

# Cohesion and Learning in a Tutorial Spoken Dialog System\*

**Arthur Ward**

Learning Research and Development Center  
University of Pittsburgh  
Pittsburgh, Pa., 15260, USA  
artward@cs.pitt.edu

**Diane Litman**

Learning Research and Development Center  
University of Pittsburgh  
Pittsburgh, Pa., 15260, USA  
litman@cs.pitt.edu

## Abstract

Two measures of lexical cohesion were developed and applied to a corpus of human-computer tutoring dialogs. For both measures, the amount of cohesion in the tutoring dialog was found to be significantly correlated to learning for students with below-mean pretest scores, but not for those with above-mean pre-test scores, even though both groups had similar amounts of cohesion. We also find that only cohesion between tutor and student is significant: the cohesiveness of tutor, or of student, utterances is not. These results are discussed in light of previous work in textual cohesion and recall.

## Introduction

One-on-one tutoring with a human tutor often yields significantly higher learning gains than classroom instruction (Bloom 1984). Because human tutors are expensive, however, research has focused on replicating their advantages in computerized Intelligent Tutoring Systems (ITSs). One important line of ITS research hypothesizes that there is something about the natural language interaction used by human tutors that makes learning easier. If we could identify what features of natural tutorial dialog enhance learning, we could use this knowledge to build tutors that encourage these features. For example, Graesser, Person, & Magliano (1995) argue that learning in tutoring is linked to certain dialog acts such as question answering and explanatory reasoning. Similarly, Forbes-Riley *et al.* (2005) found correlations between deep student answers and learning in a hand-tagged corpus of tutorial dialogs with a computer tutor. This is a promising approach, but it is currently problematic to deploy in an ITS because of the difficulty of automatically tagging the dialog input. Other approaches look for dialog features which could more easily be detected by a computer tutor during

the tutoring session. For example, Litman *et al.* (2004) examine surface-level dialog features such as average turn length and ratio of student to tutor words. While such features had correlated with learning in prior studies of typed interactions primarily with human tutors, they did not correlate with learning in Litman *et al.*'s corpora of spoken dialogs with both human and computer tutors.

In this paper we examine a different feature of tutorial dialog, cohesion, which we define as lexical co-occurrence between turns. Cohesion has been studied both within the computational linguistics community and by researchers in discourse comprehension. In computational linguistics, the emphasis is often on using cohesion as a guide for text segmentation or summarization (Barzilay & Elhadad 1997; Hearst 1994). This work often focuses on text rather than dialog, and has not, so far, related cohesion to learning. The work in discourse comprehension, on the other hand, has related textual coherence to learning, but has used only carefully controlled experimental texts (McNamara & Kintsch 1996). Our work is novel in that it takes automatically computable measures of lexical cohesion from computational linguistics, applies them to tutorial dialog, and relates the result to student learning.

Similarly to important results in text comprehension (McNamara & Kintsch 1996), we find that measures of lexical cohesion are strongly associated with learning among students with below mean pre-test scores. In addition, we find that only cohesion between tutor and student dialog contributions is correlated with learning. Cohesion between tutor or between student turns is not. Finally, we show that the mean amounts of tutor-to-student cohesion were not significantly different between the groups of high and low pre-testers. The low pre-testers learned in proportion to the cohesion in their dialogs, while the high pre-testers did not.

## Motivation

Our interest in measuring cohesion in tutorial dialog is motivated in part by the success researchers in discourse com-

---

\*This research is supported by the NSF (0325054) and ONR (N00014-04-1-0108). We thank the ITSPPOKE group and Pamela Jordan for their support and useful comments.  
Copyright © 2006, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

Stem Group	Tokens
packag	package, packages
packet	packet, packets
speed	speed, speeding, speeds
veloc	veloc, velocities, velocity, velocitys
horizont	horizontal, horizontally
displac	displace, displaced, displacement, displacements, displacing
find	find, finding
so	so
thu	thus

Table 1: Examples of how tokens are grouped by stems

prehension have had studying the relationship of coherence to learning in expository text. In particular, McNamara and Kintsch (1996) have found an interaction between the coherence of an expository text and the aptitude of the reader, as measured by pre-test score. We will describe their findings in some detail, because our results suggest a similar interaction in our tutorial dialogs.

McNamara and Kintsch instructed students to read both high and low coherence texts. Their experiments used an expository text in its original form as the “low coherence” text, and a modified version as the “high coherence” text. The original text tended to use a variety of terms for the same referent, and to assume some background knowledge. The “high coherence” version was this same text, but altered to use consistent referring expressions, to identify anaphora, and supply some background information missing in the original. In their work “coherence” refers to the extent to which relations in the text are made explicit, rather than having to be inferred by the reader. Our measure of “cohesion,” as described below, is a measure of the extent to which the same words reoccur in succeeding turns. These are related, but not exactly the same. In this paper we sometimes use the term “cohesion” to include both meanings.

McNamara and Kintsch’s measures of learning were designed in light of van Dijk and Kintsch’s (1983) three layer model of memory for discourse. This model suggests that three types of memory trace are created while reading expository text: the “surface code,” the “text base,” and the “situation model.” The surface code is a rapidly decaying memory for the exact words in the sentence being read. The text base is a memory of the explicitly stated propositional content of the sentence. The situation model is a longer lived memory for the deeper content, built by inferential processes which use both world knowledge and information in the text. McNamara and Kintsch designed learning measures to test recall at the text base and situation model levels separately. They found that increasing the coherence of the text being

read helped low pre-test readers form both text base and situation model level memories. For the high pre-testers, however, a more coherent text had a small benefit for text-base memory, but actually reduced recall at the situation model level.

In our work we use a corpus of previously collected tutoring dialogs, in which cohesion has been created by tutor and student contributions. Because we cannot manipulate cohesion in this corpus, we look instead to the computational linguistics literature for ways to measure cohesion in existing text.

Hearst (1994) uses cohesion as a guide for topic segmentation in text. She locates topic boundaries by comparing the word-count similarity of adjacent spans of text. Olney and Cai (2005) extend topic segmentation from text into dialog. They compare the use of several measures of lexical cohesion, including Hearst’s, in finding boundaries between topics in tutorial dialog.

Morris and Hirst (1991) measure cohesion in text using lexical chains of words with related thesaurus entries. This measure of cohesion is used to segment the text by intentional structure. Barzilay and Elhadad (1997) automate this lexical chain method using WordNet senses rather than thesaurus synonyms, and apply it to the problem of text summarization.

In this work we develop measures of cohesion that are similar to those used by Hearst (1994) in that they compare spans of text based on word-count similarity. However, we apply these measures to a corpus of tutorial dialogs, rather than text, and examine their correlation with learning.

## The Corpora of Tutorial Dialogs

To train and test our model, we used two different corpora of tutoring transcripts collected by the ITSPoKE intelligent tutoring system project (Litman & Silliman 2004). ITSPoKE is a speech enhanced version of the Why2-Atlas qualitative physics tutoring system (VanLehn *et al.* 2002). Our training corpus consisted of transcripts collected in the Fall of 2003. The testing corpus was collected in the Spring of 2005 using a slightly updated computer tutor, which had been given a larger language model to improve speech recognition.

Both experiments had identical tutoring methodologies, and both taught conceptual physics to students who had never taken college physics. The students first read instructional materials about physics, then were given a pre-test to gauge their physics knowledge. At the start of interactive tutoring the student was given a qualitative physics problem and asked to write an essay about it. The computer tutor would interpret the essay, identify a key point that was missing or wrong, and engage the student in a tutorial dialog to teach that point. The student then would be asked to revise the essay, and the tutorial cycle would repeat until all

Token w/stop(14)		Token, no stop (9)	Stem, no stop(11)
packet, horizontal, the, it, is, of, only, force, acting, on, there, will, still, after		packet, horizontal, only, force, acting, there, will, still, after	packet, horizont, onli, forc, act, acceler, vertic, there, will, still, after
Student Essay	No. The airplane and the packet have the same horizontal velocity. When the packet is dropped, the only force acting on it is g, and the net force is zero. The packet accelerates vertically down, but does not accelerate horizontally. The packet keeps moving at the same velocity while it is falling as it had when it was on the airplane. There will be displacement because the packet still moves horizontally after it is dropped. The packet will keep moving past the center of the swimming pool because of its horizontal velocity.		
Computer Tutor	Uh huh. There is more still that your essay should cover. Maybe this will help you remember some of the details need in the explanation. After the packet is released, the only force acting on it is gravitational force, which acts in the vertical direction. What is the magnitude of the acceleration of the packet in the horizontal direction?		

Table 2: Two consecutive turns, counting cohesive ties at the token and stem levels

points had been covered. Each student covered five problems this way, then was given a post-test. The post test consisted of problems designed to cover the same concepts as the pre-test, but without sharing any text with either the pre-test or the training problems. In terms of the McNamara and Kintsch study described above, this post test emphasized the situation model level, rather than the text-base or propositional content.

There were twenty students in the 2003 study, who did five problems each. Five dialogs were excluded from our current analysis because the computer tutor had accepted the student's initial essay without engaging in any dialog, leaving ninety-five dialogs in our training corpus. There were 34 students in the 2005 study, doing five problems each. Seven of those dialogs were removed for the same reason, which resulted in a testing corpus of 163 dialogs.

Both corpora exhibited strict turn taking between tutor and student. Tutor utterances alternated with student contributions, which could be an utterance, an essay submission, or an empty turn.<sup>1</sup>

## Measuring Cohesion in our Corpora

In this section we describe the measures of cohesion developed for use in our tutoring dialogs. Our final suite of measures is the product of four sets of decisions. First, we had to decide how to identify cohesion in the transcripts. Second, we had to determine in which conversational partner's contributions to study cohesion. Third, we had to pick in which group of students to look. Fourth, we chose among a number of additional processing steps to determine if any would improve our measurements. We will describe each of these decisions in turn.

<sup>1</sup>Empty student turns sometimes occurred when the student exceeded the computer tutor's pre-set time out period before responding. The tutor would then re-prompt the student, leaving two consecutive tutor turns in the transcript. In these cases, the tutor turns were merged to maintain the alternation of turns.

## Identifying Cohesion

We identify cohesion in our corpus at two different levels of abstraction. These levels can be understood in terms of a framework presented by Halliday and Hasan in *Cohesion in English* (Halliday & Hasan 1976). They divide textual cohesion into two general categories, grammatical and lexical. Grammatical cohesion includes various kinds of reference, while lexical cohesion includes various kinds of reiteration. Reiteration is further divided into several repetition types, ranging from exact word repetition, through repetition of synonyms, near synonyms, superordinate class terms, and general referring nouns. Halliday and Hasan suggest that the cohesiveness of a text can be measured by counting the number of "cohesive ties" that it contains, where a "cohesive tie" consists of two words joined by one of these devices, such as repetition. We examine counting two kinds of cohesive ties: "token" and "stem group," which are modeled after their first two types of reiteration.

At the "token" level, a cohesive link is counted if exactly the same word form (after stripping punctuation, and ignoring case) appears both in one turn and the next. Results were collected at this level both with and without stop-words being counted. This corresponds to the first of Halliday and Hasan's reiteration types.

At the "stem" level, a link is counted if two words in consecutive turns are given the same stem by a standard Porter stemmer (Porter 1997). Table 1 gives examples of how tokens are grouped by stems. Tokens to which the stemmer assigns a common stem appear together in the second column, with their stem in the first column.

Table 2 shows how these definitions of a cohesive tie can affect the amount of cohesion counted. Two consecutive turns are shown at the bottom of the table. The three columns at the top of the table show the matches counted at each level, and their total count. For example, the "Token w/stop" level counts 14 exact word repetitions. The "token, no stop" level matches tokens after removing stop words. This level counts 9 cohesive ties between these turns. The "stem, no stop" level matches stems after removing stop

words. This level counts 11 cohesive ties: the same 9 as at the “token, no stop” level, plus the additional stems “acceler” and “vertic.” This allows the stem level to match the tokens “accelerates” to “acceleration,” and “vertically” to “vertically.” These matches were not found at the token levels.

### Looking for Effects of Interactivity

Our second decision concerned in which dialog participant’s contributions to look for cohesion. For ITS development, we would like to find tutor behaviors associated with learning, however many recent results suggest that the student’s contribution is at least as important. For example, Chi *et al.* (2001) compared student-centered, tutor-centered and interactive hypotheses of learning and found that students learned just as effectively when tutor feedback was suppressed. Their evidence suggested that deep learning was facilitated by the student’s self-construction of knowledge, rather than simply by tutor actions alone. Also Forbes-Riley *et al.* (2005) find that student utterances which display reasoning, and reasoning questions asked by the computer tutor, both correlate with learning. These considerations lead us to look separately at cohesion between tutor utterances, between student utterances, and between tutor and student utterances. We do this by creating two additional corpora, one with only tutor turns, and one with only student turns, and running on them the same tests as on the full, interactive corpus. These comparisons should offer evidence about the importance of interactivity in tutoring.

### Looking for an Aptitude Treatment Interaction

As mentioned above, McNamara and Kintsch (1996) demonstrated that textual coherence affected recall differently for high and low pre-test readers, as categorized by mean pre-test score. We decided to split our data in an identical way to see if it revealed a similar aptitude interaction. This divided our 2003 data into 13 “low” pre-testers and 7 “high” pre-testers. Our 2005 data was divided into 18 “low” pre-testers, and 16 “high” pre-testers.

### Choosing other Processing Steps

In addition to the choices described above, we experimented with several other kinds of processing in our training corpus. Briefly, we investigated measuring cohesion between spans of various sizes, for example comparing spans of two turns each. We also experimented with removing stop words (high frequency, low meaning words such as “it” and “is”), with pre-processing our transcripts using TF-IDF normalization, and with counting only turns that had been given a “substantive” tag.<sup>2</sup> Finally, we tried both raw cohesion counts

<sup>2</sup>See (Forbes-Riley *et al.* 2005) for a description of the tag set used. “Substantive” turns included only turns with physics related

Students	Tests			
	Train: 2003 Data		Test: 2005 Data	
	R	P-Value	R	P-Value
Grouped by Token (with stop words)				
All Students	0.380	0.098	0.207	0.239
Low Pretest	0.614	<b>0.026</b>	0.448	0.062
High Pretest	0.509	0.244	0.014	0.958
Grouped by Token (stop words removed)				
All Students	0.431	0.058	0.269	0.124
Low Pretest	0.676	<b>0.011</b>	0.481	<b>0.043</b>
High Pretest	0.606	0.149	0.132	0.627
Grouped by Stem (stop words removed)				
All Students	0.423	0.063	0.261	0.135
Low Pretest	0.685	<b>0.010</b>	0.474	<b>0.047</b>
High Pretest	0.633	0.127	0.121	0.655

Table 3: Results for all turns, comparing one-turn spans, using turn-normalization

and “turn normalized” counts, in which the total number of cohesive ties was divided by the number of turns in the dialog. We experimented extensively in the training corpus to determine what combination of these options would be used. When we had finalized our set of options, we used them without alteration on our testing corpus. All the results we report use this final set of options: they compare one-turn spans, use turn normalization, and use neither TF-IDF normalization nor substantive turn selection.

## Results

Our major test is a partial correlation of post-test score with cohesion, controlling for the effect of pre-test score. We use this test because post-test and pre-test scores are correlated in our training data ( $R = .462$ ,  $P\text{-Value} = .04$ ), making it necessary to control for pre-test score by regressing it out of our correlations.

Results for interactive data are shown in table 3. The table is divided vertically, with columns two and three showing results for our training data, and columns four and five showing results on our test data. The table is grouped horizontally by the level (token w/stops, token w/no stops, stem w/no stops) at which cohesion is being measured. Within each of those groups, results are shown for the three divisions of our data designed to test for an aptitude-treatment interaction. The top row in each group shows results for all students. Below are results for students with below mean pretest scores, then results for students with above mean pretest scores. Statistically significant results ( $P \leq .05$ ) are shown in bold.

content

## Learning and Cohesion

Table 3 shows that we had significant correlations with learning using both the “token” and “stem” types of measurement. However, it also shows that while removing stops offered some improvement, there was little difference between the “token” and “stem” levels. There are several possible explanations for this. First, it may be that there is simply not much variation in word choice in our dialogs. In this case, being able to see the similarity between “velocity,” “velocities” and “velocitys,” for example, wouldn’t help much because students almost always say “velocity.” Another possibility is that a more sophisticated “sense” level matching scheme would help. This is a matter for future work to resolve.

## Aptitude and Cohesion

As shown in Table 3, the same pattern is apparent for all our measures of cohesion, and in both our training and test sets. Cohesion is found to be strongly correlated with learning *only* for students with below mean pre-test scores. These correlations hold in the testing data, although the “token with stop words” level is reduced to a trend. In both data sets, cohesion is a highly *non* significant predictor of learning among high-pretest students, and is never more than a trend among students taken as a whole.

This result is not due simply to the high and low pre-test groups having different amounts of lexical cohesion. Table 4 shows the difference in mean amounts of cohesion between the high pre-testers and low pre-testers, using each of our three ways of counting cohesive ties. This table shows that, using the turn-normalized measures we report in this paper, there is no significant difference in the amount of cohesion found in the high pre-test vs. low pre-test groups. However, cohesion is correlated with learning for the low pre-testers, but not for the high pre-testers.

## Interactivity and Cohesion

To investigate the role of interactivity in our corpus, we re-ran these tests on corpora made up of tutor-only and student-only turns. Table 5 shows results for the same tests run on the tutor-only and student-only corpora, organized identically to table 3. There are no significant correlations, and no trends. These results suggest that the cohesion between adjacent tutor utterances has no correlation to learning. We also have no evidence that the cohesion between adjacent student utterances is correlated with learning. It should be noted, however, that students can be much less verbose with computer than with human tutors (Litman *et al.* 2004), and this may make it more difficult to find a correlation. In our computer tutor corpora only the cohesiveness between student and tutor is significantly correlated with learning.

	Mean Cohesion		P-Val
	High Pre	Low Pre	
Token (with Stop words)	9.978	9.449	0.581
Token (Stops removed)	5.375	5.209	0.768
Stem (Stops removed)	5.713	5.611	0.867

Table 4: Mean turn-normalized cohesion measurements for high and low pre-testers in 2003 data. The P-Values indicate no significant difference between the group means.

Students	Tests			
	Tutor Only		Student Only	
	R	P-Value	R	P-Value
Grouped by Token (with Stop words)				
All Students	0.060	0.800	0.242	0.304
Low Pretest	0.121	0.695	0.325	0.279
High Pretest	0.531	0.220	0.451	0.310
Grouped by Token (Stops removed)				
All Students	0.004	0.987	0.111	0.640
Low Pretest	0.114	0.710	0.010	0.974
High Pretest	0.351	0.440	0.493	0.261
Grouped by Stem (Stops removed)				
All Students	0.010	0.967	0.113	0.637
Low Pretest	0.107	0.727	0.009	0.976
High Pretest	0.465	0.293	0.501	0.252

Table 5: Tutor-Only and Student-Only turns, 2003 data

## Discussion

Learning gains were significant for both the low and high pre-testers. In the 2003 data, a two-way ANOVA with condition (hi vs low pre-tester) by repeated test (pre vs post-test) design, there was a robust main effect for test phase,  $F(1,18)=22.061$ ,  $p=0.000$ ,  $Mse=0.015$ , indicating that students in both conditions learned a significant amount during tutoring. Results were similar for the 2005 data,  $F(1,32)=106.007$ ,  $p=0.000$ ,  $Mse = .004$ . So, both low and high pre-testers successfully learned from these dialogs. Our measures of cohesion, however, seem to reflect only what the low pre-testers are doing to achieve their gains. They capture none of what the high pre-testers are doing to learn.

Given the strong resemblance between our results in dialog and those of McNamara and Kintsch in text, described above, it is interesting to speculate on their relationship. In the McNamara and Kintsch experiment, texts were manipulated to contain two distinct levels of coherence, with the semantic relationships more explicit in the high coherence version. In our data we do not have these distinct levels, but instead measure cohesion as the extent to which the tutor and student use the same words in adjacent turns.

In both these studies, the low pre-testers learned more with higher cohesion. In text, this may be because the high coherence version supplied them semantic relationships they were unable to infer from the low coherence version. In di-

alog, higher cohesion means the student and tutor are using more of the same words, which may be evidence that the student has made whatever inferences were necessary to learn their use.

The comparison is more difficult for the high pre-test students. McNamara and Kintsch's high pre-testers showed a negative correlation between textual coherence and learning. Ours showed no correlation at all. McNamara and Kintsch's negative correlation may be because a higher coherence text supplied fewer "inference triggers." The high pre-testers therefore made fewer inferences, and learned less. This mechanism doesn't seem to hold for the high pre-testers in our data.

It may be that cohesion is also associated with learning for our high pre-testers, but we are measuring the wrong type of cohesion. Possibly cohesion measured at the sense level would be significant for the high pre-testers. Or, it might be necessary to measure cohesive ties between groups of concepts (Ward & Litman 2005), rather than between individual terms. These are aims for our future work.

Another matter for future work will be to investigate various hypothesis about the cause of these observed correlations. For example, perhaps the inferences the low pretesters may be making from the high cohesion dialogs are about the meaning and use of domain relevant terms. We might be able to find evidence for this hypothesis in a correlation between the use of such terms and learning among our low pretesters.

## Conclusions and Future Work

We have shown that simple measures of lexical cohesion are correlated with learning in our tutorial dialogs, but only for students with below mean pretest scores. These results replicated on a second corpus of transcripts. We also show that only the cohesion between tutor and student turns is correlated with learning. Neither cohesion between tutor utterances, nor between student productions is correlated with learning in our corpora.

Future work will include assessing more sophisticated measures of cohesion, which we hope will reveal correlations for our high pre-test students, too. In particular, we hope to capture "sense" level cohesion using WordNet senses, similarly to Barzilay and Elhadad (1997). Another useful approach to capturing cohesion at this level may be Latent Semantic Analysis (Olney & Cai 2005). After these alternative measures of cohesion have been evaluated, we hope to manipulate tutor utterances to determine if varying dialog cohesion experimentally affects student learning.

## References

Barzilay, R., and Elhadad, M. 1997. Using lexical chains for text summarization. In *Proceedings of the Intelligent Scalable*

*Text Summarization Workshop (ISTS'97), ACL, Madrid, Spain.*, 10–17.

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

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

Forbes-Riley, K.; Litman, D.; Huettner, A.; and Ward, A. 2005. Dialogue-learning correlations in spoken dialogue tutoring. In *Proceedings 12th International Conference on Artificial Intelligence Education (AIED)*, Amsterdam, Netherlands.

Graesser, A.; Person, N.; and Magliano, J. 1995. Collaborative dialog patterns in naturalistic one-on-one tutoring. *Applied Cognitive Psychology* 9:359–387.

Halliday, M. A. K., and Hasan, R. 1976. *Cohesion in English*. English Language Series. Pearson Education Limited.

Hearst, M. 1994. Multi-paragraph segmentation of expository text. In *32nd. Annual Meeting of the Association for Computational Linguistics*, 9–16.

Litman, D., and Silliman, S. 2004. ITSPOKE: An intelligent tutoring spoken dialogue system. In *Companion Proc. of the Human Language Technology Conf: 4th Meeting of the North American Chap. of the Assoc. for Computational Linguistics*.

Litman, D. J.; Rosé, C. P.; Forbes-Riley, K.; VanLehn, K.; Bhembé, D.; and Silliman, S. 2004. Spoken versus typed human and computer dialogue tutoring. In *Proceedings of the 7th International Conference on Intelligent Tutoring Systems(ITS)*. Macceio, Brazil.

McNamara, D. S., and Kintsch, W. 1996. Learning from text: Effects of prior knowledge and text coherence. *Discourse Processes* 22:247–287.

Morris, J., and Hirst, G. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics* 17(1):21–48.

Olney, A., and Cai, Z. 2005. An orthonormal basis for topic segmentation in tutorial dialog. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, 971–978. Vancouver.

Porter, M. F. 1997. An algorithm for suffix stripping. *Readings in information retrieval* 313–316.

vanDijk, T. A., and Kintsch, W. 1983. *Strategies of Discourse Comprehension*. New York, Academic Press.

VanLehn, K.; Jordan, P. W.; Rose, C. P.; Bhembé, D.; Boettner, M.; Gaydos, A.; Makatchev, M.; Pappuswamy, U.; Ringenberg, M.; Roque, A.; Siler, S.; and Srivastava, R. 2002. The architecture of why2-atlas: A coach for qualitative physics essay writing. In *Proc. 6th Int. Conf. on Intelligent Tutoring Systems*, volume 2363 of LNCS, 158–167. Springer.

Ward, A., and Litman, D. 2005. Predicting learning in tutoring with the landscape model of memory. In *Proceedings of the 2nd Workshop on Building Educational Applications using NLP*, 21–24.