Content-Learning Correlations in Spoken Tutoring Dialogs at Word, Turn and Discourse Levels

Amruta Purandare and **Diane Litman**
Intelligent Systems Program
University of Pittsburgh
{amruta,litman}@cs.pitt.edu

Abstract

We study correlations between dialog content and learning in a corpus of human-computer tutoring dialogs. Using an online encyclopedia, we first extract domain-specific concepts discussed in our dialogs. We then extend previously studied shallow dialog metrics by incorporating content at three levels of granularity (word, turn and discourse) and also by distinguishing between students' spoken and written contributions. In all experiments, our content metrics show strong correlations with learning, and outperform the corresponding shallow baselines. Our word-level results show that although verbosity in student writings is highly associated with learning, verbosity in their spoken turns is not. On the other hand, we notice that content along with conciseness in spoken dialogs is strongly correlated with learning. At the turn-level, we find that effective tutoring dialogs have more content-rich turns, but not necessarily more or longer turns. Our discourse-level analysis computes the distribution of content across larger dialog units and shows high correlations when student contributions are rich but unevenly distributed across dialog segments.

Motivation

One important line of research in Intelligent Tutoring Systems (ITS) focuses on integrating and enhancing natural language dialog support in computer-aided tutoring (Graesser *et al.* 2004) (VanLehn *et al.* 2002) (Litman & Silliman 2004) (Pon-Barry *et al.* 2004) (Evens *et al.* 2001). This has indirectly motivated researchers to study how specific dialog features affect learning, so that system development can be founded on a solid empirical analysis. For example, (Core, Moore, & Zinn 2003) design a Socratic style tutor to encourage student initiative, based on the observation that student language production is often correlated with learning.

Previous studies on Dialog-Learning correlations have primarily looked at surface-level (shallow) dialog features (Litman *et al.* 2004) and deep dialog moves (Graesser, Person, & Magliano 1995) (Rosé *et al.* 2003) (Forbes-Riley *et al.* 2005). Shallow measures (such as dialog length or average words per turn etc.) can be computed automatically, and hence, are easy to incorporate in an adaptive tutoring system. Studies of (Litman *et al.* 2004), however, show that these shallow measures are mostly associated with learning on human-human typed dialogs, and do not generalize to human-computer and/or spoken tutoring dialogs. Dialog moves (such as feedback, hint or type of question/answer) have a strong impact on learning (Forbes-Riley *et al.* 2005) (Jackson, Person, & Graesser 2004) (Core, Moore, & Zinn 2003) (Rosé *et al.* 2003), however, are not easy to measure automatically and often require manual labeling.

In this paper, we propose content-based dialog features that are more informative than shallow dialog metrics, and also easy to compute automatically. Recent studies of (Forbes-Riley *et al.* 2007) have shown that simple semantic features that account for dialog content are some of the best predictors of student learning, more so than advanced features like dialog moves. While (Forbes-Riley *et al.* 2007) compare the relative utility of different feature types, our work exclusively focuses on content metrics and their correlations with learning. We extend previously studied shallow dialog features with *content* at three levels of granularity: word, turn and discourse. Our analysis also addresses multiple aspects of student-tutor interaction by considering each speaker's individual contributions as well as students' written essays, in addition to their spoken dialogs. Our content metrics outperform their corresponding shallow baselines at all three levels of granularity and thus support our hypothesis that learning is better associated with dialog content than its length. As none of our metrics rely on human annotations, they can be easily applied to any corpora or domain of interest.

Data

Our corpus consists of a total of 100 human-computer spoken tutoring dialogs between 20 university students and ITSPOKE (Litman & Silliman 2004), a speech enabled version of the Why-2 Atlas conceptual physics tutoring system (VanLehn *et al.* 2002). At the beginning of each dialog, the tutor first presents a student with a qualitative physics problem that requires an essay-like descriptive answer. After the

student enters the initial essay, the tutor engages the student in a spoken dialog to elaborate missing details, corrects student misconceptions, and eventually asks the student to revise the essay. The cycle then continues until the student writes a correct and complete essay. The current system is designed to tutor 5 physics problems, i.e. each student participates in 5 dialogs. The essays are written, whereas, the dialogs are spoken. There are a total of 5239 spoken dialog turns and 401 essays in our corpus. The turn-count in individual dialogs varies from 10 to 160. The overall corpus has a total of 178,569 words. The number of words in individual dialogs ranges from 400 to 5000.

In addition to analyzing entire tutoring sessions, we also look at each speaker's individual contributions, similar to previous studies of (Litman *et al.* 2004). Moreover, we also distinguish between students' spoken and written contributions by analyzing their spoken dialog turns and written essays, separately as well as together. Thus our analysis involves 6 cases as described below:

- **S+T+E**: Entire tutoring session, including the spoken dialog between a student and the tutor, and the essays written by the student.
- **S+T**: Only the spoken dialog between a student and the tutor.
- **S+E**: Total student contributions, including students' spoken dialog turns and written essays.
- **S**: Students' spoken dialog turns.
- **T**: Tutor's spoken dialog turns.
- **E**: Students' written essays.

These cases represent our method of analyzing student-centered (S, E, S+E), tutor-centered (T) and interaction-centered (S+T, S+T+E) hypotheses of (Chi *et al.* 2001).

Table 1 shows a summary of our corpus for all six cases. Rows represent various metrics presented in this paper (discussed later) and the numbers show their averages over all 20 students. For example, the entry at #Words and S+T+E shows that there are approximately 8928 words on an average discussed per student, considering the entire tutoring session. Of these, 2583 are contributed by the student (S+E) and 6345 by the tutor (T).

Automatic Concept Extraction

To automatically extract domain-specific physics concepts that appear in our corpus, we use an online physics dictionary from *"Eric Weisstein's World of Physics"* (http://scienceworld.wolfram.com/physics). There are a total of 3040 physics concepts defined in this dictionary, including multi-word concepts such as *Newton's Second Law, Frictional Force* etc. At present, our notion of *concepts* is limited to simple unigram (single word) concepts in their singular noun-forms as defined in the above dictionary. In our tutoring corpus, there are 56 unique physics concepts discussed in total. Figure 1 shows a dialog excerpt from our corpus with physics concepts underlined.

Tutor: Right. Let's try the original question again. How does the horizontal displacement of the pumpkin compare with the man's while it is flying?
Student: Same
Tutor: Yep. So where will the pumpkin land relative to the man?
Student: Same spot
Tutor: Correct. To summarize: the basic strategy that we used to solve this problem was to divide it in three different time intervals: before the toss, during the toss, and after the toss. For the second and third time intervals we looked at the forces applied on the pumpkin, the net force, and the acceleration of the pumpkin. Finally, for the flight of the pumpkin we also analyzed its velocity and displacement and by following that chain of reasoning we were able to conclude that it would land in the man's hand. What we've talked about should help you. Please try to write an explanation now. Please remember to press submit.
Essay: The pumpkin will land in the same place relative to the man. While the man is carrying the pumpkin its velocity is the same as the man's. Here the vertical velocity is zero. In other words, the velocity is constant and in the horizontal direction. Now, during the toss, the two vertical forces exerted on the pumpkin are gravity (down) and the man (up). And the NET force is in the upward direction. The direction of the pumpkin's acceleration is vertically up. The vertical component of the pumpkin's velocity will increase and the change in the horizontal component of the pumpkin's velocity will be zero. The horizontal velocities are the same and their horizontal displacements are the same and therefore, it will land in the man's hand.
. . . [Discourse Boundary] . . .
Tutor: You have the correct answer but you need to show even more of your reasoning in your essay. Maybe this will help you remember some of the details need in the explanation. What is the relationship between the velocity of the runner and the velocity of the pumpkin, which the runner is holding, before the runner releases the pumpkin?
Student: Same

Figure 1: Dialog Excerpt

Experiments

Our corpus was collected using the same experimental procedure as in other studies: students first take a pre-test to test their prior physics knowledge, work with the tutor using a voice and web interface, and finally take a post-test that is similar to the pre-test, after finishing all tutoring sessions. Students showed significant learning in our corpus ($F(1,19)=26.94$, $p=0.000$), however, their pre-test and post-test scores are significantly correlated ($R=0.462$, $p=0.04$). We, therefore, measure the partial correlation between students' post-test scores and various metrics (presented below), controlled for their pre-test scores by regressing it out of the correlation. These correlations are computed on all 20 students in our data. Also, all the metrics presented here are computed over all 5 dialog sessions with a student. Correla-

Table 1: Corpus Summary: Table shows Mean values per Student

Word-Level	S+T+E	S+T	S+E	S	T	E
#Words	8928.5	6641.8	2583.5	296.8	6345	2286.7
#Concepts	686.3	482.5	233.2	29.4	453.1	203.8
Concepts/Words	0.0754	0.0715	0.0887	0.0984	0.0702	0.0872
Coverage	27.4	25.7	15.3	8.9	24.9	12.8
Avg Repeats	25	18.7	14.9	3.3	18.1	15.1
Turn-Level	S+T+E	S+T	S+E	S	T	E
#Turns	282	262	134.5	114.5	147.5	20
Words/Turn	31.7	25.5	19.1	2.6	43.1	110.2
Zero	122.2	121.3	92.2	91.3	30	0.9
One+	159.9	140.7	42.4	23.2	117.5	19.2
Concepts/Turn	2.4	1.8	1.7	0.26	3	9.7
Discourse-Level	S+T+E	S+T	S+E	S	T	E
Avg-Words	472.8	356.7	132.3	16.2	340.5	116.1
Avg-Concepts	36.1	25.8	11.8	1.6	24.3	10.2
Max-Words	1611.5	1484.9	237.2	81.8	1421.5	210
Max-Concepts	145.4	134.6	24.7	10.9	125.2	20.1
StdDev-Words	409.5	382	76.5	20.3	363.3	67.9
StdDev-Concepts	36.1	33.9	7.4	2.5	31.7	6.4

tions significant at 0.05 level ($p<=0.05$) are marked in boldface, and those below 0.01 ($p<=0.01$) are shown in boldface and italics. In the next three sections, we present our word, turn and discourse level concept metrics by measuring their correlations with learning.

Word Level Concept Analysis

First, as a baseline, we tried a simple shallow metric, the number of words (#Words) in dialogs, that was used by (Litman *et al.* 2004). Similar to their results, we see that there is no strong correlation between #Words in dialog turns (S, T and S+T) and learning (See Table 2). However, we do see that the number of words in student essays (E) and in the overall student contributions (S+E) are significantly correlated with learning. This result suggests that in human-computer tutoring, *even if verbosity in spoken dialogs isn't associated with learning, verbosity in students' writings is*.

Next, instead of counting all words, we ignore non-physics words, and count only domain-specific physics concepts (#Concepts) discussed with each student. As one can notice in Table 2, the #Concepts shows better correlation to learning (than the #Words) on almost all cases. This suggests that *learning is more associated with the content discussed in dialogs, than dialog lengths*. Also note that, there is a significant correlation between the #Concepts (but not the #Words) in student turns (S) and their performance. This leads us to conclude that *content and not verbosity in student turns is associated with their learning*. We also see that the #Concepts discussed in the overall tutoring session (S+T+E) is strongly correlated with learning, but not the #Words.

We next compute the concept-to-word ratio, by normalizing #Concepts over #Words, for each student. This metric essentially measures the content-to-verbosity ratio, and a higher value indicates an ability to convey more content in a concise manner. Note in Table 2 that, for the overall spoken dialog (S+T) and tutor turns (T), #Words and #Concepts did not show a significant correlation to learning, but the Concept/Word ratio does. On the other hand, on S+E and E, we see an opposite pattern; even though #Words and #Concepts in S+E and E showed a significant correlation, the Concept/Word ratio doesn't. Also notice that, on all six cases, there is a complementary pattern of significant correlations for #Words and the Concept/Word ratio. This makes sense as the #Words measures verbosity, whereas the Concept/Word ratio prefers conciseness. Column S indicates that *not only the amount of content but also conciseness in student turns has an association with their learning*. Columns S, T and S+T together suggest that *content-rich but concise spoken dialogs are more effective in tutoring*.

The metrics discussed so far count physics tokens. We now measure the *Coverage* by counting concept types (or the number of unique concepts) per student. For example, the dialog excerpt in figure 1 has a total of 17 physics tokens (#Concepts), but only 6 of them are unique (Coverage). As Table 2 shows, there are no significant correlations between the coverage of concepts and learning, for any parts of dialogs. This along with our previous results on #Concepts leads us to hypothesize that even if the coverage of concepts during tutoring sessions has no correlation to learning, the number of times they are repeated does. To test this, we measure the average number of times a concept is repeated per student (See *Avg Repeats* in Table 2). This is essentially the ratio of #Concepts to Coverage for each student. The significant correlation for tutor turns (T) proves that *the more the tutor repeats the concepts, the better the students learn*. In other words, we find that concept repetitions in dialogs are positively associated with learning, as evidenced by high correlations on dialog turns (S+T+E, S+T and T).

Table 2: Results: Word-Level Correlations with Learning

Shallow	S+T+E	S+T	S+E	S	T	E
#Words	R=0.44 p=0.059	R=0.297 p=0.217	**R=0.516** **p=0.024**	R=0.393 p=0.096	R=0.287 p=0.234	**R=0.504** **p=0.028**
Content	S+T+E	S+T	S+E	S	T	E
#Concepts	**R=0.522** **p=0.022**	R=0.427 p=0.069	**R=0.54** **p=0.017**	***R=0.685*** ***p=0.001***	R=0.385 p=0.103	**R=0.498** **p=0.03**
Concept/Word	***R=0.671*** ***p=0.002***	***R=0.608*** ***p=0.006***	R=0.394 p=0.095	**R=0.509** **p=0.026**	**R=0.564** **p=0.012**	R=0.258 p=0.286
Coverage	R=0.061 p=0.803	R=0.053 p=0.831	R=0.371 p=0.118	R=0.437 p=0.061	R=0.083 p=0.734	R=0.427 p=0.068
Avg Repeats	**R=0.551** **p=0.014**	**R=0.511** **p=0.025**	R=0.431 p=0.066	R=0.268 p=0.267	**R=0.455** **p=0.05**	R=0.407 p=0.084

Table 3: Results: Turn-Level Correlations with Learning

Shallow	S+T+E	S+T	S+E	S	T	E
#Turns	R=0.243 p=0.316	R=0.234 p=0.335	R=0.233 p=0.336	R=0.213 p=0.381	R=0.25 p=0.302	R=0.258 p=0.286
Words/Turns	R=0.262 p=0.279	R=-0.029 p=0.906	R=0.345 p=0.149	R=0.203 p=0.404	R=-0.032 p=0.897	R=0.446 p=0.056
Content	S+T+E	S+T	S+E	S	T	E
Zero	R=0.007 p=0.978	R=0.006 p=0.981	R=0.027 p=0.911	R=0.026 p=0.916	R=-0.051 p=0.837	R=0.035 p=0.886
One+	**R=0.467** **p=0.044**	**R=0.472** **p=0.042**	***R=0.649*** ***p=0.003***	***R=0.693*** ***p=0.001***	R=0.345 p=0.148	R=0.242 p=0.319
Concepts/Turns	**R=0.53** **p=0.02**	**R=0.502** **p=0.028**	R=0.44 p=0.059	**R=0.545** **p=0.016**	**R=0.456** **p=0.05**	**R=0.474** **p=0.04**

Table 4: Results: Discourse-Level Correlations with Learning

Shallow	S+T+E	S+T	S+E	S	T	E
Avg-Words	R=0.114 p=0.643	R=-0.082 p=0.74	R=0.466 p=0.055	R=0.049 p=0.841	R=-0.09 p=0.715	R=0.436 p=0.062
Max-Words	R=-0.1 p=0.683	R=-0.129 p=0.598	***R=0.641*** ***p=0.003***	R=0.299 p=0.214	R=-0.152 p=0.534	***R=0.643*** ***p=0.003***
StdDev-Words	R=-0.003 p=0.99	R=-0.028 p=0.908	**R=0.491** **p=0.033**	R=0.263 p=0.277	R=-0.045 p=0.854	**R=0.485** **p=0.035**
Content	S+T+E	S+T	S+E	S	T	E
Avg-Concepts	R=0.319 p=0.182	R=0.081 p=0.742	**R=0.534** **p=0.018**	R=0.445 p=0.056	R=0.046 p=0.852	**R=0.469** **p=0.043**
Max-Concepts	R=0.288 p=0.231	R=0.216 p=0.375	***R=0.711*** ***p=0.001***	**R=0.516** **p=0.024**	R=0.166 p=0.497	***R=0.609*** ***p=0.006***
StdDev-Concepts	R=0.35 p=0.142	R=0.308 p=0.199	***R=0.594*** ***p=0.007***	***R=0.607*** ***p=0.006***	R=0.253 p=0.295	**R=0.479** **p=0.038**

Turn Level Concept Analysis

Following (Litman *et al.* 2004), we again create shallow baselines by counting the total number of turns (S+T+E), the number of spoken dialog turns (S+T), student turns (S+E), student's spoken turns (S), tutor turns (T), student's written essays (E)[1], and also the average number of words per turn (Words/Turns) for each student. The results on these metrics are shown in Table 3, and as we can notice, these shallow metrics do not show any significant correlations to learning.

We then compute for each student, the number of turns that discuss no physics concept (Zero), at least one concept (One+), and the average number of concepts per turn (Concepts/Turns). For example, in our dialog excerpt (figure 1), there are a total of 8 turns, of which 4 have no physics concepts (1 tutor and 3 student) and 4 have at least one concept (3 tutor and 1 essay). The Concepts/Turns ratio for this excerpt is 2.1 (17/8).

As Table 3 shows, the number of turns with no physics concepts (Zero) have no strong correlation with learning. But we do see many significant correlations for the number of turns that discuss at least one physics concept (One+). Thus by separating dialog turns that discuss physics concepts from those that don't, we notice significant improvements in correlations. Also notice that, the average number of words per turn has no significant correlations to learning, but the average number of concepts per turn does. This result suggests that *effective tutoring dialogs have more content-rich turns* (as exhibited by strong correlations on One+ and Concepts/Turns metrics), *but not necessarily more or longer turns* (as there are no significant correlations on shallow metrics, #Turns and Words/Turns).

Discourse Level Concept Analysis

We now extend our content metrics to span over larger dialog units (beyond words and turns). While (Rotaru & Litman 2006) study more complex hierarchical discourse structures, we, here, apply a simple linear discourse segmentation scheme by marking discourse boundaries after each essay submission. Our dialogs are structured in such a way that there is a change in the discourse purpose (Grosz & Sidner 1986) (Grosz & Hirschberg 1992) after each essay submission. Once a student submits an essay, the tutor scans it, discusses missing details and misconceptions with the student, asks the student to revise the answer, and the cycle continues. Thus, essays mark break-points or milestones in our tutoring dialogs, where a new round of discussion starts over. Note that, this simple and automatic discourse segmentation scheme captures only high-level segments, and may miss fine-level discourse changes between two essays.

Using this dialog segmentation scheme, we then compute for each student, the number of physics concepts per dialog segment. For example, our excerpt in figure 1 shows one discourse boundary, and of the 17 physics concepts that appear in this excerpt, 15 belong to the first segment and 2 to the second. Figure 2 shows a plot of #physics-concepts (y-axis) in each dialog segment (x-axis) for one of the students in our corpus. The segments are numbered in the same order

[1]Each essay is counted as one turn.

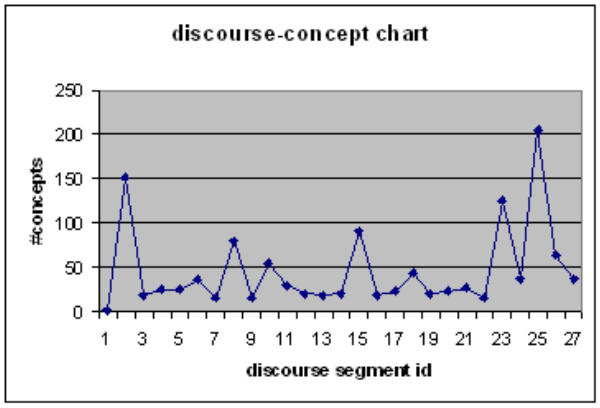

Figure 2: Discourse Analysis: Plot of #Concepts per Discourse Segment

as they appear in the dialogs. From this chart, we then compute for each student, the average number of concepts discussed per segment (Avg), the maximum (Max) number of concepts discussed in one segment (this corresponds to the highest peak in the graph), and the standard deviation which distinguishes between uniform and skewed distribution of concepts across dialog segments. Similar to the word and turn level analyses, we create shallow baselines by counting the number of words per segment. In table 3, we saw that there was no significant correlation between the number of essay revisions (or discourse segments) and learning (see the entry at #Turns and E).

Table 4 shows that the average number of words per discourse segment are not strongly associated with learning, but we see a significant correlation between the average number of concepts per segment and learning, for total student contributions (S+E) and essays (E). Metrics Max and StdDev show strong correlations on both words and concepts in S+E and E. On students' spoken turns (S), however, we see no correlations for metrics Max-Words and StdDev-Words, but we do see strong correlations on Max-Concepts and StdDev-Concepts. This suggests that *dialogs that show content-rich but unevenly distributed student contributions are more effective than those in which students contributions are uniformly poor*. Note that, the discourse analysis shows some of the best correlations on student contributions (S+E, E), compared to the word and turn level results. This makes us believe that this higher level analysis is important and should be accounted for in dialog-learning analysis.

Conclusions

In this paper, we examined correlations between dialog content and learning, in a corpus of human-computer tutoring dialogs, at three levels of granularity; word, turn and discourse. Our content features are simple domain-specific concepts that are extracted automatically using an online encyclopedia. Our results support that dialog content is strongly correlated with learning, more so than shallow features like dialog length. We also notice that content, con-

ciseness as well as concept repetitions lead to better learning, and that effective tutoring dialogs show more content-rich turns, but not necessarily more or longer turns. We also showed that the distribution of student contributions across discourse segments is associated with their learning.

Our work is also novel to distinguish between students' spoken and written contributions, and highlights some interesting complementary patterns on the two. For example, we notice that verbosity in student writings is associated with learning, but verbosity in their spoken dialog turns is not. On the other hand, both content and conciseness in spoken dialogs (but not in written essays) are strongly correlated with learning. In short, we find that most metrics that show significant correlations on spoken dialogs do not show the same on written essays, and vice-a-versa.

While our notions of *content* and *discourse* are currently very simple, our analysis is fully automated and shows many interesting results. In the future, we would like to explore features based on n-gram patterns of concepts as well as more sophisticated probabilistic content models like (Barzilay & Lee 2004), to take into account the order in which concepts are discussed in dialogs. We also plan to extend our discourse-level analysis with hierarchical discourse structures similar to (Rotaru & Litman 2006), and study other features of discourse such as cohesion (Ward & Litman 2006) and coherence (Higgins *et al.* 2004). Finally, we would like to use our current correlation results to generate hypotheses that test for causality, i.e. consciously manipulate the tutor to encourage content-rich, concise dialogs or verbose essays, and see if that actually improves learning.

Acknowledgments

Authors would like to thank the anonymous reviewers and members of the ITSPOKE group for their valuable comments and feedback. This research is partially supported by the ONR Grant N00014-04-1-0108.

References

Barzilay, R., and Lee, L. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of the HLT/NAACL Conference*, 113–120.

Chi, M.; Siler, S.; Jeong, H.; Yamauchi, T.; and Hausmann, R. 2001. Learning from human tutoring. *Cognitive Science* 25:471–533.

Core, M.; Moore, J.; and Zinn, C. 2003. The role of initiative in tutorial dialogue. In *Proceedings of the EACL Conference*.

Evens, M.; Brandle, S.; Chang, R.; Freedman, R.; Glass, M.; Lee, Y.; Shim, L.; Woo, C.; Zhang, Y.; Zhou, Y.; Michaeland, J.; and Rovick, A. 2001. Circsim-tutor: An intelligent tutoring system using natural language dialogue. In *Proceedings of the Midwest Artificial Intelligence and Cognitive Science Conference*, 16–23.

Forbes-Riley, K.; Litman, D.; Huettner, A.; and Ward, A. 2005. Dialogue-learning correlations in spoken dialogue tutoring. In *Proceedings of the AIED Conference*.

Forbes-Riley, K.; Litman, D.; Purandare, A.; Rotaru, M.; and Tetreault, J. 2007. Comparing linguistic features for modeling learning in computer dialogue tutoring. In *Proceedings of the AIED Conference*.

Graesser, A.; Lu, S.; Jackson, G.; Mitchell, H.; Ventura, M.; Olney, A.; and Louwerse, M. 2004. Autotutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers* 36:180–193.

Graesser, A.; Person, N.; and Magliano, J. 1995. Collaborative dialog patterns in naturalistic one-on-one tutoring. *Applied Cognitive Psychology* 9:1–28.

Grosz, B., and Hirschberg, J. 1992. Some intonational characteristics of discourse structure. In *Proceedings of the International Conference on Spoken Language Processing*, volume 1, 429–432.

Grosz, B., and Sidner, C. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics* 12(3):175–204.

Higgins, D.; Burstein, J.; Marcu, D.; and Gentile, C. 2004. Evaluating multiple aspects of coherence in student essays. In *Proceedings of the HLT/NAACL Conference*.

Jackson, G.; Person, N.; and Graesser, A. 2004. Adaptive tutorial dialogue in autotutor. In *Proceedings of Workshop on Dialog-based Intelligent Tutoring Systems at Intelligent Tutoring Systems*.

Litman, D., and Silliman, S. 2004. ITSPOKE: An intelligent tutoring spoken dialogue system. In *Companion Proceedings of the HLT/NAACL Conference*.

Litman, D.; Rose, C.; Forbes-Riley, K.; VanLehn, K.; Bhembe, D.; and Silliman, S. 2004. Spoken versus typed human and computer dialog tutoring. In *Proceedings of the International Conference on Intelligent Tutoring Systems*.

Pon-Barry, H.; Clark, B.; Bratt, E.; Schultz, K.; and Peters, S. 2004. Evaluating the effectiveness of SCoT: A spoken conversational tutor. In *Proceedings of Workshop on Dialogue-based Intelligent Tutoring Systems: State of the Art and New Research Directions*.

Rosé, C.; Bhembe, D.; Siler, S.; Srivastava, R.; and VanLehn, K. 2003. The role of why questions in effective human tutoring. In *Proceedings of the AIED Conference*.

Rotaru, M., and Litman, D. 2006. Exploiting discourse structure for spoken dialogue performance analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

VanLehn, K.; Jordan, P.; Rosé, C.; Bhembe, D.; Bottner, M.; Gaydos, A.; Makatchev, M.; Pappuswamy, U.; Ringenberg, M.; Roque, S.; Siler, R.; Srivastava, R.; and Wilson, R. 2002. The architecture of why2-atlas: A coach for qualitative physics essay writing. In *Proceedings of Intelligent Tutoring Systems Conference*.

Ward, A., and Litman, D. 2006. Cohesion and learning in a tutorial spoken dialog system. In *Proceeding 19th International FLAIRS (Florida Artificial Intelligence Research Society) Conference*.