Semantic Cohesion and Learning

Arthur Ward and Diane Litman

University of Pittsburgh, Pittsburgh, Pa., 15260, USA

Abstract. A previously reported measure of dialog cohesion was extended to measure cohesion by counting semantic similarity (the repetition of meaning) as well as lexical reiteration (the repetition of words) cohesive ties. Adding semantic similarity ties improved the algorithm's correlation with learning among high pre-testers in one of our corpora of tutoring dialogs, where the lexical reiteration measure alone had correlated only for low pre-testers. Counting cohesive ties which have increasing semantic distance increases the measure's correlation with learning in that corpus. We also find that both directions of tie, student-to-tutor and tutor-to-student, are equally important in producing these correlations. Finally, we present evidence suggesting that the correlations we find may be with deeper "far transfer" learning.

1 Introduction

Researchers in Intelligent Tutoring Systems often study tutorial dialog for clues to the effectiveness [1] of human tutors. This research has focused on many aspects of the tutoring interaction, from tutorial support [2] to the occurrence of various dialog acts [3, 4]. Because deep dialog act features such as question answering are difficult for tutoring systems to identify automatically, our own research has also focused on shallow dialog features (such as word count or turn length) which are more automatically computable [5]. Unfortunately these shallow features tend to have poorer correlations with learning than the deeper features [6]. We look for correlations between dialog act features and learning both because we want to be able to detect learning during tutoring, and because we want to be able to design effective tutorial dialog interventions later on.

Much work on tutorial interventions, however, suggests that their effectiveness is often dependent on the preparedness level of the student. For example, VanLehn et al. [7] present evidence that tutoring is only better than reading when the reading material is too difficult for that particular student. Kalyuga et al. describe a number of instructional techniques which work for low-knowledge students, but not for high-knowledge students [8]. Similarly, Conati and VanLehn [9] find that providing rich scaffolding for self-explanation helps low-knowledge students, but can actually impede learning for high-knowledge students. McNamara and Kintsch [10] find that increasing the coherence of text can aid recall for low, but not for high-knowledge readers.

In a previous paper [11], we found a similar interaction with student preparedness. In that work, we measured the cohesion of tutorial dialog in a way

similar to the lexical reiteration cohesion baseline described in Section 4. We found that the amount of cohesion in our tutorial dialogs predicted learning for our below-mean, but not our above-mean pre-testers. In that paper [11], we speculated that maybe cohesion could predict learning for our high-knowledge students, but we were measuring the wrong kind of cohesion. Perhaps measuring the reiteration of senses (meanings), rather than of words, would correlate with learning in high-knowledge students. In this work we have implemented that idea, and report on the results.

We find that in one corpus of tutoring dialogs, adding a count of semantic similarity reiterations to our lexical reiteration measure does indeed improve its correlation with learning among above-mean pre-test students. We find that lowering a similarity threshold so that more semantically distant pairs are counted as cohesive ties improves this correlation. We also find that tutor-to-student and student-to-tutor cohesive ties are equally well correlated with learning. Finally, we present suggestive evidence that our correlations may be with the deeper learning measured by "far-transfer," as opposed to "near-transfer" questions.

2 Related Work

In Section 5, we discuss a method by which dialog cohesiveness can be calculated using a WordNet [12] based measure of semantic similarity. Many measures of semantic similarity based on the WordNet taxonomy have been described in the computational linguistics literature. These measures range from counting edges [13], to adjusting edge counts with other information such as depth in the taxonomy [14] or information content calculated from a corpus [15].

These systems are typically evaluated by comparing them to human judgments or by seeing how they perform in tasks such as spelling correction [16]. We differ from the semantic similarity work mentioned above in that we apply our measure to tutorial dialog, and evaluate it by how strongly it correlates with learning in our corpora of tutorial dialogs. Pending future work, we use the simplest reported measure of semantic similarity.

Other work examining the cohesiveness of tutorial dialog has been done by the AutoTutor group at the University of Memphis. In [17], they use the CohMetrix [18] cohesion analysis tool to analyze the cohesiveness of tutor and student dialog contributions along many dimensions. Our semantic measures are similar in spirit, but where they use LSA to gauge the distributional similarity between two turns, we use a WordNet similarity metric to locate specific pairs of similar words between turns. Their "argument overlap" metric is also very similar to our lexical reiteration measure. However, we look for correlations between dialog cohesion and learning, whereas [17] examines cohesion differences between tutoring and other types of discourse.

3 Tutoring Dialog Corpora

We test our model on two corpora of tutoring transcripts collected by the Itspoke intelligent tutoring system project [5] in 2003 and 2005. Itspoke is a speech enhanced version of the Why2 qualitative physics tutoring system [19]. In both

experiments, the Itspoke tutor taught qualitative physics to students who had never taken physics before. The tutor was identical in each case, except that the version used in 2005 had a larger language model to improve speech recognition during dialog. In this work, we use dialog transcriptions, rather than speech recognizer output. Students for the 2003 study were recruited by flyer, whereas students in 2005 were recruited via their "Introduction to Psychology" instructor, as well as by flyer.

In each experiment, the students would first read instructional material about physics, and take a pre-test to gauge their physics knowledge. The tutor would then present a problem in qualitative physics, which the student would answer in essay form. The computer tutor would interpret the essay and engage the student in a spoken dialog to teach a missing or incorrect point. This would repeat until the tutor was satisfied that all points were covered. Each student worked through five problems this way, then took a post-test.

The 2003 and 2005 pre and post tests contained 26 questions in common. The tests used in 2005 contained an additional 14 questions re-used from other experiments. A pilot tagging study suggests that the full 40 question set (40Q) used in 2005 contains about 50% "far" transfer questions that were non-isomorphic to the tutored problems. However, the 26 question set (26Q) is about 27%, and the 14 question set (14Q) is over 90% "far" transfer. A firm classification of the questions into "near" and "far" is left for a more formal tagging study. Here, we present results for the three 2005 question sets separately, and suggest that their differences may be because of differing proportions of "far" transfer questions.

	2003		**2005 40Q**		**2005 26Q**		**2005 14Q**	
Group	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
All Pre	0.48	0.17	0.54	0.17	0.49	0.18	0.61	0.18
All Post	0.69	0.18	0.71	0.14	0.70	0.16	0.70	0.15
High Pre	0.67	0.14	0.68	0.09	0.65	0.10	0.75	0.09
High Post	0.79	0.13	0.82	0.09	0.82	0.10	0.78	0.14
Low Pre	0.38	0.06	0.41	0.10	0.35	0.09	0.45	0.10
Low Post	0.64	0.18	0.61	0.11	0.61	0.14	0.60	0.09

Table 1. Test Scores

There were twenty students in the 2003 study, who completed a total of ninety-five dialogs. A mean pre-test split divided these students into 13 "low" pre-testers and 7 "high" pre-testers. There were 34 students in the 2005 study, who completed 163 dialogs.[1] A mean pre-test split using the 40Q or 26Q test results divides these students into 18 "low" and 16 "high" pre-testers. Using the 14Q set divides them into 16 "low" and 18 "high" pre-testers. In each experiment, pre-test splits were done relative to the question set being used. Mean (M) pre and post-test scores, with standard deviations (SD) are shown in Table 1 for each pre-test group (All students, High & Low pre-testers).

[1] Dialogs were not always collected for every one of a student's five problems, because the computer tutor would sometimes accept the initial essay without discussion.

4 Baseline Cohesion Measure - Lexical Reiteration

As mentioned in Section 1, we want to know if measuring cohesion at the "sense" level will improve our previous lexical cohesion measure. Our baseline, therefore, will be a lexical reiteration cohesion measure similar to the one used in [11]. This measures the cohesiveness of a dialog by counting the number of token and stem matches between utterances, after removing stop words. A stem is the "root" form of a word, which we find using a standard Porter stemmer. An illustration of this measure is shown in Table 2. The top two lines of the table show two consecutive utterances from one of our dialogs. Nine cohesive ties can be counted between these utterances at the token level. The tokens matched are shown in row three of the table, and in bold in the utterances. Cohesive ties can also be counted at the stem level, by counting a tie whenever one utterance and the next contain words with the same stem. An example of this is shown in row four of Table 2, where the tokens "force" and "forces" have matched by stem. In both of our measures, cohesive ties are counted between all consecutive pairs of utterances, ie: both from tutor-to-student and student-to-tutor. This measure is a close approximation to the "exact word repetition" type of cohesion described by Halliday and Hassan [20] in *Cohesion in English* .

Speaker	**Utterance**
Student	Before the **release** of the **keys**, the man's and the keys velocity are the same. **After** the release the **only force** on the keys and **man** is downward force of earth's **gravity**, **so** they are in freefall. We **can** ignore the **forces** that the air exerts on these objects since they are dense. Therefore, at every **point** in time the keys will remain in front of the man's face during their whole trip down.
Tutor	**So** you **can** compare it to your response, here's my summary of a missing **point**: **After** the **release**, the **only force** on the **person**, **keys**, and elevator is the **force** of **gravity**. Kindly correct your essay. If you're finished, press the submit button.
Level	**Cohesive Ties Counted between Utterances, at each level**
Token	so-so, release-release, point-point, only-only, keys-keys, gravity-gravity, can-can, after-after, force-force
Stem	forces-force
Sem	man-person

Table 2. Token, Stem, and Semantic Similarity (Sem) Matches

We count the total number of cohesive ties for each dialog as described above. We then line normalize the count, dividing it by the total number of lines in the dialog. We do this to remove the possibility that the count of cohesive ties correlates with learning simply because the longer dialogs had more cohesive ties. However, neither the total number of tutor turns, student turns, tutor words, or student words are correlated with learning in spoken dialogs with our computer

tutor [5]. In the example shown in Table 2, we count a total of 10 cohesive ties at the token and stem levels, line-normalizing the count (as if this were an entire dialog) would give a score of $10/2 = 5$.

Finally, we sum the line normalized counts over all dialogs for each student, resulting in a per-student cohesion measure which we correlate with learning. As in previous work [11], we use partial correlations of post-test score with our cohesion count measures, controlling for pre-test score. We control for pre-test score because it is significantly correlated with post-test score in both our 2003 corpus ($r(18)$=.462, p=.04) and in our 2005 corpus (40Q:$r(32)$=.817; 26Q:$r(32)$=.741; 14Q:$r(32)$=.698, all $p<$ .001).

5 New Cohesion Measure - Semantic Similarity

We next extend the baseline measure described above by counting semantic similarity, as well as lexical reiteration cohesive ties. We count a cohesive tie at the semantic similarity level whenever one utterance and the next have different words with similar meanings, and we measure similarity using WordNet [12]. WordNet is a large semantic lexicon with several features that are useful here. First, it groups words into groups of synonyms called "synsets." Second, it organizes synsets into an "is-a" (hypernym/hyponym) taxonomy. If we know the sense in which a word is being used, and therefore its relevant synset, we can use WordNet to find its relationship to other synsets in the taxonomy. An example of these relationships can be seen in Figure 1, which reproduces a portion of the WordNet taxonomy. For one sense of "man," we can see in WordNet that man is a type of male, and that male is a type of animal. Man is also a type of adult, which is a type of person.

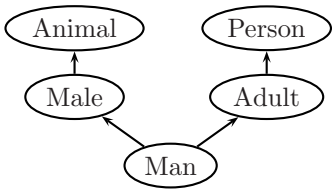

Fig. 1. Wordnet paths

We do not attempt to do word sense disambiguation before measuring the similarity of two words in WordNet. Instead, for each potential pair of words, we choose the senses in which the words are most similar.

We measure semantic similarity as a function of the distance between two concepts in the WordNet hierarchy. In this work, we use the simplest method of measuring this distance: Path Distance Similarity, as implemented in NLTK [21]. This measure calculates the similarity between two concepts as 1/1+N, where N is the number of edges in the shortest path between them. Scores range from zero to one, where zero means no similarity and one indicates exact synonyms. In Figure 1, the shortest path between "man" and "person" has two edges, and so their similarity is 1/1+2, or.333.

5.1 Identifying Semantic Ties

Finding semantic similarity ties is slightly more complicated than finding lexical reiteration ties because a word in one utterance may have non-zero similarities to several words in the other utterance. Therefore, we use the following algorithm

to find the best set of ties. For each word in utterance B, we look up the WordNet path similarity values between it and each word in utterance A. After collecting the set of all possible word pairs this way, we sort them by their similarity values. Starting at the high end of the list, for each pair we remove all lower pairs which have the same token in the same position. This process is illustrated in Table 3. To keep the example small, we have selected only two tutor and three student words from the example shown in Table 2. This produces six possible pairs, which are shown in columns two and three of Table 3, sorted by their similarity values. Starting at the top of the list, we consider first the pair: "man-person." We remove all instances below of "man" in position A and of "person" in position B. This step is shown under "Step 1" in Table 3. In step 2, we move down to the next remaining pair, "release-elevator." We remove all instances below that of "release" in position A and of "elevator" in position B. There are no pairs remaining to be considered in our example, so we stop and count two semantic cohesive ties: "man-person" with a similarity of .33, and "release-elevator" with a similarity of .13.

	Start		**Step 1**		**Step 2**	
Sim	**Tok A**	**Tok B**	**Tok A**	**Tok B**	**Tok A**	**Tok B**
0.33	man	person	**man**	**person**	man	person
0.13	release	person	release			
0.13	release	elevator	release	elevator	**release**	**elevator**
0.11	velocity	person	velocity		velocity	
0.09	man	elevator		elevator		
0.07	velocity	elevator	velocity	elevator	velocity	

Table 3. Finding the best semantic ties

This method can count cohesive ties with a broad range of similarity scores. We will investigate whether the stronger ties are more useful by instituting a threshold, and only counting cohesive ties for pairs with similarity values above the threshold. In the example shown in Table 3, a threshold of .3 would count the tie between "person" and "man" but not between "elevator" and "release."

A threshold $> .5$ counts cohesive ties only for word pairings which are listed in WordNet as being exact synonyms, and which therefore have a similarity score of one (note from the path similarity formula that scores between .5 and 1 are impossible). A threshold reduced to .3 allows cohesive ties with slightly more semantic distance in the pair, and a threshold of 0 allows all pairs found by our algorithm. Examples of cohesive ties counted at each of these thresholds are shown in Table 4. In these

Threshold		
> 0.5	**0.3**	**0**
5-five	motion-contact	remains-same
remain-stay	man-person	man-runner
speed-velocity	decrease-acceleration	force-magnitude
conclude-reason	acceleration-change	summarize-point
package-packet	travel-flying	submit-pull

Table 4. Example Semantic ties

examples we can see that the matches counted become more distant and less sensible as the threshold is reduced.

5.2 Semantic Similarity Measure

We count the number of cohesive ties between two utterances by first counting all the exact token matches between them, then counting ties based on stem matches as described in Section 4. After these lexical reiteration ties are identified, we look for semantic similarity ties among the remaining words.

An example of an additional cohesive tie counted at the semantic similarity level is shown in row five of Table 2. Here a tie between the tokens "man" and "person" has been counted which, as shown in Table 3, have a semantic similarity of .33. Adding this tie to the baseline measure from Section 4 brings our cohesive tie count to 10.33, and our normalized cohesion score for the example to 5.16.

6 Results

pre-test Group	**2003**		**2005 40Q**		**2005 26Q**		**2005 14Q**	
	Cor	**pVal**	**Cor**	**pVal**	**Cor**	**pVal**	**Cor**	**pVal**
	Lexical Only							
All	0.474	**0.035**	0.273	0.118	0.185	0.295	0.289	0.098
Low	0.682	**0.005**	0.606	**0.013**	0.279	0.263	0.462	0.072
High	0.798	0.105	0.152	0.546	0.084	0.756	0.333	0.177
Lexical plus WordNet Similarity Threshold = .5 to .99								
All	0.470	**0.036**	0.276	0.114	0.187	0.289	0.298	0.087
Low	0.686	**0.005**	0.605	**0.012**	0.286	0.250	0.473	0.064
High	0.825	0.085	0.159	0.527	0.084	0.757	0.336	0.173
Threshold = .3								
All	0.470	**0.037**	0.277	0.112	0.182	0.303	0.308	0.076
Low	0.689	**0.004**	0.613	**0.011**	0.271	0.276	0.495	0.051
High	0.899	**0.038**	0.153	0.543	0.070	0.797	0.341	0.166
Threshold = 0								
All	0.451	**0.046**	0.286	0.100	0.183	0.301	0.337	0.051
Low	0.665	**0.007**	0.607	**0.012**	0.259	0.300	0.519	**0.039**
High	0.984	**0.002**	0.161	0.522	0.082	0.763	0.378	0.122

Table 5. Learning-Cohesion Correlations

The top section of Table 5 shows results for our lexical measure alone. Here cohesive ties are counted for token and stem matches between utterances, but not for semantic similarity matches. In the 2003 corpus (cols 2 & 3), this measure produces significant correlations with learning for below mean pre-testers, and for the group of all students. It does not produce significant correlations with learning for above mean pre-testers. This pattern is similar to the one reported in [11] [2].

Next we examine results for the 2003 corpus after adding the semantic similarity measure to the previous lexical reiteration measure, shown in the lower three sections of Table 5. As the threshold is reduced, correlations for high pre-testers become significant and increasingly stronger. Fisher's z-test indicates

[2] The correlations shown here are slightly different from those reported in [11] because of small differences in implementation.

that the improvement in high pre-tester correlations between the > .5 and 0 thresholds is significant (p <= .0003).

Results for the 2005 corpus are shown in the right three sections of Table 5. Unfortunately, the success of the semantic similarity measure among high pre-testers does not replicate in this corpus. Our measure correlates with learning only among low pre-testers. However, comparing results from different question sets gives us some insight into what sort of learning is correlating. Note that we get strong, significant correlations for low pre-testers in the 40Q question set (cols 4 & 5), which we have argued includes 50% far transfer questions. For the 26Q set (cols 6 & 7), which has fewer far transfer, we get no correlations. However for the 14Q set (cols 8 & 9), which is probably almost all far transfer, we see a significant correlation for the semantic measure at a threshold of zero, but not for the lexical measure. This suggests that our semantic measure may be correlating with the deeper learning measured by far transfer questions.

As described in Section 4, the results presented in Table 5 are bi-directional, ties are counted both when the student's utterance follows the tutor's, and the other way around. It is interesting, however, to consider whether one direction of tie is more important than the other. Results for uni-directional cohesive ties are shown in Table 6, for the 2003 corpus and for the 2005 corpus with the full question set. For brevity we present only significant correlations with a threshold of .3. Comparing these results to those shown in Table 5 for the same threshold indicates that both directions are equally responsible for our results among low pre-testers. This suggests the possibility of increasing learning by altering tutor word choice to manipulate dialog cohesion.

pre test	**2003**		**2005 40Q**	
Group	**Cor**	**pVal**	**Cor**	**pVal**
Threshold = .3				
Tutor to Student				
Low	0.682	**0.005**	0.602	**0.014**
Student to Tutor				
Low	0.634	**0.011**	0.593	**0.015**

Table 6. Directional Correlations

7 Discussion

As mentioned in Section 1, this work grew out of speculations we developed when trying to explain the results reported in [11]. We hypothesized that the lexical reiteration cohesive ties we were counting signaled inferences which led to learning among our low pre-testers. We wondered if being able to recognize cohesive ties between different words with similar meanings would allow us to detect the deeper inferences that might lead to learning among high pre-testers. For example, perhaps hearing (or producing) "person" when the dialog partner just used "man" is associated with deeper inference than simply re-using "man."

Results from both corpora suggest a relationship between inference and cohesion. In the 2003 corpus, counting semantic reiteration cohesive ties does improve learning correlations for high pre-testers. Also, counting more semantically distant pairs, which may represent deeper inferences, improves this correlation.

In our 2005 corpus, correlations with "near transfer" learning seem to be weaker than correlations with "far transfer" learning, using a preliminary division of questions into "near" and "far". Also, correlations with "far transfer

only" (14Q) learning improve with lower thresholds, becoming significant at a threshold of zero. This also supports a link between cohesion, inference and learning. Performance on "far transfer" tasks is often thought to be impeded if the original knowledge is encoded too simply [22]. Counting ties at low thresholds may partially measure the inferential elaboration which aids far transfer.

The comparative weakness of the 05 results, especially among high pre-testers is unexplained. However, the 05 corpus has a stronger pre-to-post test score correlation than the 03 corpus (.741(26Q) vs .462), suggesting that perhaps the tutorial dialog is responsible for correspondingly less of the learning gain. If so, we might also expect lower correlations between dialog features and learning. Other work [23] has also found these corpora to be quite different with respect to a suite of quantitative evaluation metrics.

8 Conclusions and Future Work

We have shown that measuring cohesion as the repetition of meanings, as opposed to only the repetition of words (or stems), improves the measure's correlation with learning among high pre-testers in one of our corpora of tutoring dialogs. Also, these correlations improve as we allow more semantically distant matches, which presumably require deeper inference to produce or understand.

Results from our other corpus suggest that counting semantic repetition cohesive ties predicts the deeper learning measured by "far transfer" problems.

In future work we hope to make a more formal division of questions into "near" and "far" to confirm the relationship between cohesion and far transfer learning. We are also currently extending this work to see if it will replicate in other corpora of tutoring dialogs. If these results replicate successfully, we will experiment with using tutor word choice to manipulate cohesion during tutoring.

9 Acknowledgments

This research is supported by the ONR (N00014-07-1-0039). We gratefully thank our anonymous reviewers, Bob Hausmann, Chas Murray, Mihai Rotaru and the ITSPOKE group for many helpful comments.

References

1. Bloom, B.S.: The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher **13** (1984) 4 – 16
2. Merrill, D., Reiser, B., Ranney, M., Trafton, J.: Effective tutoring techniques: A comparison of human tutors and intelligent tutoring systems. The Journal of the Learning Sciences **2** (1992) 277 – 306
3. Graesser, A.C., Person, N., Magliano, J.P.: Collaborative dialogue patterns in naturalistic one-to-one tutoring. Applied Cognitive Psychology **9** (1995) 495 – 522
4. Forbes-Riley, K., Litman, D., Huettner, A., Ward, A.: Dialogue-learning correlations in spoken dialogue tutoring. In: Proceedings 12th International Conference on Artificial Intelligence Education (AIED),Amsterdam, Netherlands. (July 2005)
5. Litman, D.J., Rose, C.P., Forbes-Riley, K., VanLehn, K., Bhembe, D., Silliman, S.: Spoken versus typed human and computer dialogue tutoring. International Journal of Artificial Intelligence in Education **16** (2006) 145 – 170

6. Forbes-Riley, K., Litman, D., Amruta Purandare, M.R., Tetreault, J.: Comparing linguistic features for modeling learning in computer tutoring. In: Proceedings 13th International Conference on Artificial Intelligence Education (AIED), Los Angeles, Ca. (2007)
7. VanLehn, K., Graesser, A., Jackson, G., Jordan, P., Olney, A., Rose, C.: When are tutorial dialogues more effective than reading? Cognitive Science **30** (2006) 1 – 60
8. Kalyuga, S., Ayres, P.: The expertise reversal effect. Educational Psychologist **38** (2003) 23 – 31
9. Conati, C., VanLehn, K.: Further results from the evaluation of an intelligent computer tutor to coach self-explanation. 5th International Conference on Intelligent Tutoring Systems (2000) 304 – 313
10. McNamara, D.S., Kintsch, W.: Learning from text: Effects of prior knowledge and text coherence. Discourse Processes **22** (1996) 247–287
11. Ward, A., Litman, D.: Cohesion and learning in a tutorial spoken dialog system. In: Proceedings of the 19th International FLAIRS Conference. (2006) 533–538
12. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: Introduction to wordnet: An on-line lexical database. International Journal of Lexicography (special issue) **3 (4)** (1990) 235–312
13. Rada, R., Mili, H., Bicknell, E., Bletner, M.: Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics **19(1)** (1989) 17 – 30
14. Sussna, M.: Word sense disambiguation for free-text indexing using a massive semantic network. In: CIKM '93: Proceedings of the second international conference on Information and knowledge management, New York, NY, USA, ACM (1993) 67–74
15. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. Proceedings of the 14th International Joint Conference on Artificial Intelligence (August 1995) 448 – 453
16. Budanitsky, A., Hirst, G.: Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. Workshop on WordNet and Other Lexical Resources, in the North American Chapter of the Association for Computational Linguistics (June 2001)
17. Graesser, A., Jeon, M., Yan, Y., Cai, Z.: Discourse cohesion in text and tutorial dialogue. Information Design Journal **15** (2007) 199 – 213
18. Graesser, A.C., McNamara, D.S., Louwerse, M.M., Cai, Z.: Coh-metrix: Analysis of text on cohesion and language. Behavior, Research Methods, Instruments, and Computers **36** (2004) 192 – 202
19. VanLehn, K., Jordan, P.W., Rose, C.P., Bhembe, D., Boettner, M., Gaydos, A., Makatchev, M., Pappuswamy, U., Ringenberg, M., Roque, A., Siler, S., Srivastava, R.: 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., Springer (2002) 158–167
20. Halliday, M.A.K., Hasan, R.: Cohesion in English. English Language Series. Pearson Education Limited (1976)
21. Loper, E., Bird, S.: Nltk: The natural language toolkit (2002)
22. Stark, R., Mandl, H., Gruber, H., Renkl, A.: Instructional means to overcome transfer problems in the domain of economics: empirical studies. International Journal of Educational Research **31** (1999) 591 – 609
23. Ai, H., Litman, D.: Comparing real-real, simulated-simulated, and simulated-real spoken dialogue corpora. In: Proceedings of the AAAI Workshop on Statistical and Empirical Approaches for Spoken Dialogue Systems, Boston, MA. (2006)