Cohesion, Semantics and Learning in Reflective Dialog

Arthur WARD, John CONNELLY, Sandra KATZ, Diane LITMAN, Christine WILSON
Learning Research and Development Center, University of Pittsburgh

Abstract. A corpus of reflective tutorial dialogs was tagged for cohesive relationships between student and tutor. We describe our tagging scheme, and show that certain cohesive features of tutoring dialog are correlated with learning in our corpus. In particular, our *semantic* cohesive relationship tags are significant predictors of learning, while our *lexical* tag is not. We find that "abstractive" dialog moves, in which the student or tutor repeats the other's previous utterance but at a greater level of generality, are significant positive predictors of learning. We also find that tutor moves which repeat the student's previous utterance but in a less abstract way predict learning in our corpus. These findings suggest that tracking student dialogue moves can enhance student modeling and guide planning of effective natural-language dialogues.

Keywords. Intelligent Tutoring, Learner Modeling, Discourse Analysis

1. Introduction

Interactive tutorial dialog with a human tutor has been shown to be a very effective form of instruction [1,2]. Many researchers have hypothesized that the very *interactivity* of tutorial dialog contributes to the effectiveness of one-on-one tutoring, and there is substantial empirical support for this hypothesis [3–5]. Although we have some idea what interactivity looks like from the perspective of exchange level analysis [5,6], we know little about what specific discourse mechanisms contribute to interactivity, or how they affect learning. Identifying discourse mechanisms that correlate with learning might help us both to improve our tutoring system dialogs, and also to improve our student models by helping us recognize knowledge gaps and learning during tutoring.

Based on previous work [7,8], we suspect that "cohesion" is an important discourse mechanism in tutoring. Following others [9], we consider cohesion to be the connectedness of a text. Cohesive devices such as pronoun reference and word repetition tell us what elements to include in our mental model and how to connect them. Zwaan and Radvansky [10] consider text to be a "set of processing instructions on how to construct a mental representation of the described situation" (p. 162). The result of following these instructions *may* be a coherent mental model. However, work in textual cohesion has shown that not all readers respond to these processing instructions in the same way. In a series of experiments (e.g.: [11,12]), McNamara and her colleagues have shown that low knowledge readers gain in both comprehension and recall from reading a high, but not

a low cohesion text. On the other hand, high knowledge readers, particularly those with low comprehension skill, show better comprehension gains when given a low cohesion text.

Cohesion in *dialog* can be considered a record of the participants' "collaboration toward coherence" [13]. Dialog participants use various cohesive devices to establish common ground [14], negotiate references [15], and coordinate their mental models [16]. Just as high cohesion text can indicate more detailed instructions for building a mental model (relative to low cohesion text), high cohesion dialogue may signal more detailed collaboration between dialog participants, in building a shared mental model.

Our previous work provided some evidence that cohesion in tutorial dialog interacts with student prior knowledge in a way similar to that of cohesion in expository text. We have found that simple automatically detectable cohesive devices such as word and word-stem repetition between tutor and student predicted learning for low knowledge students, but not for high knowledge students [7]. We later found that also counting cohesive ties between words that were lexically different but semantically related in a hypernym/hyponym hierarchy improved the correlation with far-transfer learning for high pre-testers [8]. In that study far-transfer learning was evaluated using questions that were non-isomorphic to the tutored problems. We discuss a related definition of far transfer for the current corpus in Section 3.

The previous, automatically detectable cohesive ties fit under Halliday and Hasan's [9] category of "lexical cohesion," which includes word, synonym and superordinate-class reiteration. Our implementations, however, were limited to recognizing simple lexical relations between single words. In the current work we use a similar tag which counts simple lexical repetition (our "Exact" tag, Section 2). However we also count more sophisticated *semantic* relations, and recognize ties between multiple-word spans. We find that in our corpus, these manually tagged measures are in fact better predictors of learning than the simple lexical measure. Specifically, we find that tags indicating tutor or student abstraction are significant predictors of learning in our corpus. A tag indicating tutor specialization is also a significant predictor of learning. Similarly to our previous work [7, 8], we find that student response to cohesion varies with both student preparedness and with the type of learning being measured. Our results suggest that abstraction and specialization are important cohesive devices in tutorial dialog. We argue in Section 5 that this has implications for both student modeling and dialog planning.

We describe our tagging scheme in Section 2, our corpus in Section 3 and correlations between tags and learning in Section 4.

2. Tagging for Cohesion

Our previous, automatically computable tags attempted to identify when the tutor and student were referring to each other's contributions. When selecting our expanded tag set, we again focused on ties between tutor and student contributions which might indicate their types of interactivity.

Our final tag set is largely a subset of Halliday and Hasan's [9] taxonomy of cohesive devices. Tags and their brief definitions are listed below. The bracketed numbers after the tag name indicate the tag's total count in tutor (**T:**) and in student (**S:**) turns.

- **Exact** [T:899 S:512] is used when one utterance and the next contain the same word, either in identical or inflected forms.
- **Synonym** [T:67 S:36] is applied when two words with similar meanings are used.
- **Paraphrase** [T:605 S:205] is used for phrase repetitions with word substitution or with different word order.
- **Pronoun** [T:327 S:153] repetition is used when a pronoun such as "it" in one utterance refers to a discourse entity in the previous utterance.
- **Superordinate-class** [T:236 S:50] is used when one speaker uses a more general or abstract referring expression. Examples from our corpus include "force" as a more general reference to "weight," and "velocity" when it follows the more specific "horizontal components of velocity."
- **Class-member** [T:206 S:214] is used when a *more* specific word or phrase such as "horizontal" is used after a less specific one such as "direction."
- **Collocation** [T:121 S:55] is the use of lexical items that regularly co-occur. We follow Halliday and Hasan (who refer to collocation as "the most problematical part of lexical cohesion") and emphasize collocations that stand in some relation of complementarity, such as "left-right" and "up-down." Although collocations are often between individual words, we also recognize the relationship between phrases when they have the complementarity relation.
- **Negation** [T:46 S:25] is used when one speaker directly contradicts the previous speaker.

In choosing this tag set, we selected cohesive devices from Halliday and Hasan (H&H) [9] which could identify common reference between tutor and student, and which seemed to be present in our corpus. We combined some devices which had been distinct in H&H but which were poorly represented in our corpus. For example, our "pronoun" category includes devices such as "nominal reference" ("this") and other types of substitution (e.g. "one"). Our categories of "exact," "synonym," "superordinate-class" and "class-member" correspond to types of lexical reiteration in H&H. Our "paraphrase" tag, however, has no corresponding device in H&H. It is designed to recognize when tutor and student use entire phrases to refer to the other's contribution, and can often contain other types of ties, such as ellipsis, synonym and collocation.

Table 1 contains examples of most of these tags taken from our corpus, edited slightly for clarity. A tutor utterance and the student utterance that followed it are shown at the top of the table. Below them are shown the spans identified in each utterance and the tags given to those spans. In the middle of the table the student utterance and the tutor utterance that followed it are shown. Again, the spans identified in each utterance are shown below them, with their tags. For example, there are two cohesive ties shown[1] between the first two utterances: "superordinate-class" and "exact."

As can be seen from the above definitions and examples, many of our tags required the identification of spans of words that were being paraphrased, elided or otherwise referred to. Identifying spans turned out to be difficult. Spans can be split (as when the referents of "those" are in separated clauses of the preceding utterance), and can even overlap. An example of overlapping spans is in Table 1, where "coming down" is tagged as a collocation, and is also part of the paraphrase "faster coming down." An important and difficult part of applying this set of tags is therefore identifying appropriate spans.

[1] Other ties were removed from the example for clarity.

Using this tag set, two coders tagged a training corpus of 518 student and tutor turns, iteratively refining tag definitions and re-tagging. During this initial tagging phase, the coders relied largely on lexical features. That is, a cohesive tie would be tagged if the words in one span, taken by themselves, could be construed to have a cohesive relationship to the words in the other span. No attempt was made at this stage to judge if the spans referred to the same discourse item or if the relationship made sense in context.

Following this initial coding, we performed a second coding pass in which we re-evaluated spans which had been previously tagged "superordinate-class" "class-member" or "collocation." The remaining tags will be checked later, as time allows. In the new pass we required that the ties previously selected using only lexical features also make sense, and we eliminated the ones that didn't. Ties were eliminated when their spans seemed to have mis-matched topics or referents. Ties were also eliminated if they were not original to the second speaker. For example, if the first speaker had used both "weight" and "force," and the second speaker had also used "force," we would no longer count a superordinate-class tie between "weight" and "force" in the second utterance. We did this in order to distinguish between lexical repetition and knowledge co-construction or elaboration on the part of the second speaker.

Tutor: Good. And the effect on the water is the same. What about the horizontal components of the velocity (of the ball or of the water - either?)		
Student: Velocity is in the same direction as acceleration so the ball is faster coming down		
Tut. Span	**Stu. Span**	**Tag**
horizontal components of the velocity	velocity	superordinate-class
ball	ball	exact
Student: Velocity is in the same direction as acceleration so the ball is faster coming down		
Tutor: It slows down going up and it speeds up coming down - but all the time the horizontal components of the velocity stay unchanged. Horizontal components of velocity are unaffected by gravity. Ok?		
Stu. Span	**Tut. Span**	**Tag**
the ball	it	pronoun
faster coming down	speeds up coming down	paraphrase
velocity	horizontal components of the velocity	class-member
same	unchanged	synonym
coming down	going up	collocation
direction	horizontal	class-member

Table 1.Example Cohesion Tags

One coder re-tagged all instances of these three tags, and a second tagger coded a randomly selected 10% of them for agreement analysis. Kappa on these tags was .57. Agreement was counted when both taggers identified the same textual span *and* applied the same tag to it. Due to the difficulty of reaching agreement on spans, they were counted as the same if they had substantial overlap (no more than one word different at either end, not counting stop-words).

3. Corpus

Our corpus was collected as part of a study of the effectiveness of post-practice, reflective discussions [17]. This study had three conditions. In each condition, students solved a

series of physics problems in the Andes physics tutoring system [18]. After the Andes session, the students were asked "reflection questions" that invited them to elaborate on a specific part of the solution. For example, the following reflection question changes one variable in a previous problem about a jumper hanging motionless from a bungee cord:

> Suppose the maximum tension that the bungee cord could maintain without snapping was 700 N. What would happen to the bungee jumper if he hung stationary on the cord?

After answering reflection questions, students in two conditions were given either canned text feedback or no feedback. In the third condition, however, students used a chat interface to engage a human tutor in dialogue about the reflective questions. Our corpus is taken from these dialogs. The experimental procedure used and the resulting dialogue corpus can be summarized as follows.

Sixteen students answered a questionnaire, then were given a math test and a physics pre-test. After the pre-test, students reviewed a workbook chapter on kinematics developed for the experiment, and received training in the use of the Andes tutoring system. They then solved 12 physics problems in Andes. Following each problem, they were given between three and eight reflection questions. They would type their answer to each question, or state that they could not answer the question, and then engage a human tutor in reflective dialog about the answer, using a chat interface. Fifteen students participated in 60 reflective dialogs each, while the sixteenth participated in 53 dialogs. This created a corpus of 953 reflective dialogs, containing 2,218 student turns and 2,136 tutor turns. A post-test was given after the reflective dialogs. The pre- and post- tests covered the same topics and contained 36 questions: 9 quantitative mechanics questions similar to those that the students worked on in Andes, and 27 qualitative questions that tested their ability to apply mechanics concepts and principles to new problems that were dissimilar to those tutored under Andes. The pre- and post- tests were administered in a counterbalanced order. Overall, the researchers found that students in both dialog treatment conditions learned more than students in the no-dialogue control condition, as measured by pre-test to post-test gain score, but the canned feedback and human feedback conditions did not differ significantly.

In Section 4 we show that some of our tags predict learning measured by the qualitative but not the quantitative questions. Because the qualitative questions were less similar to the training problems, we argue that they measure farther transfer of learning than do the quantitative questions. While this transfer is probably not all that "far" in the taxonomy described by Barnett and Ceci [19], it seems probable that success on these problems required the construction of a more abstract representation of the material than was needed for the quantitative problems.

As noted in Section 1, students of different knowledge levels may respond to cohesive cues differently, so we ran statistics on our "high" and "low" pre-testers separately, as divided by their median pre-test score. When using the quantitative questions, this division results in eight low and eight high pre-testers. Using either the qualitative questions or the set of all questions results in seven low and nine high pre-testers.

4. Results

We used linear models to look for relationships between each of our cohesion tags and learning, as measured by pre- and post-test scores (total scores as well as quantitative and

	All Questions											
	All Students				Low Pre-test				High Pre-test			
Tag Name:	**Pre**	**Mth**	**Tag**	**Mod**	**Pre**	**Mth**	**Tag**	**Mod**	**Pre**	**Mth**	**Tag**	**Mod**
S:Super-Ord	.061	.070	*.054*	**.005**	.259	.562	.152	.109	.466	.146	.420	.272
	Qualitative Questions											
	All Students				**Low Pre-test**				**High Pre-test**			
Tag Name:	**Pre**	**Mth**	**Tag**	**Mod**	**Pre**	**Mth**	**Tag**	**Mod**	**Pre**	**Mth**	**Tag**	**Mod**
S: Super-Ord	.274	.006	**.005**	**.002**	.987	.072	*.066*	.111	.966	.159	.229	.295
T: Class Mem	.205	.006	**.015**	**.005**	.983	.620	.794	.629	.537	.097	**.032**	*.059*
T: Super-Ord	.296	.049	.262	**.045**	.029	.011	**.017**	**.032**	.488	.185	.748	.550
	Quantitative Questions											
	All Students				**Low Pre-test**				**High Pre-test**			
Tag Name:	**Pre**	**Mth**	**Tag**	**Mod**	**Pre**	**Mth**	**Tag**	**Mod**	**Pre**	**Mth**	**Tag**	**Mod**
T: Class Mem	.035	.671	.873	*.093*	.058	.527	**.050**	.153	.283	.533	.614	.155

Table 2. P-values for individual predictor variables (Pre-test, Math and tag count) and whole linear model

qualitative subscores). We regressed post-test score (or relevant sub-score) on pre-test score (or relevant sub-score), math test score, and normalized tag count. We included pre-test scores as predictors because they are significantly correlated with post-test scores in our corpus[2], and math test scores because they were shown to be a significant predictor of learning in previous work with the Andes tutor [17]. We use normalized tag counts to control for the effect of longer tutorial dialog on learning. For example, the tag "Student Superordinate-class" (S:Super-Ord) is the total count of "superordinate-class" tags for a student, divided by that student's total number of turns.

Results are shown in Table 2. The table is divided horizontally by which measure of learning gain is used in the regression model. The models at the top use all 36 questions, the models in the middle use the 27 qualitative questions, and the models at the bottom use the 9 quantitative questions. For each measure of learning, the tags that were significant predictors (or strong trends) in at least one model are shown in the left most column. P-values for the linear models that use that tag are shown on the same row. For each linear model, the table shows the individual p-value for each predictor variable: "Pre" = relevant pre-test score, "Mth" = math score, "Tag" = the cohesion tag used in the model, "Mod" = the p-value for the whole model. Significant tag and model p-values are shown in bold, and trends are italicized. For each tag, we ran regressions using the entire student sample as well as on the subgroups of high and low pre-testers.

The only tag that approached significance when predicting total learning gains (under "All Questions") was Student Superordinate-class. This tag had a p-value of .054 for All Students, but was not even a trend for the low or high pre-testers taken separately. Similarly only one tag, "Tutor Class-member," was significant in predicting quantitative learning gain, but it occurred in a non-significant model (p = .153).

For qualitative learning, however, three different tags proved to be significant predictors. As shown in the center rows of Table 2, the Student Superordinate-class tag was

[2]The pre- to post-test correlation for all questions is .67 (p =.0045), for the quantitative (near) questions: .63 (p = .009), and for the qualitative (far) questions: .51 (p = .043)

significant for all students and achieved a trend in the sub-group of low pre-testers, although in a model which was not significant overall. Similarly, the Tutor superordinate-class tag was a significant predictor for low pre-testers. Tutor class-member was a significant predictor for all students. It was also significant for high pre-testers, but in a model that fell slightly short of being significant ($p = .059$).

5. Discussion and Future Work

The goal of this study was to expand upon our previous work which had suggested the importance of cohesion in tutorial dialog. Those studies had found that automatic measures of interactivity, which measured when tutor and student used similar words or words that were related in WordNet's is-a hierarchy, are correlated with student learning. The shallow measures we used then, however, could only provide limited insight about exactly what mechanisms account for the value of interactivity. Our new tags capture semantic relationships between phrases which were invisible when counting only shallow lexical relationships between individual words.

The current study has broadened and reinforced our earlier work by showing that different measures of tutor-to-student cohesion also positively predict learning in a new corpus of tutorial dialogs. In addition, it provides insight into exactly what mechanisms are involved. By tagging manually rather than automatically, we were able to recognize a broad set of semantic relationships between tutor and student utterances. Some of these cohesive relationships did not seem to be related to learning in our corpus. The ones that *were* related involve changes in the level of concreteness being used. In particular, tutor or student abstraction seems to be a particularly valuable cohesive device. We suggest that this type of cohesive tie tends to happen when the student is building a more abstracted mental model. This model is then more useful in answering far-transfer questions, as shown by our results in predicting "qualitative" learning.

These results can also be seen as adding detail to previous work by Katz et al. [17], who found that the number of reflective dialogs that abstracted from the previous problems were correlated with learning.

It is also interesting to note that the "Exact" tag was not significant in any model, even though this tag is similar to the lexical reiteration measure which correlated with learning in other corpora [7]. In the current corpus of reflective dialogs, only tags that were sensitive to the semantic content of the utterances were significant predictors of learning.

Our results using the Tutor Class-member and Tutor Superordinate-class tags suggest that we might be able to improve learning by manipulating tutor and student utterances. In future work we hope to test this experimentally, by making tutor utterances more concrete or more abstract at appropriate places in the tutorial dialog, and by prompting students to do the same. Further work will be required to tell where those places are.

Our results using the Student Superordinate-class tag suggest that we might be able to improve student modeling in our tutor by measuring student abstraction during tutoring. We hope to explore this possibility by using cohesion within certain dialog segments to predict correctness on particular post-test questions. If we can in fact build better student models through more sophisticated automated cohesion analysis, this model could then be used to guide more effective tutorial dialog planning.

6. Acknowledgments

Funding for this work was provided by the Office of Naval Research (ONR), Grant Number N000140710039, and by the Learning Research and Development Center. These organizations do not necessarily endorse the data or views presented here.

References

[1] B. S. Bloom. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13:4 – 16, 1984.

[2] A. Corbett. Cognitive computer tutors: Solving the two sigma problem. *Proceedings of the Eighth International Iser Modelling Conference*, pages 137 – 147, 2001.

[3] Michelene T.H. Chi, Stephanie A. Siler, Heisawn Jeong, Takashi Yamauchi, and Robert G. Hausmann. Learning from human tutoring. *Cognitive Science*, 25:471 – 533, 2001.

[4] Michelene Chi, Marguerite Roy, and Robert Hausmann. Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive Science: A Multidisciplinary Journal*, 32:301 – 341, 2008.

[5] Arthur C. Graesser, Natalie Person, and Joseph P. Magliano. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, 9:495 – 522, 1995.

[6] Kate Forbes-Riley, Diane Litman, Alison Huettner, and Arthur Ward. Dialogue-learning correlations in spoken dialogue tutoring. In *Proceedings 12th International Conference on Artificial Intelligence Education (AIED),Amsterdam, Netherlands*, July 2005.

[7] Arthur Ward and Diane Litman. Cohesion and learning in a tutorial spoken dialog system. In *Proceedings of the 19th International FLAIRS Conference (FLAIRS-19)*, pages 533–538, May 2006.

[8] Arthur Ward and Diane Litman. Semantic cohesion and learning. In *Proceedings 9th International Conference on Intelligent Tutoring Systems (ITS)*, pages 459–469, Ann Arbor, June 2008.

[9] M. A. K. Halliday and Ruqaiya Hasan. *Cohesion in English*. English Language Series. Pearson Education Limited, 1976.

[10] Rolf A. Zwaan and Gabriel A. Radvansky. Situation models in language comprehension and memory. *Psychological Bulletin*, 123:162 – 185, 1998.

[11] Danielle S. McNamara and Walter Kintsch. Learning from text: Effects of prior knowledge and text coherence. *Discourse Processes*, 22:247–287, 1996.

[12] Danielle McNamara. Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology*, 55:51–62, 2001.

[13] *Coherence in Spontaneous Text*. Typological Studies in Language, 31. John Benjamins, Philadelphia, 1995.

[14] Hebert H. Clark and S. A. Brennan. Grounding in communication. In L. B. Resnick, J. M. Levine, and S. D. Teasley, editors, *Perspectives on socially shared cognition*. 1991.

[15] Peter A. Heeman and Graeme Hirst. Collaborating on referring expressions. *Computational Linguistics*, 21:351–382, 1995.

[16] Martin J Pickering and Simon Garrod. Toward a mechanistic psychology of dialogue. In *Behavioral and Brain Sciences*, volume 27, 2004.

[17] Sandra Katz, David Allbritton, and John Connelly. Going beyond the problem given: How human tutors use post-solution discussions to support transfer. *International Journal of Artificial Intelligence in Education*, 13:79 – 116, 2003.

[18] K. VanLehn, C. Lynch, K. Schulze, J.A. Shapiro, R. Shelby, L. Taylor, D. Treacy, A. Weinstein, and M. Wintersgill. The andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence and Education*, 15:147 – 204, 2005.

[19] Susan Barnett and Stephen Ceci. When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin*, 128:612 – 637, 2002.