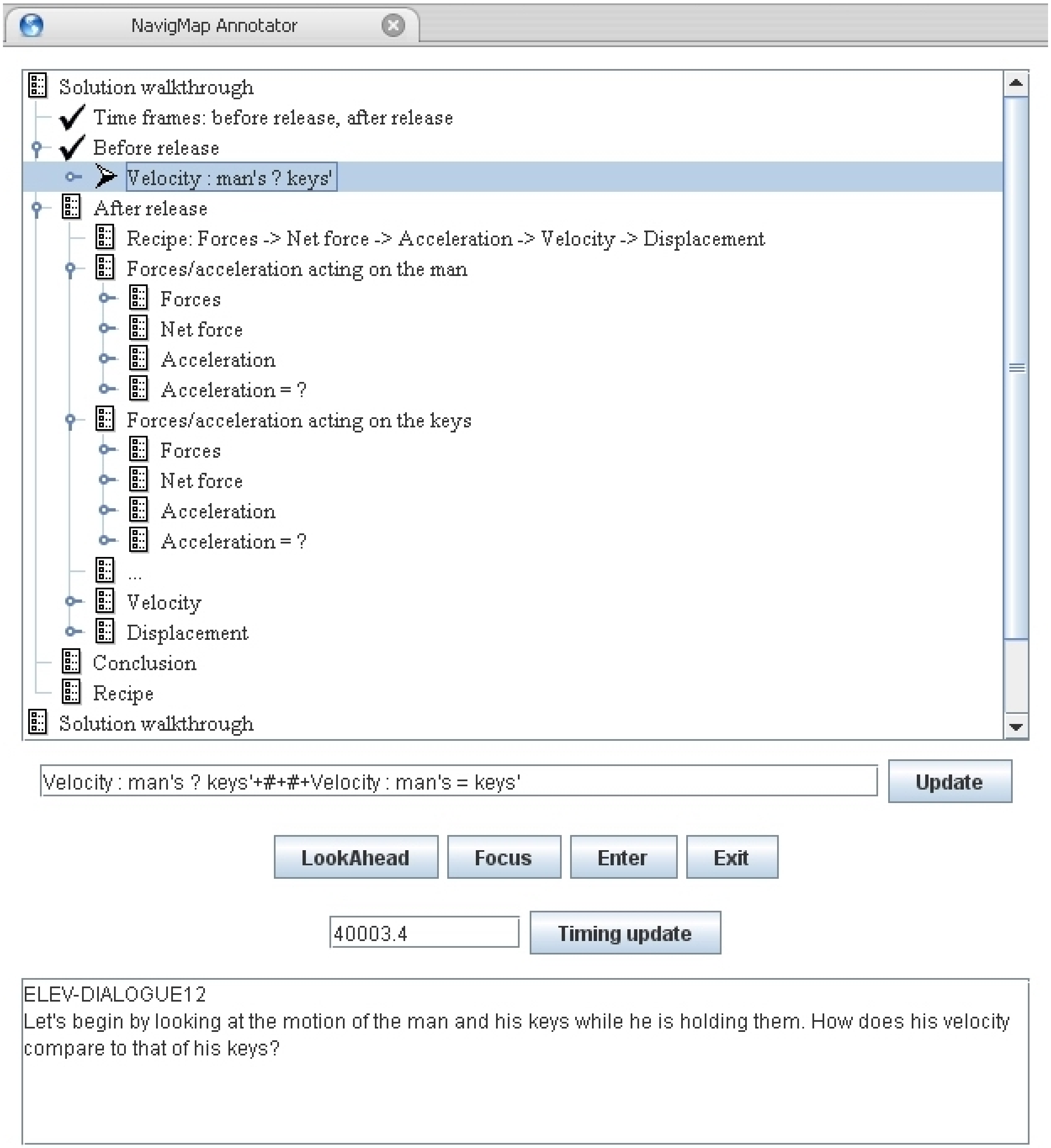


Figure 46. Screenshot of the NM annotation tool

BIBLIOGRAPHY

A. Abella, J. H. Wright and A. L. Gorin. 2004. *Dialog Trajectory Analysis.* In Proc. of ICASSP.

H. Ai and D. Litman. 2007. *Knowledge Consistent User Simulations for Dialog Systems.* In Proc. of Interspeech.

H. Ai, D. Litman, K. Forbes-Riley, M. Rotaru, J. Tetreault and A. Purandare. 2006. *Using System and User Performance Features to Improve Emotion Detection in Spoken Tutoring Dialogs.* In Proc. of Interspeech.

G. Aist, B. Kort, R. Reilly, J. Mostow and R. Picard. 2002. *Experimentally augmenting an intelligent tutoring system with human-supplied capabilities.* In Proc. of Intelligent Tutoring Systems.

P. Albacete. 1999. *An intelligent tutoring system for teaching fundamental physics concepts*. Ph.D. dissertation, University of Pittsburgh, Intelligent Systems Program

P. Albacete and K. VanLehn. 2000. *Evaluating the effectiveness of a cognitive tutor for fundamental physics concepts.* In Proc. of 22nd Annual Meeting of the Cognitive Science Society.

J. Allen, D. Byron, M. Dzikovska, G. Ferguson, L. Galescu and A. Stent. 2000. An Architecture for a Generic Dialogue Shell. *Natural Language Engineering, 6*(3-4).

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. Ang, R. Dhillon, A. Krupski, A. Shriberg and A. Stolcke. 2002. *Prosody-based automatic detection of annoyance and frustration in human-computer dialog.* In Proc. of ICSLP.

R. K. Atkinson, R. E. Mayerb and M. M. Merrill. 2005. Fostering social agency in multimedia learning: Examining the impact of an animated agent's voice. *Contemporary Educational Psychology, 30*(1).

R. Baker, J. Walonoski, N. Heffernan, I. Roll, A. Corbett and K. Koedinger. 2008. Why Students Engage in "Gaming the System" Behavior in Interactive Learning Environments. *Journal of Interactive Learning Research, 19*(2).

A. Batliner, K. Fischer, R. Huber, J. Spilker and E. Nöth. 2003. How to Find Trouble in Communication. *Speech Communication, 40 (1-2).*

K. Bhatt, M. Evens and S. Argamon. 2004. *Hedged Responses and Expressions of Affect in Human/Human and Human/Computer Tutorial Interactions.* In Proc. of Cognitive Science (CogSci).

N. Blaylock and J. Allen. 2006. *Fast hierarchical goal schema recognition.* In Proc. of AAAI.

B. S. Bloom. 1984. The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *13*(6).

D. Bohus. 2007. *Error Awareness and Recovery in Conversational Spoken Language Interfaces*. Ph.D. Dissertation, Carnegie Mellon University, School of Computer Science

D. Bohus and A. Rudnicky. 2003. *RavenClaw: Dialog Management Using Hierarchical Task Decomposition and an Expectation Agenda.* In Proc. of Eurospeech.

D. Bohus and A. Rudnicky. 2005. *Sorry, I Didn't Catch That! - An Investigation of Non-understanding Errors and Recovery Strategies.* In Proc. of Workshop on Discourse and Dialogue (SIGdial).

A. Boozer, S. Seneff and M. Spina. 2003. *Towards Recognition of Emotional Speech in Human-Computer Dialogues*. CSAIL Research Abstract.

J. Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics, 22*.

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.

M. Chi and K. VanLehn. 2008. *Eliminating the Gap between the High and Low Students through Meta-Cognitive Strategy Instruction.* In Proc. of Intelligent Tutoring Systems (ITS).

M. T. H. Chi, S. A. Siler, H. Jeong, T. Yamauchi and R. G. Hausmann. 2001. Learning from human tutoring. *Cognitive Science, 25*.

M. Core and J. Allen. 1997. *Coding Dialogs with the DAMSL Annotation Scheme.* In Proc. of AAAI Fall Symposium on Communicative Action in Humans and Machines.

R. Cowie and R. Cornelius. 2003. Describing the emotional states that are expressed in speech. *Speech Communication, 40*(1-2).

S. D. Craig, A. C. Graesser, J. Sullins and B. Gholson. 2004. Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media, 29*(3).

S. K. D'Mello, S. D. Craig, B. Gholson, S. Franklin, R. Picard and A. C. Graesser. 2005. *Integrating affect sensors in an intelligent tutoring system.* In Proc. of Affective Interactions: The Computer in the Affective Loop Workshop at 2005 International Conference on Intelligent User Interfaces (IUI).

M. Daneman and P. A. Carpenter. 1980. Individual Differences in Working Memory and Reading. *Journal of Verbal Learning and Verbal Behavior, 19*(4).

M. Danieli and E. Gerbino. 1995. *Metrics for evaluating dialogue strategies in a spoken language system.* In Proc. of AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation.

C. P. Doncaster and A. J. H. Davey. 2007. *Analysis of Variance and Covariance: How to Choose and Construct Models for the Life Sciences*. Cambridge University Press, Cambridge

K. Forbes-Riley. 2005. *Correlating Student Acoustic-Prosodic Profiles with Student Learning in Spoken Tutoring Dialogues.* In Proc. of 9th European Conference on Speech Communication and Technology (Interspeech-2005/Eurospeech).

K. Forbes-Riley and D. Litman. 2005. *Using Bigrams to Identify Relationships Between Student Certainness States and Tutor Responses in a Spoken Dialogue Corpus.* In Proc. of SIGdial.

K. Forbes-Riley and D. Litman. 2006. *Modelling User Satisfaction and Student Learning in a Spoken Dialogue Tutoring System with Generic, Tutoring, and User Affect Parameters.* In Proc. of HLT/NAACL.

K. Forbes-Riley, D. Litman, A. Purandare, M. Rotaru and J. Tetreault. 2007. *Comparing Linguistic Features for Modeling Learning in Computer Tutoring Dialogues.* In Proc. of International Conference on Artificial Intelligence in Education (AIED).

K. Forbes-Riley, D. Litman and M. Rotaru. 2008. *Responding to Student Uncertainty during Computer Tutoring: An Experimental Evaluation.* In Proc. of 9th International Conference on Intelligent Tutoring Systems (ITS).

K. Forbes-Riley, D. Litman, S. Silliman and J. Tetreault. 2006. *Comparing Synthesized versus Pre-recorded Tutor Speech in an Intelligent Tutoring Spoken Dialogue System.* In Proc. of Florida Artificial Intelligence Research Society (FLAIRS) Conference.

K. Forbes-Riley, M. Rotaru and D. Litman. 2008. The Relative Impact of Student Affect on Performance Models in Spoken Dialogue Tutoring Systems. *User Modeling and User-Adapted Interaction, 18*(1-2).

K. Forbes-Riley, M. Rotaru, D. Litman and J. Tetreault. 2007. *Exploring Affect-Context Dependencies for Adaptive System Development.* In Proc. of Human Language Technology / North American Chapter of the Association for Computational Linguistics Conference (HLT/NAACL).

M. Gabsdil and O. Lemon. 2004. *Combining Acoustic and Pragmatic Features to Predict Recognition Performance in Spoken Dialogue Systems.* 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).

A. Graesser, N. Person and D. Harter. 2001. Teaching tactics and dialog in AutoTutor. *International Journal of Artificial Intelligence in Education*.

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

B. Grosz, S. Weinstein and A. Joshi. 1995. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics, 21*(2).

A. Gruenstein, S. Seneff and C. Wang. 2006. *Scalable and Portable Web-Based Multimodal Dialogue Interaction with Geographical Databases.* In Proc. of Interspeech ICSLP.

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, D. Litman and M. Swerts. 2004. Prosodic and Other Cues to Speech Recognition Failures. *Speech Communication, 43*(1-2).

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

L. Hirschman, D. A. Dahl, D. P. McKay, L. M. Norton and M. C. Linebarger. 1990. *Beyond class A: a proposal for automatic evaluation of discourse.* In Proc. of Workshop on Speech and Natural Language.

K. S. Hone and R. Graham. 2000. Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering, 6*(3/4).

E. Horvitz and T. Paek. 1999. *A Computational Architecture for Conversation.* In Proc. of User Modeling.

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

X. Huang, F. Alleva, H.-w. Hon, M.-y. Hwang and R. Rosenfeld. 1993. The SPHINX-II Speech Recognition System: An Overview. *Computer, Speech and Language, 7*.

N. Iglewicz and D. Hoaglin. 1993. *How to Detect and Handle Outliers*. American Society for Quality Control, Milwaukee, WI

G. T. Jackson and A. Graesser. 2007. *Content Matters: An investigation of feedback categories within an ITS.* In Proc. of International Conference on Artificial Intelligence in Education (AIED).

P. Jordan, C. Rosé and K. VanLehn. 2001. *Tools for Authoring Tutorial Dialogue Knowledge.* In Proc. of Artificial Intelligence in Education (AIED).

D. Jurafsky and J. H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, Upper Saddle River, NJ

S. Katz, D. Allbritton and J. Connelly. 2003. Going Beyond the Problem Given: How Human Tutors Use Post-Solution Discussions to Support Transfer. *International Journal of Artificial Intelligence in Education (IJAIED), 13*.

P. A. Kirschner, J. Sweller and R. E. Clark. 2006. Why Minimal Guidance During Instruction Does Not Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching. *Educational Psychologist, 41*(2).

R. Kumar, G. Gweon, M. Joshi, Y. Cui and C. P. Rose. 2007. *Supporting students working together on Math with Social Dialogue.* In Proc. of Workshop on Speech and Language Technology in Education (SLaTe).

R. Kumar, C. P. Rosé, Y. C. Wang, M. Joshi and A. Robinson. 2007. *Tutorial Dialogue as Adaptive Collaborative Learning Support.* In Proc. of Artificial Intelligence in Education (AIED).

S. Larsson and D. Traum. 2000. Information state and dialogue management in the TRINDI Dialogue Move Engine Toolkit. *Natural Language Engineering,* (Special Issue on Best Practice in Spoken Language Dialogue Systems Engineering).

C. M. Lee, S. S. Narayanan and R. Pieraccini. 2002. *Combining acoustic and language information for emotion recognition.* In Proc. of ICSLP.

G. E. Legge, D. G. Pelli, G. S. Rubin and M. M. Schleske. 1985. Psychophysics of reading--I. Normal vision. *Vision Research, 25*(2).

E. Levin and R. Pieraccini. 2006. *Value-based optimal decision for dialogue systems.* In Proc. of IEEE/ACL Workshop on Spoken Language Technology (SLT).

E. Levin, R. Pieraccini and W. Eckert. 2000. A Stochastic Model of Human Machine Interaction for Learning Dialog Strategies. *IEEE Transactions on Speech and Audio Processing, 8:1*.

G.-A. Levow. 1998. *Characterizing and recognizing spoken corrections in human-computer dialogue.* In Proc. of COLING-ACL.

G.-A. Levow. 2004. *Prosodic Cues to Discourse Segment Boundaries in Human-Computer Dialogue.* In Proc. of SIGdial.

D. Litman and K. Forbes-Riley. 2004. *Predicting student emotions in computer-human tutoring dialogues.* In Proc. of Assoc. for Computational Linguistics (ACL).

D. Litman and K. Forbes-Riley. 2006. Recognizing Student Emotions and Attitudes on the Basis of Utterances in Spoken Tutoring Dialogues with both Human and Computer Tutors. *Speech Communication, 48*(5).

D. Litman, J. Hirschberg and M. Swerts. 2000. *Predicting Automatic Speech Recognition Performance Using Prosodic Cues.* In Proc. of NAACL.

D. Litman and S. Pan. 2002. Designing and Evaluating an Adaptive Spoken Dialogue System. *User Modeling and User-Adapted Interaction, 12*(2/3).

D. Litman, C. Rose, K. Forbes-Riley, K. VanLehn, D. Bhembe and S. Silliman. 2004. *Spoken Versus Typed Human and Computer Dialogue Tutoring.* In Proc. of Intelligent Tutoring Systems.

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

C. Lynch, K. Ashley, N. Pinkwart and V. Aleven. 2007. *Argument diagramming as focusing device: does it scaffold reading?* In Proc. of Workshop on AIED Applications for Ill-Defined Domains at the 13th International Conference on Artificial Intelligence in Education.

C. Mancini, D. Scott and S. Buckingham Shum. 2006. Visualising Discourse Coherence in Non-Linear Documents. *Traitement Automatique des Langues, Special Issue on Computational Approaches to Document and Discourse, 47*(2).

I. Mani. 2001. *Automatic Summarization*. John Benjamins Publishing,

W. C. Mann and S. A. Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. Technical Report ISI/RS-87-190, USC, Information Sciences Institute.

R. J. Marzano, B. B. Gaddy and C. Dean. 2000. *What works in classroom instruction*. Aurora, CO: Mid-continent Research for Education and Learning.,

D. S. McNamara and W. Kintsch. 1996. Learning from Texts: Effects of Prior Knowledge and Text Coherence. *Discourse Processes, 22*(3).

G. Möhler and J. Mayer. 2001. *A discourse model for pitch-range control.* In Proc. of ISCA workshop on Speech Synthesis.

S. Möller. 2005. *Parameters for Quantifying the Interaction with Spoken Dialogue Telephone Services.* In Proc. of SIGDial.

S. Möller and N. Ward. 2008. *A Framework for Model-based Evaluation of Spoken Dialog Systems.* In Proc. of Workshop on Discourse and Dialogue (SIGDial).

S. Y. Mousavi, L. Renae and J. Sweller. 1995. Reducing cognitive load by mixing auditory and visual presentation modes. *Journal of Educational Psychology, 87*(2).

R. J. Nadolski, P. A. Kirschner and J. J. G. V. Merriënboer. 2005. Optimizing the number of steps in learning tasks for complex skills. *British Journal of Educational Psychology, 75*(2).

S. Oviatt, P. Cohen, L. Wu, J. Vergo, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landry, J. Larson and D. Ferro. 2000. Designing the user interface for multimodal speech and pen-based gesture applications: State-of-the-art systems and future research di- rections. *Human Computer Interaction, 15(4).*

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.

T. Paek and E. Horvitz. 2000. *Conversation as Action Under Uncertainty.* In Proc. of Uncertainty and Artificial Intelligence (UAI).

T. Paek and E. Horvitz. 2004. *Optimizing Automated Call Routing by Integrating Spoken Dialog Models with Queuing Models.* In Proc. of HLT-NAACL.

S. Pan. 1999. *Modeling Prosody Automatically in Concept-to-Speech Generation.* In Proc. of AAAI/IAAI.

K. Papineni, S. Roukos, T. Ward and W. J. Zhu. 2002. *BLEU: a method for automatic evaluation of machine translation.* In Proc. of Association for Computational Linguistics (ACL).

R. Passonneau and D. Litman. 1997. Discourse segmentation by human and automated means. *Computational Linguistics, 23*(Special Issue on Empirical Studies in Discourse Interpretation and Generation).

R. Picard. 2003. Affective Computing: Challenges. *International Journal of Human-Computer Studies, 59*(1-2).

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.*

A. Purandare and D. Litman. 2008. *Content-Learning Correlations in Spoken Tutoring Dialogs at Word, Turn and Discourse Levels.* In Proc. of 21st International FLAIRS Conference.

A. Raux, D. Bohus, B. Langner and M. Eskenazi. 2006. *Doing Research in a Deployed Spoken Dialog System: One Year of Let's Go! Public Experience.* In Proc. of Interspeech.

A. Raux and M. Eskenazi. 2008. *Optimizing Endpointing Thresholds using Dialogue Features in a Spoken Dialogue System.* In Proc. of 9th SIGdial Workshop on Discourse and Dialogue.

A. Raux, B. Langner, D. Bohus, A. Black and M. Eskenazi. 2005. *Let's Go Public! Taking a Spoken Dialog System to the Real World.* In Proc. of Interspeech.

M. Rayner, B. A. Hockey, N. Chatzichrisafis, K. Farrell and J.-M. Renders. 2005. *A Voice Enabled Procedure Browser for the International Space Station.* In Proc. of ACL.

B. Reeves and C. Nass. 1996. *The Media Equation: How People Treat Computers, Television and New Media Like Real People and Places*.

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

C. Rich, C. L. Sidner and N. Lesh. 2001. COLLAGEN: Applying collaborative discourse theory to human-computer interaction. *AI Magazine*, *22*(Issue).

C. P. Rosé, B. D. Eugenio, L. S. Levin and C. V. Ess-Dykema. 1995. *Discourse processing of dialogues with multiple threads.* In Proc. of Proceedings of the 33rd annual meeting on Association for Computational Linguistics (ACL).

C. P. Rosé, A. Gaydos, B. S. Hall, A. Roque and K. VanLehn. 2003. *Overcomming the Knowledge Engineering Bottleneck for Understanding Student Language Input.* In Proc. of Articial Intelligence in Education (AIED).

M. Rotaru and D. Litman. 2005. *Interactions between Speech Recognition Problems and User Emotions.* In Proc. of Interspeech.

M. Rotaru and D. Litman. 2006. *Dependencies between Student State and Speech Recognition Problems in Spoken Tutoring Dialogues.* In Proc. of ACL.

M. Rotaru and D. Litman. 2006. *Discourse Structure and Speech Recognition Problems.* In Proc. of Interspeech.

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

M. Rotaru and D. Litman. 2007. *The Utility of a Graphical Representation of Discourse Structure in Spoken Dialogue Systems.* In Proc. of ACL.

N. Roy, J. Pineau and S. Thrun. 2000. *Spoken dialogue management using probabilistic reasoning.* In Proc. of 38th Annual Meeting on Association for Computational Linguistics (ACL).

A. Rudnicky, E. Thayer, P. Constantinides, C. Tchou, R. Stern, K. Lenzo, W. Xu and A. Oh. 1999. *Creating natural dialogs in the Carnegie Mellon Communicator System.* In Proc. of Eurospeech.

K. Scherer. 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication, 40*(1-2).

R. Shelby, K. Schulze, D. Treacy, M. Wintersgill, K. VanLehn and A. Weinstein. 2001. *An assessment of the Andes tutor.* In Proc. of Physics Education Research Conference.

S. Singh, D. Litman, M. Kearns and M. Walker. 2002. Optimizing Dialogue Management with Reinforcement Learning: Experiments with the NJFun System. *Journal of Artificial Intelligence Research,* (16).

G. Skantze. 2005. Exploring human error recovery strategies: Implications for spoken dialogue systems. *Speech Communication, 45*(3).

R. Smith and S. Gordon. 1997. Effects of Variable Initiative on Linguistic Behavior in Human-Computer Spoken Natural Language Dialogue. *Computational Linguistics, 23*(1).

H. Soltau and A. Waibel. 2000. *Specialized acoustic models for hyperarticulated speech.* In Proc. of ICASSP.

A. Stent. 2000. *Rhetorical structure in dialog.* In Proc. of 2nd International Natural Language Generation Conference (INLG).

A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, M. Meteer and C. V. Ess-Dykema. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics, 26:3.*

J. Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science, 11.*

J. Sweller, J. J. G. v. Merrienboer and F. G. W. C. Paas. 1998. Cognitive Architecture and Instructional Design. *Educational Psychology Review, 10*(3).

M. Swerts and E. Krahmer. 2005. Audiovisual Prosody and Feeling of Knowing. *Journal of Memory and Language, 53.*

M. Swerts, D. Litman and J. Hirschberg. 2000. *Corrections in Spoken Dialogue Systems.* In Proc. of ICSLP.

J. Tetreault and D. Litman. 2006. *Comparing the Utility of State Features in Spoken Dialogue Using Reinforcement Learning.* In Proc. of Human Language TechnologyConference/North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL).

K. VanLehn, A. C. Graesser, G. T. Jackson, P. Jordan, A. Olney and C. P. Rose. 2007. When are tutorial dialogues more effective than reading? *Cognitive Science, 31*(1).

K. VanLehn, P. W. Jordan, C. P. Rosé, D. Bhembe, M. Böttner, A. Gaydos, M. Makatchev, U. Pappuswamy, M. Ringenberg, A. Roque, S. Siler and R. Srivastava. 2002. *The Architecture of Why2-Atlas: A Coach for Qualitative Physics Essay Writing.* In Proc. of Intelligent Tutoring Systems (ITS).

K. VanLehn, C. Lynch, K. Schulze, J. A. Shapiro, R. Shelby, L. Taylor, D. Treacy, A. Weinstein and M. Wintersgill. 2005. *The Andes physics tutoring system: Five years of evaluations.* In Proc. of Artificial Intelligence in Education (AIED).

K. VanLehn, S. Siler, C. Murray, T. Yamauchi and W. B. Baggett. 2003. Why do only some events cause learning during human tutoring? *Cognition and Instruction, 21*(3).

M. Walker, D. Litman, C. Kamm and A. Abella. 1997. *PARADISE: A Framework for Evaluating Spoken Dialogue Agents.* In Proc. of Association for Computational Linguistics (ACL).

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

M. Walker, R. Passonneau and J. Boland. 2001. *Quantitative and Qualitative Evaluation of Darpa Communicator Spoken Dialogue Systems.* In Proc. of ACL.

M. Walker, A. Rudnicky, R. Prasad, J. Aberdeen, E. Bratt, J. Garofolo, H. Hastie, A. Le, B. Pellom, A. Potamianos, R. Passonneau, S. Roukos, G. Sanders, S. Seneff and D. Stallard. 2002. *DARPA Communicator: Cross-System Results for the 2001 Evaluation.* In Proc. of ICSLP.

M. Walker, J. Wright and I. Langkilde. 2000. *Using natural language processing and discourse features to identify understanding errors in a spoken dialogue system.* In Proc. of ICML.

A. Ward and D. Litman. 2006. *Cohesion and Learning in a Tutorial Spoken Dialog System.* In Proc. of 19th International FLAIRS (Florida Artificial Intelligence Research Society) Conference.

A. Ward and D. Litman. 2008. *Semantic Cohesion and Learning.* In Proc. of 9th International Conference on Intelligent Tutoring Systems (ITS).

S. Young, J. Schatzmann, K. Weilhammer and H. Ye. 2007. *The Hidden Information State Approach to Dialog Management.* In Proc. of ICASSP.

M. X. Zhou, Z. Wen and V. Aggarwal. 2005. *A Graph-Matching Approach to Dynamic Media Allocation in Intelligent Multimedia Interfaces.* In Proc. of ACM Conference on Intelligent User Interfaces (IUI).

V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T. J. Hazen and L. Hetherington. 2000. Jupiter: A Telephone-Based Conversational Interface for Weather Information. *IEEE Transactions on Speech and Audio Processing, 8*(1).