# Discourse Structure and Speech Recognition Problems

*Mihai Rotaru and Diane J. Litman*

Department of Computer Science
University of Pittsburgh, 210 S. Bouquet, Pittsburgh, PA, 15260, USA
mrotaru, litman @ cs.pitt.edu

## Abstract

We study dependencies between discourse structure and speech recognition problems (SRP) in a corpus of speech-based computer tutoring dialogues. This analysis can inform us whether there are places in the discourse structure prone to more SRP. We automatically extract the discourse structure by taking advantage of how the tutoring information is encoded in our system. To quantify the discourse structure, we extract two features for each system turn: depth of the turn in the discourse structure and the type of transition from the previous turn to the current turn. The $\chi^2$ test is used to find significant dependencies. We find several interesting interactions which suggest that the discourse structure can play an important role in several dialogue related tasks: automatic detection of SRP and analyzing spoken dialogues systems with a large state space from limited amounts of available data.

**Index Terms**: discourse structure, speech recognition analysis, spoken dialogue systems.

## 1. Introduction

With recent advancements in spoken dialogue system technologies, researchers have turned their attention to more complex domains. Interactions in these domains result in more complex dialogues with a richer underlying discourse structure. One example of such a domain is tutoring [1]. In typical information access dialogue systems, the discourse structure is relatively simple: get the information from the user and return the query results with the minimal complexity added by confirmation dialogues. In contrast, a tutoring dialogue system has to discuss concepts, laws and relationships and to engage in complex subdialogues to correct student misconceptions.

This paper is part of our ongoing work that investigates the importance of discourse structure for spoken dialogue design. In [2], we have shown that discourse structure is helpful for spoken dialogue performance modeling. Here, we study the relationship between discourse structure and speech recognition problems (**SRP**). Previous work [3] has shown that the number of SRP is negatively correlated with overall user satisfaction. Given the negative impact of SRP, there has been a lot of work in trying to understand this phenomenon through predictive models [4-6]. Acoustic, prosodic and lexical features are commonly used in these models. Usage of the discourse structure information is limited to local features (e.g. dialogue act sequencing information [4]) or flattens the discourse structure (e.g. the number of confirmation subdialogues [6]).

The main question behind our study 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 *hierarchical* aspect of the discourse structure to define the notion of "places in the dialogue", extending over previous work that ignores this information [4, 6]. We exploit the hierarchical aspect by looking at the depth and transitions in the discourse structure. To find "places" with more SRP, we use the Chi Square ($\chi^2$) test to find dependencies between depth/transition and SRP.

We find that student answers at lower levels in the discourse structure have more SRP and that certain transitions have specific interaction patterns with SRP. Our results suggest that discourse structure can play an important role in several dialogue related tasks: automatic detection of SRP and analyzing spoken dialogue systems with a large state space from limited amounts of available data.

## 2. Corpus and annotation

The corpus analyzed in this paper consists of 95 experimentally obtained spoken tutoring dialogues between 20 students and **ITSPOKE** (**I**ntelligent **T**utoring **SPOKE**n dialogue system). ITSPOKE [1] is a speech-enabled version of the text-based Why2-Atlas conceptual physics tutoring system [7]. When interacting with ITSPOKE, students first type an essay answering a qualitative physics problem using a graphical user interface. ITSPOKE then engages the student in *spoken* dialogue (using speech-based input and output) to correct misconceptions and elicit more complete explanations, after which the student revises the essay, thereby ending the tutoring or causing another round of tutoring/essay revision.

### 2.1. Speech Recognition Problems (SRP)

Our corpus is annotated for two types of SRP: rejections and ASR misrecognitions. Here we provide a brief description of each SRP; for more details see [8]. Rejections occur when ITSPOKE is not confident enough in the recognition hypothesis and asks the student to repeat (Figure 1, $TUTOR_6$). When the speech recognition hypothesis is different from what the student actually said but the system is confident in its hypothesis, we call this an ASR Misrecognition (a binary version of the commonly used Word Error Rate). For each type of SRP we define a binary variable. The top part of Table 1 lists the distribution for our two SRP variables.

### 2.2. Discourse structure

We base our annotation of discourse structure on the Grosz & Sidner theory of discourse structure [9]. A critical ingredient of this theory is the intentional structure. According to the theory,

each discourse has a discourse purpose/intention. Satisfying the main discourse purpose is achieved by satisfying several smaller purposes/intentions organized in a hierarchical structure. As a result, the discourse is segmented in discourse segments each with an associated discourse segment purpose/intention. This theory has inspired several generic dialogue managers for spoken dialogue systems [10].

Table 1. *Variable distribution*

| Variable | Values | Student turns (2334) |
|---|---|---|
| Speech recognition problems | | |
| ASR | AsrMis | 25.4% |
| MIS | noAsrMis | 74.6% |
| REJ | Rej | 7.0% |
| | noRej | 93.0% |
| Discourse structure features | | |
| DEPTH | 1 | 57.3% |
| | 2 | 27.3% |
| | 3 | 10.5% |
| | 4 | 4.8% |
| TRANS | Advance | 53.4% |
| | NewTopLevel | 13.5% |
| | PopUp | 9.2% |
| | PopUpAdv | 3.5% |
| | Push | 14.5% |
| | SameGoal | 5.9% |

We automate our annotation of the discourse structure by taking advantage of the structure of the tutored information. 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 structure (see Figure 1). 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.

In Figure 1 we illustrate ITSPOKE's behavior and our discourse structure annotation. 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). 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 into 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). 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 ($Tutor_5$).

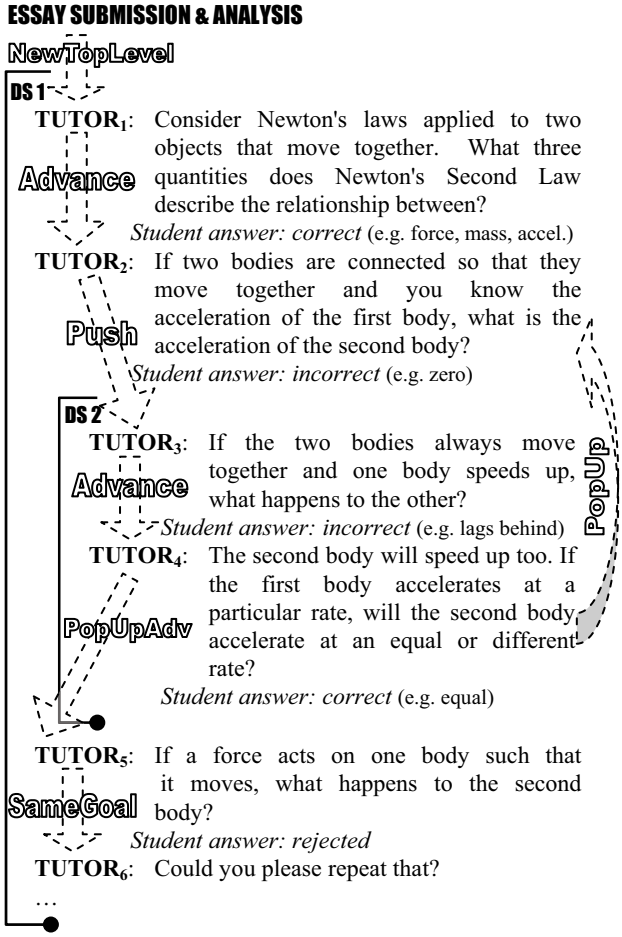


Figure 1. *The discourse structure and transition annotation*

In this paper we want to study if the position in the discourse structure interacts with SRP. We quantify the position using two features: **depth** and **transition**. The depth feature captures the vertical position in the discourse structure while the transition feature captures the horizontal relative position. For each system turn, we define its depth as the depth of its discourse segment in the discourse structure ($Tutor_{1,2,5,6}$ will have depth 1 while $Tutor_{3,4}$ will have depth 2). In our corpus the maximum depth is 6; because of the small number of turns at level 5 and 6 we choose to collapse levels 4, 5 and 6 in a single level, level 4. For our interaction experiment we define the variable DEPTH with 4 values (1-4).

The transition feature captures the position in the discourse structure of the current system turn relative to the previous system turn. We define six labels. **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 $Tutor_2$. In addition, variations of "Ok, back to the original question" are also included in the new system turn 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. $Tutor_5$). Note that while the sum of PopUp and PopUpAdv should be equal with Push, it is smaller in our corpus because in some cases ITSPOKE popped up more than one level in the discourse structure hierarchy. In case of rejections, the system question is repeated using variations of "Could you please repeat that?". We label such cases as **SameGoal** (e.g. $Tutor_6$). For our interaction experiments we define the variable TRANS with the six values.

Please note that each student dialogue has a specific discourse structure based on the dialogue that dynamically emerges based on the correctness of her answers. For this reason, the same system question in terms of content may get a different depth and transition label for different students. Also, while a human annotation of the discourse structure will be more complex but more time consuming, its advantages are outweighed by the automatic nature of our annotation.

## 3. Identifying dependencies using $\chi^2$

To discover interactions between our variables we apply the $\chi^2$ test. The test assesses whether the differences between observed and expected counts are large enough to conclude a statistically significant dependency between two variables. The $\chi^2$ test has been used by to produce interesting insights about human-human conversations [11]. It was also used successfully in our previous analysis of SRP [8].

Table 2. *Interactions between TRANS and ASRMIS*

| Combination | | Obs. | Exp. | $\chi^2$ |
|---|---|---|---|---|
| TRANS – ASRMIS | | | | 23.88 |
| NewTopLevel – AsrMis | - | 61 | 79 | 6.75 |
| PopUp – AsrMis | + | 74 | 54 | 10.56 |
| Push – AsrMis | + | 106 | 85 | 7.3 |

First we find significant dependencies between our variables. Next, for each significant dependency, we look more deeply into the overall interaction by investigating how particular variable's values interact with each other. To do that, we compute a binary variable for each variable's value and study dependencies between these variables. For example, for the value PopUp of the variable TRANS we create a binary variable with two values: PopUp and 'Anything Else' (the other five transitions). By studying the dependency between these binary variables we can understand how the interaction works.

Table 2 reports in rows 3-5 all significant interactions between the values of variables TRANS and ASRMIS. Each row shows: 1) the value for each variable, 2) the sign of the dependency, 3) the observed counts, 4) the expected counts and 5) the $\chi^2$ value. For example, in our data there are 74 PopUp transitions that will result in a AsrMis. This value is larger than the expected counts (54); the dependency between PopUp and AsrMis is significant with a $\chi^2$ value of 10.56 ($p<0.05$ for a $\chi^2$ value larger than 3.84; $p<0.01$ for $\chi^2$ value larger than 6.63). A comparison of the observed counts and expected counts reveals the direction (sign) of the dependency. In our case we see that after Push and PopUp there are more AsrMis than expected (rows 4-5), while after NewTopLevel there are less AsrMis than expected (row 3). On the other hand, there is no interaction between the other three transitions (Advance, PopUpAdv and SameGoal) and AsrMis. In other words, the interaction between the TRANS and ASRMIS is explained only by the three interactions listed in the table.

## 4. Results

In this section we present all significant dependencies ($p<0.05$) between the two discourse structure features and SRP. We begin by looking at the interaction between the depth of the system question and the presence of SRP in the subsequent student answer. We find only one significant dependency: DEPTH–REJ (Table 3). We find that at level 1 there are less rejections than expected while at level 3 and 4 there are more rejections than expected. In other words, our data indicates that rejections are more likely to happen at lower levels than at higher levels.

Table 3. *Interactions between DEPTH and REJ*

| Combination | | Obs. | Exp. | $\chi^2$ |
|---|---|---|---|---|
| DEPTH – REJ | | | | 80.30 |
| 1 – Rej | - | 52 | 93 | 47.31 |
| 3 – Rej | + | 42 | 17 | 43.89 |
| 4 – Rej | + | 20 | 7 | 21.41 |

Several hypotheses can explain this interaction. One hypothesis is that lower levels address deeper student knowledge gaps and at these levels the student is more likely to be incorrect, uncertain or even frustrated. Our ongoing work shows that the student state, in terms of correctness and affect, also correlates with SRP [12]. Another hypothesis is that the automatic speech recognition component is less competent at these levels. This might be due to the smaller size of training data for lower levels: all students go through all questions at level one but only those with more knowledge gaps go through lower levels (see the skewed distribution of DEPTH in Table 1). Also, lower levels might have more out-of-vocabulary words which is another potential source of problems.

Table 4. *Interactions between TRANS and REJ*

| Combination | | Obs. | Exp. | $\chi^2$ |
|---|---|---|---|---|
| TRANS – REJ | | | | 383.15 |
| Advance – Rej | - | 45 | 87 | 46.95 |
| NewTopLevel – Rej | - | 12 | 21 | 5.58 |
| SameGoal – Rej | + | 66 | 9 | 376.63 |

Next we look at how the type of transition to the current system question interacts with SRP in the student answer. Here we find that TRANS interact with both ASR MIS (recall Table 2) and REJ (Table 4). 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 [12].

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 [13]. Interestingly, Push and PopUp interact with AsrMis but do not interact with Rej.

In terms of rejections (Table 4), we find that starting a new tutoring dialogue or advancing at the same level 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 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 [8]: 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 (NewTopLevelGoal and Advance). This observation provides additional support for the rejection handling strategy we proposed in [8] for our domain: do not reject but keep the conversation going. This strategy is on par with observations on human-human dialogues [11].

## 5. Discussion

Our results suggest that discourse structure is an important information source for dialogue related tasks. Previous work on automatic detection of SRP has focused primarily on acoustic, prosodic, lexical and simple discourse features [4-6]. The specific interaction patterns we observe in our data suggest that a hierarchical model of discourse structure can be an informative feature for predictive models of SRP.

In terms of spoken dialogue systems analysis, discourse structure can help with data sparsity problems. Our system has 254 unique states (i.e. system questions). Given the relatively small size of our corpus, 2334 system turns, it is impossible to perform an analysis for each system state. By providing a level of abstraction over individual system states, the discourse structure allows us to perform a meaningful analysis of our corpus with interesting results. 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 tutor questions involved in the interaction is between 15 and 47 with no tutor question from this set being repeated in our corpus more than 10-15 times.

From the dialogue designer perspective, our results suggest that particular attention should be paid to specific locations in the discourse structure. For example, for our system, more effort should be spent in designing lower level subdialogues. The interactions between Push/PopUp and SRP suggest that increasing student awareness of the discourse structure through lexical and prosodic means [13] might also be beneficial.

## 6. Conclusions

We investigate the role of discourse structure in characterizing SRP via the $\chi^2$ dependency test. We automatically compute an approximation of the Grosz & Sidner discourse structure [9] by using the inherent structure of the tutoring information encoded in our system. To quantify the discourse structure, we extract two features for each system turn: depth of the turn in the discourse structure and the type of transition from the previous system turn to the current turn. The depth feature captures the vertical position in the discourse structure while the transition feature captures the horizontal relative position. We find that student answers at lower levels in the discourse structure have more SRP and that certain transitions have specific interaction patterns with SRP (e.g. Push and PopUp transitions have problematic interactions with AsrMis).

Our results suggest that the discourse structure can play an important role in several dialogue related tasks: automatic detection of SRP and analyzing spoken dialogues systems with a large state space from limited amounts of available data.

In the future, we would like to build a SRP prediction model for our system and measure the improvement offered by the discourse structure features. Testing if our results generalize to a manual annotation of the discourse structure is another important future step. We also plan to investigate if our analysis is useful for information access dialogue systems.

## 7. Acknowlegments

This work is supported by NSF Grants No. 0328431 and 0428472. We would like to thank members of the NLP@Pitt group for their helpful comments.

## 8. References

[1] D. Litman and S. Silliman, "ITSPOKE: An intelligent tutoring spoken dialogue system", HLT/NAACL, 2004.

[2] M. Rotaru and D. Litman, "Exploiting Discourse Structure for Spoken Dialogue Performance Analysis", EMNLP, 2006.

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

[4] M. Gabsdil and O. Lemon, "Combining Acoustic and Pragmatic Features to Predict Recognition Performance in Spoken Dialogue Systems", ACL, 2004.

[5] J. Hirschberg, D. Litman, and M. Swerts, "Prosodic and Other Cues to Speech Recognition Failures," *Speech Communication*, vol. 43, pp. 155-175, 2004.

[6] M. Walker, J. Wright, and I. Langkilde, "Using natural language processing and discourse features to identify understanding errors in a spoken dialogue system", ICML, 2000.

[7] K. VanLehn, P. W. Jordan, C. P. Rosé, et al., "The Architecture of Why2-Atlas: A Coach for Qualitative Physics Essay Writing", Intelligent Tutoring Systems (ITS), 2002.

[8] M. Rotaru and D. Litman, "Interactions between Speech Recognition Problems and User Emotions", Interspeech, 2005.

[9] B. Grosz and C. L. Sidner, "Attentions, intentions and the structure of discourse," *Computational Lingustics*, vol. 12, 1986.

[10] D. Bohus and A. Rudnicky, "RavenClaw: Dialog Management Using Hierarchical Task Decomposition and an Expectation Agenda", Eurospeech, 2003.

[11] G. Skantze, "Exploring human error recovery strategies: Implications for spoken dialogue systems," *Speech Communication*, vol. 45, 2005.

[12] M. Rotaru and D. Litman, "Dependencies between Student State and Speech Recognition Problems in Spoken Tutoring Dialogues", ACL, 2006.

[13] J. Hirschberg and C. Nakatani, "A prosodic analysis of discourse segments in direction-giving monologues", ACL, 1996.