An Orthonormal Basis for Topic Segmentation in Tutorial Dialogue

Andrew Olney
Department of Computer Science
University of Memphis
Memphis, TN 38152
aolney@memphis.edu

Zhiqiang Cai
Institute for Intelligent Systems
University of Memphis
Memphis, TN 38152
zcai@memphis.edu

Abstract

This paper explores the segmentation of tutorial dialogue into cohesive topics. A latent semantic space was created using conversations from human to human tutoring transcripts, allowing cohesion between utterances to be measured using vector similarity. Previous cohesion-based segmentation methods that focus on expository monologue are reapplied to these dialogues to create benchmarks for performance. A novel moving window technique using orthonormal bases of semantic vectors significantly outperforms these benchmarks on this dialogue segmentation task.

1 Introduction

Ever since Morris and Hirst (1991)'s groundbreaking paper, topic segmentation has been a steadily growing research area in computational linguistics, with applications in summarization (Barzilay and Elhadad, 1997), information retrieval (Salton and Allan, 1994), and text understanding (Kozima, 1993). Topic segmentation likewise has multiple educational applications, such as question answering, detecting student initiative, and assessing student answers.

There have been essentially two approaches to topic segmentation in the past. The first of these, lexical cohesion, may be used for either linear segmentation (Morris and Hirst, 1991; Hearst, 1997) or hierarchical segmentation (Yarri, 1997; Choi, 2000). The essential idea behind the lexical cohesion approaches is that different topics will have different vocabularies. Therefore the lexical cohesion within topics will be higher than the lexical cohesion between topics, and gaps in cohesion may mark topic boundaries. The second major approach to topic segmentation looks for distinctive textual or acoustic markers of topic boundaries, e.g. referential noun phrases or pauses (Passonneau and Litman, 1993; Passonneau and Litman, 1997). By using multiple markers and machine learning methods, topic segmentation algorithms may be developed using this second approach that have a higher accuracy than methods using a single marker alone (Passonneau and Litman, 1997).

The primary technique used in previous studies, lexical cohesion, is no stranger to the educational NLP community. Lexical cohesion measured by latent semantic analysis (LSA) (Landauer and Dumais, 1997; Dumais, 1993; Manning and Schütze, 1999) has been used in automated essay grading (Landauer, Foltz, and Laham, 1998) and in understanding student input during tutorial dialogue (Graesser et al., 2001). The present paper investigates an orthonormal basis of LSA vectors, currently used by the AutoTutor ITS to assess student answers (Hu et al., 2003), and how it may be used to segment tutorial dialogue.

The focus on dialogue distinguishes our work from virtually all previous work on topic segmentation: prior studies have focused on monologue rather than dialogue. Without dialogue, previous approaches have only limited relevance to interactive educational applications such as intelligent tutoring systems (ITS). The only existing work on topic segmentation in dialogue, Galley et al. (2003), segments recorded speech between multiple persons using both lexical cohesion and dis-

tinctive textual and acoustic markers. The present work differs from Galley et al. (2003) in two respects, viz. we focus solely on textual information and we directly address the problem of tutorial dialogue.

In this study we apply the methods of Foltz et al. (1998), Hearst (1994, 1997), and a new technique utilizing an orthonormal basis to topic segmentation of tutorial dialogue. All three are vector space methods that measure lexical cohesion to determine topic shifts. Our results show that the new using an orthonormal basis significantly outperforms the other methods.

Section 2 reviews previous work, and Section 3 reviews the vector space model. Section 4 introduces an extension of the vector space model which uses an orthonormal basis. Section 5 outlines the task domain of tutorial dialogue, and Section 6 presents the results of previous and the current method on this task domain. A discussion and comparison of these results takes place in Section 7. Section 8 concludes.

2 Previous work

Though the idea of using lexical cohesion to segment text has the advantages of simplicity and intuitive appeal, it lacks a unique implementation. An implementation must define how to represent units of text, compare the cohesion between units, and determine whether the results of comparison indicate a new text segment. Both Hearst (1994, 1997) and Foltz et al. (1998) use vector space methods discussed below to represent and compare units of text. The comparisons can be characterized by a moving window, where successive overlapping comparisons are advanced by one unit of text. However, Hearst (1994, 1997) and Foltz et al. (1998) differ on how text units are defined and on how to interpret the results of a comparison.

The text unit's definition in Hearst (1994, 1997) and Foltz et al. (1998) is generally task dependent, depending on what size gives the best results. For example, when measuring comprehension, Foltz et al. (1998) use the unit of the sentence, as opposed to the more standard unit of the proposition, because LSA is most correlated with comprehension at that level. However, when using LSA to segment text, Foltz et al. (1998) use the paragraph as the unit, to "smooth out" the local changes in cohesion and become more sensitive to more global changes of cohesion. Hearst likewise chooses a large unit, 6 token-sequences of 20 tokens (Hearst, 1994), but varies these parameters dependent on the characteristics of the text to be segmented, e.g. paragraph size.

Under a vector space model, comparisons are performed by calculating the cosine of vectors representing text. As stated previously, these comparisons reflect the cohesion between units of text. In order to use these comparisons to segment text, however, one must have a criterion in place. Foltz et al. (1998), noting mean cosines of .16 for boundaries and .43 for non-boundaries, choose a threshold criterion of .15, which is two standard deviations below the boundary mean of .43. Using LSA and this criterion, Foltz et al. (1998) detected chapter boundaries with an F-measure of .33 (see Manning and Schütze (1999) for a definition of F-measure). Hearst (1994, 1997) in contrast uses a relative comparison of cohesion, by recasting vector comparisons as depth scores. A depth score is computed as the difference between a given vector comparison and its surrounding peaks, i.e. the local maxima of vector comparisons on either side of the given vector comparison. The greater the difference between a given comparison and its surrounding peaks, the higher the depth score. Once all the depth scores are calculated for a text, those that are higher than one standard deviation below the mean are taken as topic boundaries. Using a vector space method without singular value decomposition, Hearst (1997) reports an F-measure of .70 when detecting topic shifts between paragraphs. Thus previous work suggests that the Hearst (1997) method is superior to that of Foltz et al. (1998), having roughly twice the accuracy indicated by F-measure. Although these two results used different data sets and are therefore not directly comparable, one would predict based on this limited evidence that the Hearst algorithm would outperform the Foltz algorithm on other topic segmentation tasks.

3 The vector space model

The vector space model is a statistical technique that represents the similarity between collections of words as a cosine between vectors (Manning and Schütze, 1999). The process begins by collecting text into a corpus. A matrix is created from the corpus, having one row for each unique word in the corpus and one column for each document or paragraph. The cells of the matrix consist of a simple count of the number of times word *i* appeared in document *j*. Since many words do not appear in any given document, the matrix is often sparse. Weightings are applied to the cells that take into account the frequency of word *i* in document *j* and the frequency of word *i* across all documents, such that distinctive words that appear infrequently are given the most weight. Two collections of words of arbitrary size are compared by creating two vectors. Each word is associated with a row vector in the matrix, and the vector of a collection is simply the sum of all the row vectors of words in that collection. Vectors are compared geometrically by the cosine of the angle between them.

LSA (Landauer and Dumais, 1997; Dumais 1993) is an extension of the vector space model that uses singular value decomposition (SVD). SVD is a technique that creates an approximation of the original word by document matrix. After SVD, the original matrix is equal to the product of three matrices, word by singular value, singular value by singular value, and singular value by document. The size of each singular value corresponds to the amount of variance captured by a particular dimension of the matrix. Because the singular values are ordered in decreasing size, it is possible to remove the smaller dimensions and still account for most of the variance. The approximation to the original matrix is optimal, in the least squares sense, for any number of dimensions one would choose. In addition, the removal of smaller dimensions introduces linear dependencies between words that are distinct only in dimensions that account for the least variance. Consequently, two words that were distant in the original space can be near in the compressed space, causing the inductive machine learning and knowledge acquisition effects reported in the literature (Landauer and Dumais, 1997).

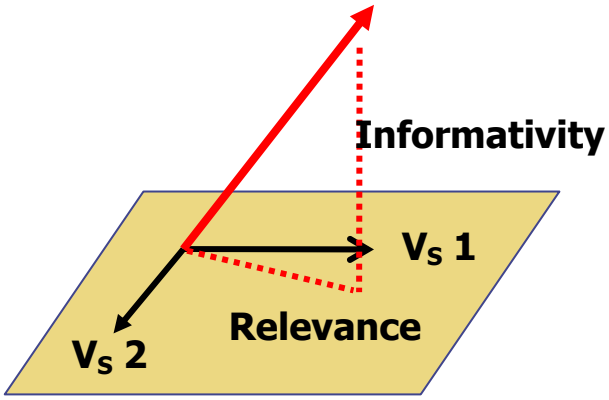

Figure 1. Projecting a new utterance to the basis

4 An orthonormal basis

Cohesion can be measured by comparing the cosines of two successive sentences or paragraphs (Foltz, Kintsch, and Landauer, 1998). However, cohesion is a crude measure: repetitions of a single sentence will be highly cohesive (cosine of 1) even though no new information is introduced. A variation of the LSA algorithm using orthonormalized vectors provides two new measures, "informativity" and "relevance", which can detect how much new information is added and how relevant it is in a context (Hu et al., 2003). The essential idea is to represent context by an orthonormalized basis of vectors, one vector for each utterance. The basis is a subspace of the higher dimensional LSA space, in the same way as a plane or line is a subspace of 3D space. The basis is created by projecting each utterance vector onto the basis of previous utterance vectors using a method known as the Gram-Schmidt process (Anton, 2000). Each projected utterance vector has two components, a component parallel to the basis and a component perpendicular to the basis. These two components represent "informativity" and "relevance", respectively. Let us first consider "relevance". Since each vector in the basis is orthogonal, the basis represents all linear combinations of what has been previously said. Therefore the component of a new utterance vector that is parallel to the basis is already represented by a linear combination of the existing vectors. "Informativity" follows similarly: it is the perpendicular component of a new utterance vector that can not be represented by the existing basis vectors. For example, in Figure 1, a new utterance creates a new vector that can be projected to the basis, forming a triangle. The leg of the triangle that lies

along the basis indicates the "relevance" of the recent utterance to the basis; the perpendicular leg indicates new information. Accordingly, a repeated utterance would have complete "relevance" but zero new information.

5 Procedure

The task domain is a subset of conversations from human-human computer mediated tutoring sessions on Newton's Three Laws of Motion, in which tutor and tutee engaged in a chat room-style conversation. The benefits of this task domain are twofold. Firstly, the conversations are already transcribed. Additionally, tutors were instructed to introduce problems using a fixed set of scripted problem statements. Therefore each topic shift corresponds to a distinct problem introduced by the tutor. Clearly this problem would be trivial for a cue phrase based approach, which could learn the finite set of problem introductions. However, the current lexical approach does not have this luxury: words in the problem statements recur throughout the following dialogue.

Human to human computer mediated physics tutoring transcripts first were removed of all markup, translated to lower case, and each utterance was broken into a separate paragraph. An LSA space was made with these paragraphs alone, approximately one megabyte of text. The conversations were then randomly assigned to training (21 conversations) and testing (22 conversations). The average number of utterances per topic, 16 utterances, and the average number of words per utterance, 32 words, were calculated to determine the parameters of the segmentation methods. For example, a moving window size greater than 16 utterances implies that, in the majority of occurrences, the moving window straddles three topics as opposed to the desired two.

To replicate Foltz et al. (1998), software was written in Java that created a moving window of varying sizes on the input text, and the software retrieved the LSA vector and calculated the cosine of each window. Hearst (1994, 1997) was replicated using the JTextTile (Choi, 1999) Java software. A variant of Hearst (1994, 1997) was created by using LSA instead of the standard vector space method. The orthonormal basis method also used a moving window; however, in contrast to the previous methods, the window is not treated just as a large block of text. Instead, the window consists of two orthonormal bases, one on either side of an utterance. That is, a region of utterances above the test utterance is projected, utterance by utterance, into an orthonormal basis, and likewise a region of utterances below the test utterance is projected into another orthonormal basis. Then the test utterance is projected into each orthonormal basis, yielding measures of "relevance" and "informativity" with respect to each. Next the elements that make up each orthonormal basis are aggregated into a block, and a cosine is calculated between the test utterance and the blocks on either side, producing a total of six measures.

Each tutoring session consists of the same 10 problems, discussed between one of a set of 4 tutors and one of 18 subjects. The redundancy provides a variety of speaking and interaction styles on the same topic.

```
Tutor: A clown is riding a
unicycle in a straight line.
She accidentally drops an egg
beside her as she continues
to move with constant veloc-
ity. Where will the egg land
relative to the point where
the unicycle touches the
ground?  Explain.
Student: The egg should land
right next to the unicycle.
The egg has a constant hori-
zontal velocity.  The verti-
cal velocity changes and
decreases as gravity pulls
the egg downward at a rate of
9.8m/s^2.  The egg should
therefore land right next to
the unicycle.
Tutor: Good! There is only
one thing I would like to
know. What can you say about
the horizontal velocity of
the egg compared to the hori-
zontal velocity of the clown?
Student: Aren't they the
same?
```

All of the 10 problems are designed to require application of Newton's Laws to be solved, and

therefore conversations share many terms such as force, velocity, acceleration, gravity, etc.

6 Results

For each method, the development set was first used to establish the parameters such as text unit size and classification criterion. The methods, tuned to these parameters, were then applied to the testing data.

6.1 Foltz et al. (1998)

In order to replicate Foltz et al.'s results, a text unit size and window size needed to be chosen. The utterance was chosen as the text unit size, which included single word utterances, full sentences, and multi-sentence utterances. To determine the most appropriate window size, results from all sizes between 1 and 16 (the average number of utterances between topic shifts) were gathered. The greatest difference between the means for utterances that introduce a topic shift versus non-shift utterances occurs when the window contains four utterances. The standard deviation is uniformly low for windows containing more than two utterances and therefore can be disregarded in choosing a window size.

The optimal cosine threshold for classification was found using logistic regression (Garson, 2003) which establishes a relationship between the cosine threshold and the log odds of classification. The optimal cutoff was found to be shift odds = .17 with associated F-measure of .49. The logistic equation of best fit is:

$$\ln(\text{shift odds}) = 1.887 + (-13.345 \cdot \text{cosine})$$

F-measure of .49 is 48% higher than the F-measure reported by Foltz et al. (1998) for segmenting monologue. On the testing corpus the F-measure is .52, which demonstrates good generalization for the logistic equation given. Compared the F-measure of .33 reported by Foltz et al. (1998), the current result is 58% higher.

6.2 Hearst (1994, 1997)

The JTextTile software was used to implement Hearst (1994) on dialogue. As with Foltz et al. (1998), a text unit and window size had to be determined for dialogue. Hearst (1994) recommends using the average paragraph size as the window size. Using the development corpus's average topic length of 16 utterances as a reference point, F-measures were calculated for the combinations of window size and text unit size in Table 1.

The optimal combination of parameters (F-measure = .17) is a unit size of 16 words and a window size of 16 units. This combination matches Hearst (1994)'s heuristic of choosing the window size to be the average paragraph length.

		Window size				
		2	**4**	**8**	**16**	**32**
Unit size	**8**	.134	.129	.130	.146	.144
	16	.142	.133	.130	.171	.140
	32	.138	.132	.130	.151	.143

Table 1. Unit vs. window size for Hearst method

On the test set, this combination of parameters yielded an F-measure of .14 as opposed to the F-measure for monologue reported by Hearst (1997), .70. For dialogue, the algorithm is 20% as effective as it is for monologue. It is unclear, however, exactly what part of the algorithm contributes to this poor performance. The two most obvious possibilities are the segmentation criterion, i.e. depth scores, or the standard vector space method.

To further explore these possibilities, the Hearst method was augmented with LSA. Again, the unit size and window size had to be calculated. As with Foltz, the unit size was taken to be the utterance. The window size was determined by computing F-measures on the development corpus for all sizes between 1 and 16. The optimal window size is 9, F-measure = .22. Given the smaller number of test cases, 22, this F-measure of .22 is not significantly different from .17. However, the Foltz method is significantly higher than both of these, $p < .10$.

6.3 Orthonormal basis

The text unit used in the orthonormal basis is the single utterance. The optimal window size, i.e. the orthonormal basis size, was determined by creating a logistic regression to calculate the maximum F-measure for several orthonormal basis sizes. The findings of this procedure are listed in Table 2.

Size	3	4	5	6	8	10	15
F	.59	.63	.65	.72	.73	.72	.73

Table 2. F-measure for orthonormal basis sizes

F-measure monotonically increases until the orthonormal basis holds six elements and holds relatively steady for larger orthonormal basis sizes. Since F-measure does not increase much over .72 for greater orthonormal basis sizes, 6 was chosen as the most computationally efficient size for the strength of the effect. The logistic equation of best fit is:

$$\begin{aligned}\ln(\text{shift odds}) = {} & 20.027 \\ & + (16.703 \cdot \text{cosine}_2) \\ & + (-30.843 \cdot \text{relevance}_1) \\ & + (-23.567 \cdot \text{informativity}_1) \\ & + (-2.698 \cdot \text{relevance}_2) \\ & + (2.771 \cdot \text{informativity}_2)\end{aligned}$$

Where the index of 1 indicates a measure on the window preceding the utterance, and an index of 2 indicates a measure on the window following the utterance. In the regression, the cosine between the utterance and the preceding window was not significant, $p = .86$. This finding reflects the intuition that the cosine to the following window varies according to whether the following window is on a new topic, whereas the cosine to the preceding window is always high. Additionally, measures of "relevance" and "informativity" correspond to vector length; all other measures did not contribute significantly to the model and so were not included.

The sign of the metrics illuminates their role in the model. The negative sign on the coefficients for relevance_1, informativity_1, and relevance_2 indicates that they are inversely correlated with an utterance signaling the start of a new topic. The only surprising feature is that informativity_1 is negatively correlated instead of positively correlated: one would expect a topic shift to introduce new information. There is possibly some edge effect here, since the last move of a topic is often a summarizing move that shares many of the physics terms present in the introduction of a new topic. On the other hand, the positive sign on cosine_2 and informativity_2 indicates that the start of a new topic should have elements in common with the following material and add new information to that material, as an overview would. Beyond the sign, the exponentials of these values indicate how the two basis metrics are weighted. For example, when informativity_2 is raised by one unit, a topic shift is 16 times more likely.

On the testing corpus the F-measure of the orthonormal basis method is .67, which is significantly different from the performance of all three methods mentioned above, $p < .05$. Table 3 compares this result with the previous results in the current study for segmenting dialogue.

Method	Hearst	Hearst + LSA	Foltz	Orth. basis
F	.14	.22	.52	.67

Table 3. Comparison of dialogue segmentation methods

7 Discussion

The relative ranking of these results is not altogether surprising given the relationships between inferencing and LSA and between inferencing and dialogue. Foltz et al. (1998) found that LSA makes simple bridging inferences in addition to detecting lexical cohesion. These bridging inferences are a kind of collocational cohesion (Halliday and Hassan, 1976) whereby words that co-occur in similar contexts become highly related in the LSA space. Therefore in applications where this kind of inferencing is required, one might expect an LSA based method to excel.

Similarly to van Dijk and Kintsch's model of comprehension (van Dijk and Kintsch, 1983), dialogue can require inferences to maintain coherence. According to Grice's Co-operative Principle, utterances lacking semantic coherence flout the Maxim of Relevance and license an inference (Grice, 1975):

```
S1: Let's go dancing.
S2: I have an exam tomorrow.
```

The "inference" in the sense of Foltz, Kintsch, and Landauer (1998) would be represented by a high cosine between these utterances, even though they don't share any of the same words. Dialogue generally tends to be less lexically cohesive and require more inferencing than expository mono-

logue, so one might predict that LSA would excel in dialogue applications.

However, LSA has a weakness: the cosine measure between two vectors does not change monotonically as new word vectors are added to either of the two vectors. Accordingly, the addition of a word vector can cause the cosine between two text units to dramatically increase or decrease. Therefore the distinctive properties of individual words can be lost with the addition of more words to a text unit. This problem can be addressed by using an orthonormal basis (Hu et al., 2003). By using a basis, each utterance is kept independent, so "inferencing" can extend over both the entire set of utterances and the linear combination of any of its subsets. Accordingly, when "inferencing" over the entire text unit is required, one would expect a basis method using LSA vectors to outperform a standard LSA method. This expectation has been put to the test recently by Olney & Cai (2005), who find that an orthonormal basis can significantly predict entailment on test data supplied by the PASCAL Textual Entailment Challenge (PASCAL, 2004).

Beyond relative performance rankings, more support for the above reasoning can be found in the difference between Hearst and Hearst + LSA. Recall that in monologue, Hearst (1997) reports a much larger F-measure than Foltz et al. (1998), .70 vs. .33, albeit on different data sets. In the present dialogue corpus, these roles are reversed, .14 vs. .52. Possible reasons for this reversal are the segmentation criterion, the vector space method, or the fact that Foltz has been trained on similar data via regression and Hearst has not. However, comparing the Hearst algorithm with the Hearst + LSA algorithm indicates that a 57% improvement stems from the addition of LSA, keeping all other factors constant. While this result is not statistically significant, the direction of the result supports the use of an "inferencing" vector space method for segmenting dialogue.

Unfortunately, the large difference in F-measure between the Foltz algorithm and the Hearst + LSA algorithm is more difficult to explain. These two methods differ by their segmentation criterion and by their training (Foltz is a regression model and Hearst is not). It may be that Hearst (1994, 1997)'s segmentation criterion, i.e. depth scores, do not translate well to dialogue. Perhaps the assignment of segment boundaries based on the relative difference between a candidate score and its surrounding peaks is highly sensitive to cohesion gaps created by conversational implicatures. On the other hand the differences between these two methods may be entirely attributable to the amount of training they received. One way to separate the contributions of the segmentation criterion and training would be to create a logistic model using the Hearst + LSA method and to compare this to Foltz.

The increased effectiveness of the orthonormal basis method over the Foltz algorithm can also be explained in terms of "inferencing". Since "inferencing" is overwhelmed by lexical cohesion (Foltz et al., 1998), the increase in window size for the Foltz algorithm deteriorates performance for a window size greater than 4. In contrast, the orthonormal basis method becomes most effective as the orthonormal basis size increases past 4. This dichotomy illustrates that the Foltz algorithm is not complementary to an "inferencing" approach in general. Use of an orthonormal basis, on the other hand, increases sensitivity to collocational cohesion without sacrificing lexical cohesion.

8 Conclusion

This study explored the segmentation of tutorial dialogue using techniques that have previously been applied to expository monologue and using a new orthonormal basis technique. The techniques previously applied to monologue reversed their roles of effectiveness when applied to dialogue. This role reversal suggests the predominance of collocational cohesion, requiring "inferencing", present in this tutorial dialogue. The orthonormal basis method, which we suggest has an increased capacity for "inferencing", outperformed both of the techniques previously applied to monologue, and demonstrates that segmentation of these tutorial dialogues most benefits from a method sensitive to lexical and collocational cohesion over large text units.

Acknowledgements

This research was supported by the National Science Foundation (SBR 9720314, REC 0089271, REC 0106965, REC 0126265) and the DoD Multidisciplinary University Research Initiative (MURI) administered by ONR under grant N00014-00-1-0600. Any opinions, findings, and

conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DoD, ONR, or NSF.

References

Anton, H. (2000). Elementary linear algebra. 8th edition. New York: John Wiley.

Barzilay, R. & Elhadad, M. (1997). Using Lexical Chains for Text Summarization. *Proceedings of the Intelligent Scalable Text Summarization Workshop*.

Choi, F. (1999). JTextTile: A free platform independent text segmentation algorithm. http://www.cs.man.ac.uk/~choif

Choi, F. (2000). Advances in domain independent linear text segmentation. In *Proceedings of the NAACL'00*, May.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of Discourse Comprehension*. New York: Academic Press.

Dumais, S. (1993). LSI meets TREC: a status report. In *Proceedings of the First Text Retrieval Conference* (TREC1), 137-152. NIST Special Publication 500-207.

Foltz, P.W., Kintsch, W. & Landauer, T.K. (1998). The measurement of textual cohesion with latent semantic analysis. *Discourse Processes, 25*, 285-307.

Galley, M., McKeown, K., Fosler-Lussier, E., & Jing, H. (2003). Discourse Segmentation of Multi-Party Conversation. *Proceedings of the ACL.*

Garson, D. Logistic Regression. Accessed on April 18th, 2003. http://www2.chass.ncsu.edu/garson/pa765/logistic

Graesser, A. C., Person, N. K., Harter, D., & the Tutoring Research Group. (2001). Teaching tactics and dialogue in AutoTutor. *International Journal of Artificial Intelligence in Education, 12*, 257-279.

Grice, H.P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds) *Syntax and Semantics* Vol 3. 41-58. New York: Academic.

Grosz, B.J. & Sidner, C.L. (1986). Attention, Intentions, and the structure of discourse. *Computational Linguistics, 12* (3), 175-204.

Halliday, M. A. & Hassan, R. A. (1976). *Cohesion in English*. London: Longman.

Hearst, M. (1994). Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd meeting of the Association for Computational Linguistics*. 9-16.

Hearst, M. (1997). Text-Tiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics,* 23(1), 33-64.

Hu, X., Cai, Z., Louwerse, M., Olney, A., Penumatsa, P., and Graesser, A. (2003). An improved LSA algorithm to evaluate contributions in student dialogue. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence* (IJCAI-03), 1489-1491.

Kozima, H. (1993). Text segmentation based on similarity between words. *In Proceedings of ACL '93*, 286-288.

Landauer, T. & Dumais, S. (1997). A solution to Plato 's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes, 25*, 259-284.

Manning, C. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

Morris, J. & 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., & Cai, Z. (2005). An Orthonormal Basis for Entailment. In Proceedings of the Eighteenth International Florida Artificial Intelligence Research Society Conference, 554-559. Menlo Park, Calif.: AAAI Press.

PASCAL. 2004. Recognising Textual Entailment Challenge. Accessed on October 4th, 2004. http://www.pascal-network.org/Challenges/RTE/

Passonneau, R. J. & Litman, D. J. (1993). Intention-based Segmentation: Human Reliability and correlation with linguistic cues. *Proceedings of the ACL*, 148-155.

Passonneau, R. J. & Litman, D. J. (1997). Discourse segmentation by human and automated means. *Computational Linguistics, 23*(1), 103–139.

Salton, G. & Allan, J. (1994). Automatic text decomposition and structuring. *In Proceedings of RIAO*, 6–29, New York, NY.

Yaari, Y. (1997). Segmentation of expository texts by hierarchical agglomerative clustering. *Proceedings of the RANLP'97.*