Running head: TEXTUAL COHERENCE USING LATENT SEMANTIC ANALYSIS

The Measurement of Textual Coherence with Latent Semantic Analysis

Peter W. Foltz

New Mexico State University

Walter Kintsch and Thomas K. Landauer

University of Colorado

Abstract

Latent Semantic Analysis is used as a technique for measuring the coherence of texts. By comparing the vectors for two adjoining segments of text in a high-dimensional semantic space, the method provides a characterization of the degree of semantic relatedness between the segments. We illustrate the approach for predicting coherence through re-analyzing sets of texts from two studies that manipulated the coherence of texts and assessed readers' comprehension. The results indicate that the method is able to predict the effect of text coherence on comprehension and is more effective than simple term-term overlap measures. In this manner, LSA can be applied as an automated method that produces coherence predictions similar to propositional modeling. We describe additional studies investigating the application of LSA to analyzing discourse structure and examine the potential of LSA as a psychological model of coherence effects in text comprehension.

The Measurement of Textual Coherence with Latent Semantic Analysis.

In order to comprehend a text, a reader must create a well connected representation of the information in it. This connected representation is based on linking related pieces of textual information that occur throughout the text. The linking of information is a process of determining and maintaining coherence. Because coherence is a central issue to text comprehension, a large number of studies have investigated the process readers use to maintain coherence and to model the readers' representation of the textual information as well as of their previous knowledge (e.g., Lorch & O'Brien, 1995)

There are many aspects of a discourse that contribute to coherence, including, coreference, causal relationships, connectives, and signals.   For example, Kintsch and van Dijk (Kintsch, 1988; Kintsch & van Dijk, 1978) have emphasized the effect of coreference in coherence through propositional modeling of texts. While coreference captures one aspect of coherence, it is highly correlated with other coherence factors such as causal relationships found in the text (Fletcher, Chrysler, van den Broek, Deaton, & Bloom, 1995; Trabasso, Secco & van den Broek, 1984).

Although a propositional model of a text can predict readers' comprehension, a problem with the approach is that in-depth propositional analysis is time consuming and requires a considerable amount of training. Semi-automatic methods of propositional coding (e.g., Turner, 1987) still require a large amount of effort. This degree of effort limits the size of the text that can be analyzed. Thus, most texts analyzed and used in reading comprehension experiments have been small, typically from 50 to 500 words, and almost all are under 1000 words. Automated methods such as readability measures (e.g., Flesch, 1948; Klare, 1963) provide another characterization of the text, however, they do not correlate well with comprehension measures (Britton & Gulgoz, 1991; Kintsch & Vipond, 1979). Thus, while the coherence of a text can be measured, it can often involve considerable effort.

In this study, we use Latent Semantic Analysis (LSA) to determine the coherence of texts. A more complete description of the method and approach to using LSA may be found in Deerwester, Dumais, Furnas, Landauer and Harshman, (1990), Landauer and Dumais, (1997), as well as in the preceding article by Landauer, Foltz and Laham (this issue).   LSA provides a fully automatic method for comparing units of textual information to each other in order to determine their semantic relatedness.   These units of text are compared to each other using a derived measure of their similarity of meaning. This measure is based on a

powerful mathematical analysis of direct and indirect relations among words and passages in a large training corpus. Semantic relatedness so measured, should correspond to a measure of coherence since it captures the extent to which two text units are discussing semantically related information.

Unlike methods which rely on counting literal word overlap between units of text, LSA's comparisons are based on a derived semantic relatedness measure which reflects semantic similarity among synonyms, antonyms, hyponyms, compounds, and other words that tend to be used in similar contexts. In this way, it can reflect coherence due to automatic inferences made by readers as well as to literal surface coreference. In addition, since LSA is automatic, there are no constraints on the size of the text analyzed. This permits analyses of much larger texts to examine aspects of their discourse structure.

In order for LSA to be considered an appropriate approach for modeling text coherence, we first establish how well LSA captures elements of coherence that are similar to modeling methods such as propositional models. A re-analysis of two studies that examined the role of coherence in readers' comprehension is described. This re-analysis of the texts produces automatic predictions of the coherence of texts which are then compared to measures of the readers' comprehension.   We next describe the application of the method to investigating other features of the discourse structure of texts. Finally, we illustrate how the approach applies both as a tool for text researchers and as a theoretical model of text coherence.

General approach for using LSA to measure coherence

The primary method for using LSA to make coherence predictions is to compare some unit of text to an adjoining unit of text in order to determine the degree to which the two are semantically related.   These units could be sentences, paragraphs or even individual words or whole books. This analysis can then be performed for all pairs of adjoining text units in order to characterize the overall coherence of the text.   Coherence predictions have typically been performed at a propositional level, in which a set of propositions all contained within working memory are compared or connected to each other (e.g., Kintsch, 1988, In press).   For LSA coherence analyses, using sentences as the basic unit of text appears to be an appropriate corresponding level that can be easily parsed by automated methods. Sentences serve as a good level in that they represent a small set of textual information (e.g., typically 3-7 propositions) and thus would be approximately consistent with the amount of information that is held in short term memory.

As discussed in the preceding article by Landauer, et al. (this issue), the power of computing semantic relatedness with LSA comes from analyzing a large number of text examples. Thus, for computing the coherence of a target text, it may first be necessary to have another set of texts that contain a large proportion of the terms used in the target text and that have occurrences in many contexts.   One approach is to use a large number of encyclopedia articles on similar topics as the target text. A singular value decomposition (SVD) is then performed on the term by article matrix, thereby generating a high dimensional semantic space which contains most of the terms used in the target text.

Individual terms, as well as larger text units such as sentences, can be represented as vectors in this space. Each text unit is represented as the weighted average of vectors of the terms it contains. Typically the weighting is by the log entropy transform of each term (see Landauer, et al., this issue).   This weighting helps account for both the term's importance in the particular unit as well as the degree to which the term carries information in the domain of discourse in general. The semantic relatedness of two text units can then be compared by determining the cosine between the vectors for the two units. Thus, to find the coherence between the first and second sentence of a text, the cosine between the vectors for the two sentences would be determined.   For instance, two sentences that use exactly the same terms with the same frequencies will have a cosine of 1, while two sentences that use no terms that are semantically related, will tend to have cosines near 0 or below. At intermediate levels, sentences containing terms of related meaning, even if none are the same terms or roots will have more moderate cosines. (It is even possible, although in practice very rare, that two sentences with no words of obvious similarity will have similar overall meanings as indicated by similar LSA vectors in the high dimensional semantic space.)

Coherence and text comprehension

This paper illustrates a complementary approach to propositional modeling for determining coherence, using LSA, and comparing the predicted coherence to measures of the readers' comprehension.   For these analyses, the texts and comprehension measures are taken from two previous studies by Britton and Gulgoz (1988), and, McNamara, et al. (1996).

In the first study, the text coherence was manipulated primarily by varying the amount of sentence to sentence repetition of particular important content words through analyzing propositional overlap. Simulating its results with LSA demonstrates the degree to which coherence is carried, or at least reflected, in the

continuity of lexical semantics, and shows that LSA correctly captures these effects. However, for these texts, a simpler literal word overlap measure, absent any explicit propositional or LSA analysis, also predicts comprehension very well.

The second set of texts, those from McNamara et al. (1996), manipulates coherence in much subtler ways; often by substituting words and phrases of related meaning but containing different lexical items to provide the conceptual bridges between one sentence and the next. These materials provide a much more rigorous and interesting test of the LSA technique by requiring it to detect underlying meaning similarities in the absence of literal word repetition. The success of this simulation, and its superiority to direct word overlap predictions, is the principal demonstration of the effectiveness of the LSA coherence measure and forms the basis of additional findings reported in the remainder of the paper.

Coherence analysis of Britton and Gulgoz texts

Using a text on the airwar in Vietnam from an Air Force training textbook, Britton and Gulgoz revised the text using several different methods. In their Principled revision of the text, they employed the Miller and Kintsch (1980) computer program to propositionalize the text and predict areas in the text where the coherence broke down.   In each place that there was an identified gap in coherence due to lack of argument overlap between propositions, they repaired the text so that there would be argument overlap. These repairs typically took the form of repeating a word that was used in a previous proposition. In a second type of revision of the text, the Heuristic revision, the text was revised by hand with the overall goal to create the best possible revision of the text. This involved such improvements as clarifying important points, reordering the presentation of ideas, and omitting information that was regarded as unimportant. In a third revision of the text, the Readability revision, they used readability formula scores to revise the original text so that it had a lower grade level readability score that was comparable to the heuristic revision of the text.

Britton and Gulgoz then assessed readers' comprehension of the original text and the three revisions of the text using a variety of measures. They found that the Principled and Heuristic revisions of the text resulted in significantly better comprehension than the Original or Readability revisions on three measures: the number of propositions recalled in free recall, the efficiency (the number of propositions recalled per minute of reading time) and scores on a multiple choice inference test. Overall, their results indicate that improving a text through

modeling propositional overlap can result in improvement in readers' comprehension of that text.

We used LSA to analyze the sentence-to-sentence coherence of the four texts from the Britton and Gulgoz' experiment. A 300 dimension semantic space was constructed based on a the first 2000 characters or less of each of 30,473 articles from Groliers' Academic American Encyclopedia (see Landauer & Dumais, 1997). After separating each of the four texts into individual sentences, the vector for each sentence was computed (as the weighted sum of it its weighted terms) and then was compared to the vector for the next sentence in the text. In determining the vectors, the 459 most frequent terms in the English language (e.g., the, and, from, etc.) were omitted from the analyses[1]. The cosine between these two vectors indicated their semantic relatedness or coherence. An overall coherence measure was then calculated for each text by averaging the cosines between the vectors for all pairs of adjoining sentences.   The average cosines for the four texts are presented in Table 1. An ANOVA on the individual sentence-to-sentence cosines comparing the four texts showed significant overall differences between the texts $F(3,181)=16.8$, $p<.001$[2]. A post-hoc Fisher's PLSD test showed that both the Heuristic and Principled revision texts had significantly higher cosines than the Original and Readability texts (Heuristic vs. Original, mean diff=.211, critical diff=.076, p<.001; Heuristic vs. Readability, mean diff=.211, critical diff=.074, p<.001; Principled vs. Original, mean diff=.155, critical diff=.070, p<.001; Principled vs. Readability, mean diff=.155, critical diff=.069, p<.001.) There were no significant differences between the Heuristic and Principled revisions and between the Original and Readability texts.

Text	LSA coherence	Weighted word overlap	No. props recalled	Efficiency (props/min.)	Inference mult. choice
Original	0.192	0.047	35.5	3.44	37.11
Readability revision	0.193	0.073	32.8	3.57	29.74
Principled revision	0.347	0.204	58.6	5.24	46.44
Heuristic revision	0.403	0.225	56.2	6.01	48.23

Table 1. <u>LSA coherence, weighted word overlap and comprehension measures for the Britton and Gulgoz texts.</u>

The averaged sentence-to-sentence cosines for each text were then compared against the three comprehension measures that showed significant differences between the texts from the Britton and Gulgoz study. The LSA coherence predictions were significantly correlated with all three measures. (Number of propositions recalled r=0.98, p<.05; Efficiency, r=0.99, p<.05; Inference multiple choice, r=1.00, p<.01). Figure 1 shows the relationship between the average cosine and the subjects' performance on the inference test. Overall, the results indicate that the coherence predictions are highly correlated with several ways of characterizing the readers' comprehension. Thus, the LSA coherence measure appears to provide an accurate measure of the comprehensibility of the texts.
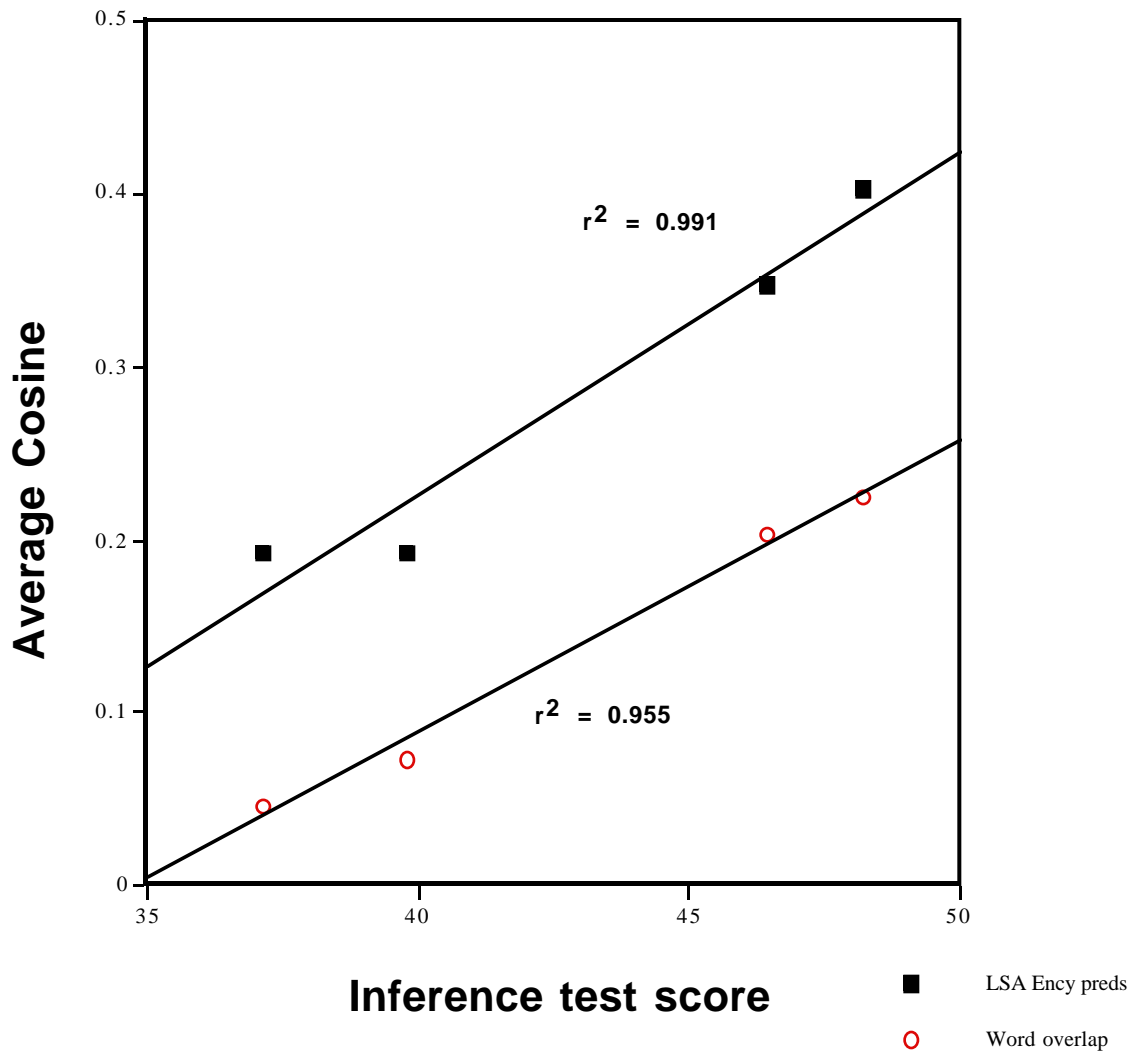
Figure 1. Average cosine versus inference test score for LSA and word overlap measures.

Part of the general effectiveness of LSA for text related applications is that it makes comparisons of textual information based on the derived semantic similarity between words. Thus, it is able to compare vectors of textual information that do not contain the same words. Nevertheless, two vectors that consist of many of the same words will tend be highly similar. In this way, LSA is highly sensitive to direct sentence-to-sentence word overlap. Therefore, it is important to determine the extent to which the coherence predictions were based on just direct word overlap as

opposed to indirect semantic overlap. To calculate a coherence measure based on literal word overlap, vectors for each sentence in the complete term by sentence matrix for each of the texts were compared to each other.   In order to keep the analysis equivalent to the LSA predictions, the same log entropy weighting was used for the terms and the most frequent terms in the English language were omitted from the analysis.   This approach was equivalent to performing the LSA coherence analysis without the dimensional reduction that is performed by the Singular Value Decomposition. As in LSA, this produces a cosine between the two vectors, although this cosine is now just a function of the number of the same words used in two adjoining sentences weighted by a function of their frequency distribution[3]. For each text these cosines were averaged to generate an overall coherence measure. The predictions for the measure of word overlap coherence are shown in Table 1 and Figure 1.

The word overlap predictions of coherence were essentially equivalent to those of the LSA coherence predictions. Additionally, the correlations with the comprehension measures were all significant (Number of propositions recalled r= 0.96, p<.05, Efficiency, r=1.00, p<.01, Inference multiple choice, r=0.98, p<.05).   The fact that the word overlap and LSA predictions are equivalent indicates that the primary change from the original text to the revised texts is the improvement in the number of literal words that overlap between sentences. Indeed, excluding the high frequency terms, in the Original text, 63 percent of the sentence transitions had no word overlap; in the Principled revision, only 16 percent of the sentence transitions had no word overlap and in the Heuristic revision only 10 percent of the sentences had no word overlap. This finding is consistent with the approach that was used by Britton and Gulgoz. In using the van Dijk and Kintsch (1978, 1983) model for identifying and repairing coherence breaks, the repair method involves inserting words that would increase the direct argument overlap in propositions between sentences.   Since LSA is highly sensitive to direct word overlap, as well as sensitive to a lesser degree to indirect semantic relatedness between words, the large effects of direct word overlap tend to overwhelm the other effects due to indirect semantic relatedness. Thus, it is no surprise that LSA is able to capture the effects of the improvements in coherence and how they affect the readers' comprehension in this rather trivial or degenerate case. However, the next case is a greater challenge.

Coherence analysis of McNamara et al. texts

The study by McNamara, et al. (1996) was designed to examine how the readers' previous knowledge interacted with the coherence of a text. In their second

experiment, they modified a student science encyclopedia article on heart disease by adding or deleting information to vary the amount of local and macro coherence. The changes to the text included such revisions as: replacing pronouns with noun phrases, adding descriptive elaborations, adding sentence connectives, replacing words to improve argument overlap, and adding topic headers and macropropositions to link paragraphs to the text and the topic. In replacing words to improve argument overlap, they did not always repeat words, but often used words of related meaning.   Their changes resulted in four texts: a maximally coherent text (CM), a text with high local coherence, but low macrocoherence (Cm), a text with low local coherence, but high macrocoherence (cM), and a text with both low local and macrocoherence (cm). Through evaluating the reader's prior knowledge on the topic, they found that readers with low knowledge benefited the most from the maximally coherent text, while high-knowledge readers benefited more from the minimally coherent text. Since low-knowledge readers were most affected by the effects of increasing coherence, their comprehension results were used to compare to the LSA coherence predictions. For our analysis using LSA, each of the four texts was separated into sentence units as in the previous analysis of the Britton and Gulgoz texts.

One question raised by using LSA to model coherence is the degree to which the initial set of texts used to create the LSA space affects the predictions. In the analysis of the Britton and Gulgoz texts, the LSA space was based on the 30,047 encyclopedia articles from Grolier's encyclopedia.   In the new analysis, in addition to using the large set of articles (large ency) for coherence predictions, a second smaller LSA space was derived from a small set of encyclopedia articles on the heart (small ency).   This space was developed by retaining 100 factors of an SVD on the matrix of 830 sentences by 2781 unique words from 24 Grolier's encyclopedia articles related to the heart and heart disease. The smaller space still contained most of the terms used in the target texts. In addition, since the SVD analysis was performed on the co-occurrence of terms across sentences (as opposed to articles in the large ency), it still provided enough text examples to permit the characterization of semantic relatedness beyond simple word overlap. The comparison of the two approaches using the larger versus the smaller set of encyclopedia articles permits a measure of the generalizability of this method to using different sets of documents to create the initial LSA space. For each of the two spaces, the vector for each sentence in each of the texts was compared to the vector for the following sentence. The cosines were

then averaged in order to provide a coherence measure. Along with these measures, a weighted word overlap measure was computed.

The three coherence measures are shown in Table 2. The two LSA measures produced comparable results, predicting the lowest coherence for the cm text, moderate coherence for the cM and Cm texts, and the greatest coherence for the CM text. These predictions are consistent with the modifications made to the texts. An analysis of variance on the individual sentence-to-sentence cosines, however, did not show any significant differences between the texts $(F(3,247)=0.77, p=.51)$. This is likely due to the high variance for the cosines for all four texts. The range of sentence-to-sentence cosines was from near 0 to 0.91, and the standard deviations for the texts ranged from 0.21 to 0.25.

Text	LSA small ency coherence	LSA large ency coherence	Weighted word overlap
cm	0.178	0.320	0.155
cM	0.209	0.346	0.147
Cm	0.203	0.374	0.152
CM	0.238	0.399	0.163

Table 2. The three coherence measures for the McNamara et al. texts.

While the LSA measures show a pattern consistent with the type of coherence revisions made to the text, the weighted word overlap measure predicted very little difference between the four texts. In fact, the method predicted slightly more coherence in the minimally coherent text than the two texts which had either high macro or local coherence. This result indicates that, although the texts had been revised to improve coherence, the coherence improvements did not involve the addition of additional argument overlap by literal repetition from sentence to sentence. Instead the improvements appear to be due to the general flow of semantic content independent of argument overlap. Thus, the LSA measures capture some effects of coherence that are not found in direct word overlap.

In order to determine the difference between the LSA and word overlap methods, we examined individual sentence transitions which have divergent

predictions. These transitions were located by computing Z-scores for the cosines for word overlap and the small ency LSA predictions and determining where these Z-scores differed.   For example, in the CM text, the Z-score difference for the transition between sentence 7 and 8 was 2.14. The sentences were:

> There are many kinds of heart disease, some of which are present at birth and some of which are acquired later.

> 1. Congenital heart disease
> A congenital heart disease is a defect that a baby is born with.

In the word overlap measure, the cosine between the two sentences was 0.09. While the words, <u>heart</u> and <u>disease</u> are repeated across the two sentences, these two terms occur with high frequency in the originally scaled encyclopedia articles and thus have very little information value using the log entropy weighting. For this reason, they contribute very little to determining the centroid of the vectors compared to other terms used in the sentences. On the other hand, for the LSA scaling, the cosine between the two sentences is 0.69. By examining individual terms in the LSA space, we can see why this prediction is much greater than that of word overlap. To a reader, seeing the word <u>birth,</u> which occurs in the first sentence, and seeing words like <u>baby</u> and <u>born</u> in the second sentence may provide markers that the two sentences are related.   Comparing individual terms in the LSA space, <u>birth</u> has a cosine of 0.56 with <u>baby</u> and a cosine of 0.33 with <u>congenital.</u> The term <u>born</u> does not occur in the original 24 encyclopedia articles used for the LSA scaling and thus does not contribute to the analysis. Had the word <u>born</u> been in those articles, it would likely only strengthen the predicted relationship since it would probably be highly related to the term <u>birth</u>. In addition, the terms <u>born</u> and <u>baby,</u> have much greater weights than terms like <u>heart</u> or <u>disease,</u> since they do not occur as frequently within the context of the heart articles. Thus, the cosine between the sentences using LSA is much greater than in word overlap because it is capturing the degree to which the sentences discuss a similar semantic content by means other than literal word repetition.

An additional issue suggested by this example, is the role of the information value of words used when computing coherence. Just because a term occurs in two propositions, does not mean that linking the two propositions should always contribute greatly to coherence. For example, repeating the term <u>heart,</u> within a text about the heart, should not contribute much to the overall coherence of a text.

Similarly with LSA, it is not just whether two terms share similar semantic content, but also the degree to which they have high information value that helps determine the amount of coherence between two sentences. Therefore, for determining coherence, it is not just that terms are repeated or are used in semantically related ways, but also the relationship of those terms to the overall text that is important.

As in the Britton and Gulgoz study, the McNamara et al. study showed that subjects with low knowledge of the topic obtained the greatest benefit from the maximally coherent text. Although they found no significant differences in the proportion of propositions recalled between texts, they did find an interaction on posttest questions between the maximally and minimally coherent texts and the level of the subjects' knowledge. Low-knowledge subjects showed the strongest effects of which text they read. Therefore, we compared the low knowledge subjects' post-test scores against the LSA and word overlap coherence predictions.

The LSA coherence measures correlated strongly with the subjects' overall posttest scores, (small ency: $r=.94$, $p=.08$, large ency: $r=.85$, $p=.21$), but did not correlate well with the word overlap measure ($r=.19$, $p=.85$). Figure 2 shows the relationship of the LSA and word overlap predictions to the posttest scores. The posttest questions were comprised of text-based, bridging inference, elaborative and problem solving questions. Both LSA measures correlated most strongly with the subjects' performance on the text-based questions. (small ency: $r=.98$, $p<.05$, large ency: $r=.84$, $p=.23$) This is consistent with the notion that a highly coherent text should be most helpful for building a well linked textbase in low-knowledge readers.
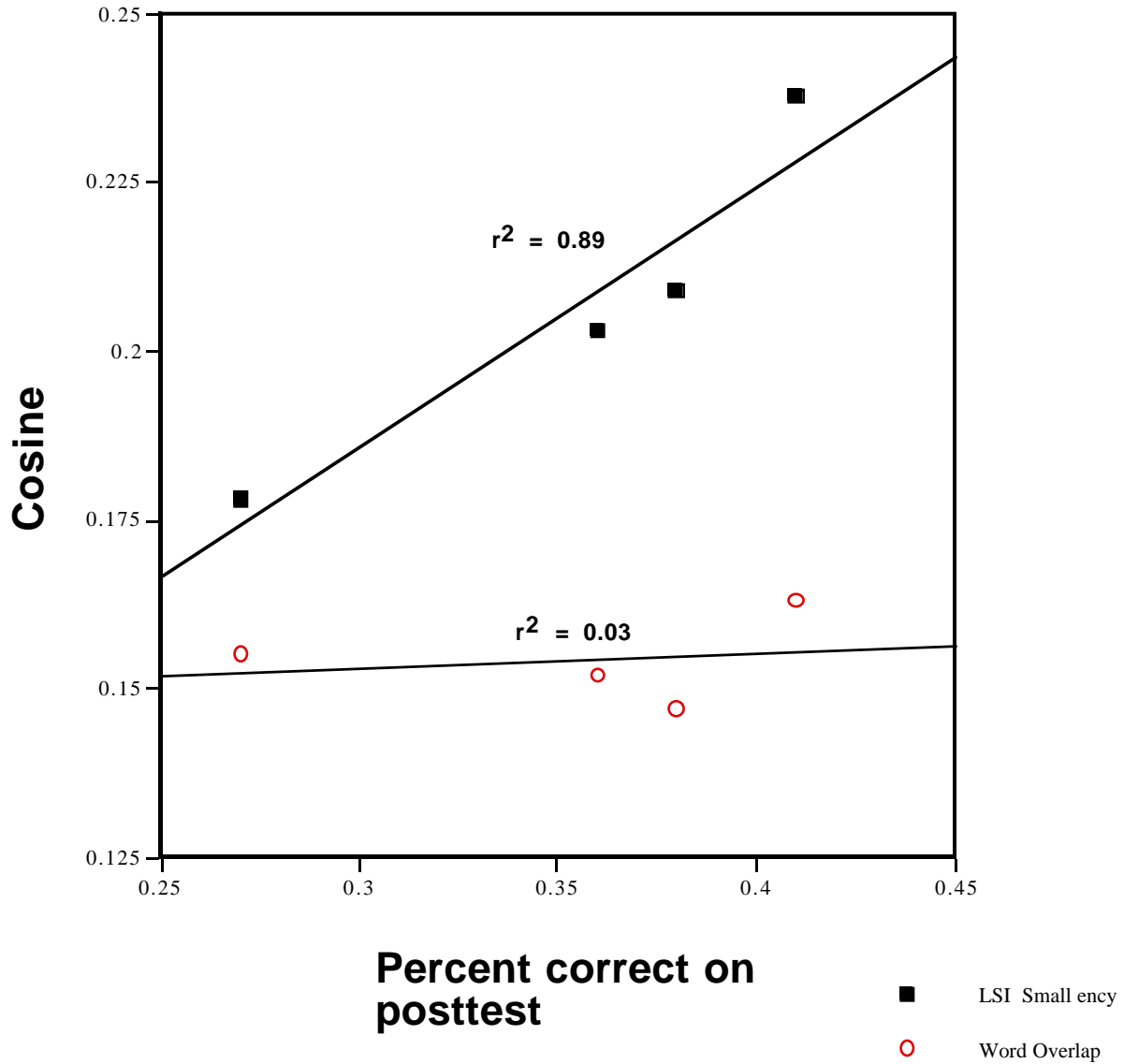
<u>Figure 2</u>. LSA and word overlap coherence measures vs. posttest performance for low-knowledge subjects.

Readability measures have also been used as an approach to characterizing the quality of texts (e.g., Flesch, 1948; Klare, 1963). The Flesch grade level score was calculated for the four texts in order to determine whether the readability measures corresponded to LSA's predictions or to the subjects' performance on the post-tests. For the readability measure, there were essentially no differences between the four texts for the Flesch grade level (CM: 7.5, cM: 7.5, Cm: 7.4, cm: 7.4). Thus, LSA's

coherence measure provided a more effective characterization of the subjects' performance on the posttests than readability measures.

As in the re-analysis of the Britton and Gulgoz data, the overall results indicate that the LSA coherence measure predicts readers' comprehension quite well.   However, unlike the Britton and Gulgoz re-analysis case, LSA here provides a much better account of the coherence of the texts than word overlap.

Additional applications for automatic computations of coherence

In the above analyses, we have illustrated how LSA can be applied to modeling discourse coherence for predicting readers' comprehension. Because LSA is an automatic method, it permits the analysis of much larger texts than are typically used in text comprehension research. Below we illustrate two additional applications for LSA. The first is predicting the discourse structure of a book by determining the breaks between chapters, while the second is an analysis of how the topic of a text changes across the text of an entire book.

Discourse segmentation.

In discourse segmentation, the goal is to identify locations in the text where topic shifts occur, so that the text can be segmented into discrete topics. Morris and Hirst (1991) have suggested that the discourse structure of a text can be determined through an analysis of lexical cohesion. Using hand coding, they used a thesaurus to identify chains of related words across sentences. Breaks in these lexical chains tended to indicate structural elements in the text, such as changes in topics and the writer's intentional structures (e.g., Grosz & Sidner, 1986). In an extension of this work, Hearst (1993) developed an automatic method that employed weighted term vectors, a sliding window and lexical disambiguation based on a thesaurus to predict readers' judgments of topic shifts within short scientific articles.

Discourse segmentation is based on the premise that the coherence should be lower in areas of the discourse where the discourse topic changes. LSA can perform a similar analysis to that of Hearst (1993), although LSA's lexical relations are based on the derived semantic similarity rather than using a thesaurus. To test LSA's ability to segment discourse, we used an introductory psychology textbook (Myers, 1995) and tried to predict the breaks between the 19 chapters. The textbook was separated into paragraphs and the matrix of 4903 paragraphs by 19160 unique terms was analyzed with LSA, retaining 300 factors.

To perform the coherence analysis, 770 paragraphs were first removed from the text. These paragraphs represented references, problem sets, and glossaries from the back of each chapter. Since these items tend not to be connected discourse, they

would have skewed the results by identifying large drops in coherence at the end of each chapter that were not actually parts of the authors' text.   Thus, the remaining 4133 paragraphs represented just the raw continuous text in which we then analyzed the paragraph-to-paragraph cosines.

Initial tests on the paragraph-to paragraph cosines indicated a lot of variability in the coherence from one paragraph to the next. In order to smooth the predictions, we used a sliding window in which we compared the last 10 paragraphs to the next 10 paragraphs. The window would then move ahead 1 paragraph to make the comparison of the next group of 10 paragraphs. This approach tends to remove the effects due to very local coherence changes that occur between paragraphs, while still detecting much larger changes in the global coherence between groups of paragraphs. Using the sliding window, the average cosine between paragraphs was 0.43 (stddev=0.14), while the average cosine between paragraphs at chapter breaks was 0.16. Thus, generally, the coherence between paragraphs at chapter breaks was significantly lower than the overall coherence of the text (p<.001).   By choosing all coherence breaks that have a cosine two standard deviations below the mean (i.e., < 0.15), the method identified nine out of the 18 breaks that were actual breaks between chapters. However, at the same time, it detected 31 other coherence breaks in the text that had a cosine of two standard deviations below the mean. Thus, although the method correctly identified half the breaks, it had false alarms on a number of other places in the text that are not chapter breaks. With a higher cutoff value, the hit rate increases, but the false alarm rate increases almost linearly with it.

An examination of the text indicates why the method is able to make the predictions, but sometimes fails. Places where the method predicted low coherence that were not chapter breaks (false alarms) tended to be places where the author had listed several (typically 5 to 10) short bullet points, questions, or summaries within the text. Since each of these points was represented as a separate paragraph but was also fairly short compared to the average length of paragraphs, they all tended to have many fewer terms that co-occur or are semantically related between them. Thus, they had a much lower average cosine.

For misses, where the coherence between paragraphs at chapter breaks was predicted to be high, the author had typically written paragraphs that linked the two chapters. For example, the two chapter breaks that had the highest predicted coherence were between the chapters on sensation and perception (cosine=0.25) and between the chapters on psychological disorders and therapy (cosine=0.29). In both

cases, the author wrote several paragraphs that identified how the two chapters were related. In this way, although there was a physical chapter break, the text actually maintained a continuous, coherent flow of ideas.

Overall, the results indicate that the method was able to identify breaks in topics. However, the breaks must be signaled by changes in the topic in the text by the author. A well written article or book may provide coherence even at these nominal breaks, making it much more difficult to identify them. Thus, topic changes are not always marked by a lack of coherence. In addition, an author may deliberately make a series of disconnected points, such as in a summary, which may not be a break in the discourse structure. Despite this variability, the method appears to be successful for discourse segmentation, especially with texts where topic breaks are more pronounced, for example, dividing the text of a newswire into distinct news articles.

Semantic distance in texts

It is also interesting to consider how the topical focus or center of meaning changes over much longer stretches of ostensibly coherent discourse, such as a textbook on a single subject. LSA yields the same kind of representation—a vector representing the average of the words it contains—for a text segment of any length. Thus one can choose any granularity one wishes for such an analysis.[4]   From the coherence analyses, we have seen that at a local level, such as from sentence to sentence or paragraph to paragraph, texts tend to be fairly coherent. Yet, over the course of a text, the topic will shift, so that any unit of text should likely be less coherent with units of text that are physically farther away. Using the same analysis of the 4903 paragraphs in the Myers introductory psychology textbook, we computed the average cosine between any two paragraphs as a function of their physical distance in the text.   For each distance d (1...4902) between paragraphs (for adjacent paragraphs, $d = 1$) there are a total of $n-d$ unique pairs of paragraphs. Figure 3 displays the average cosine between paragraphs as a function of distance. The function is remarkable in that it remains elevated over surprisingly long distances. The irregularities and the rise at the longest inter-paragraph distances are probably due, at least in part, to edge effects (the samples for paragraph pairs at the largest distances can include only material from the beginning and end of the book) and the fact that introductory and final chapters tend to share general summary discourse. It is also interesting to note that the asymptote for this graph is around 350 paragraphs. For reference, for the average paragraph in the text, 350 intervening paragraphs is approximately one and a half chapters away. Thus, on average, there

is some slight semantic similarity between any paragraph and paragraphs that are well over a chapter away from each other in the text.
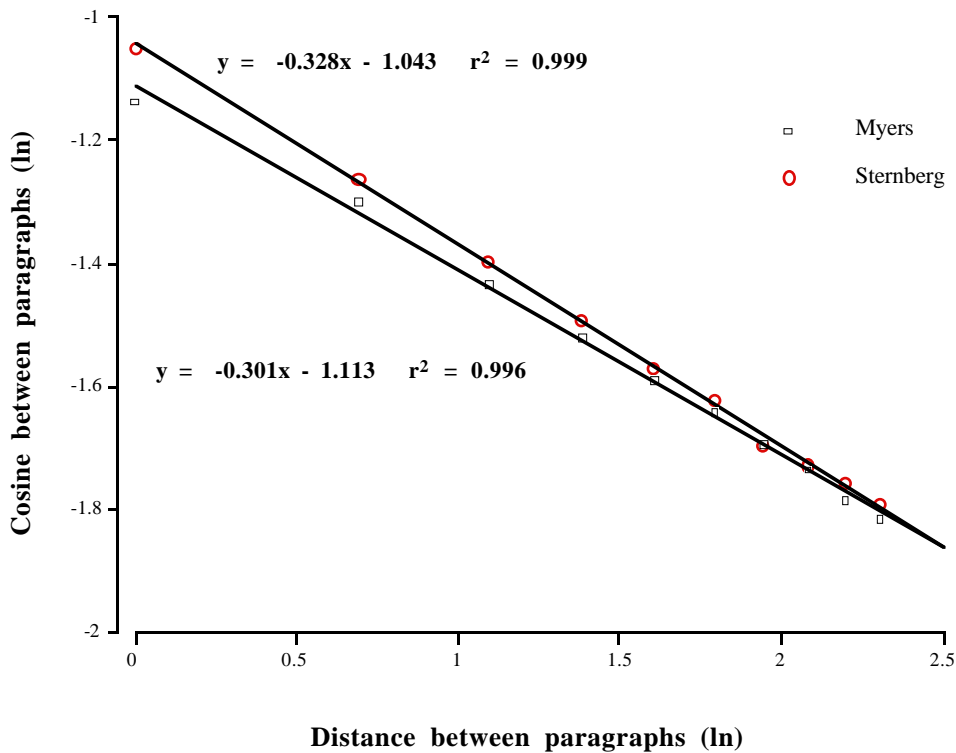

Distance between paragraphs (ln)
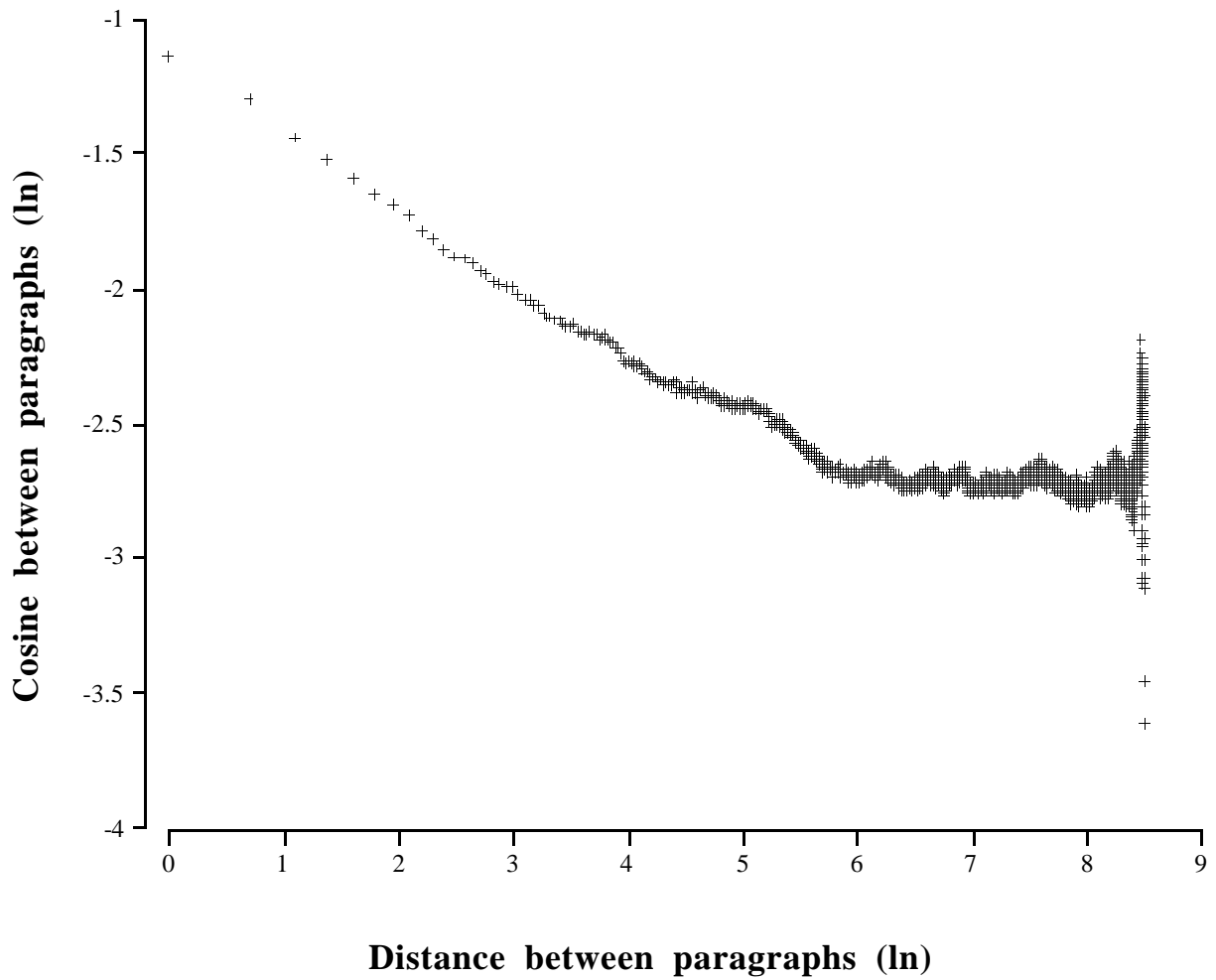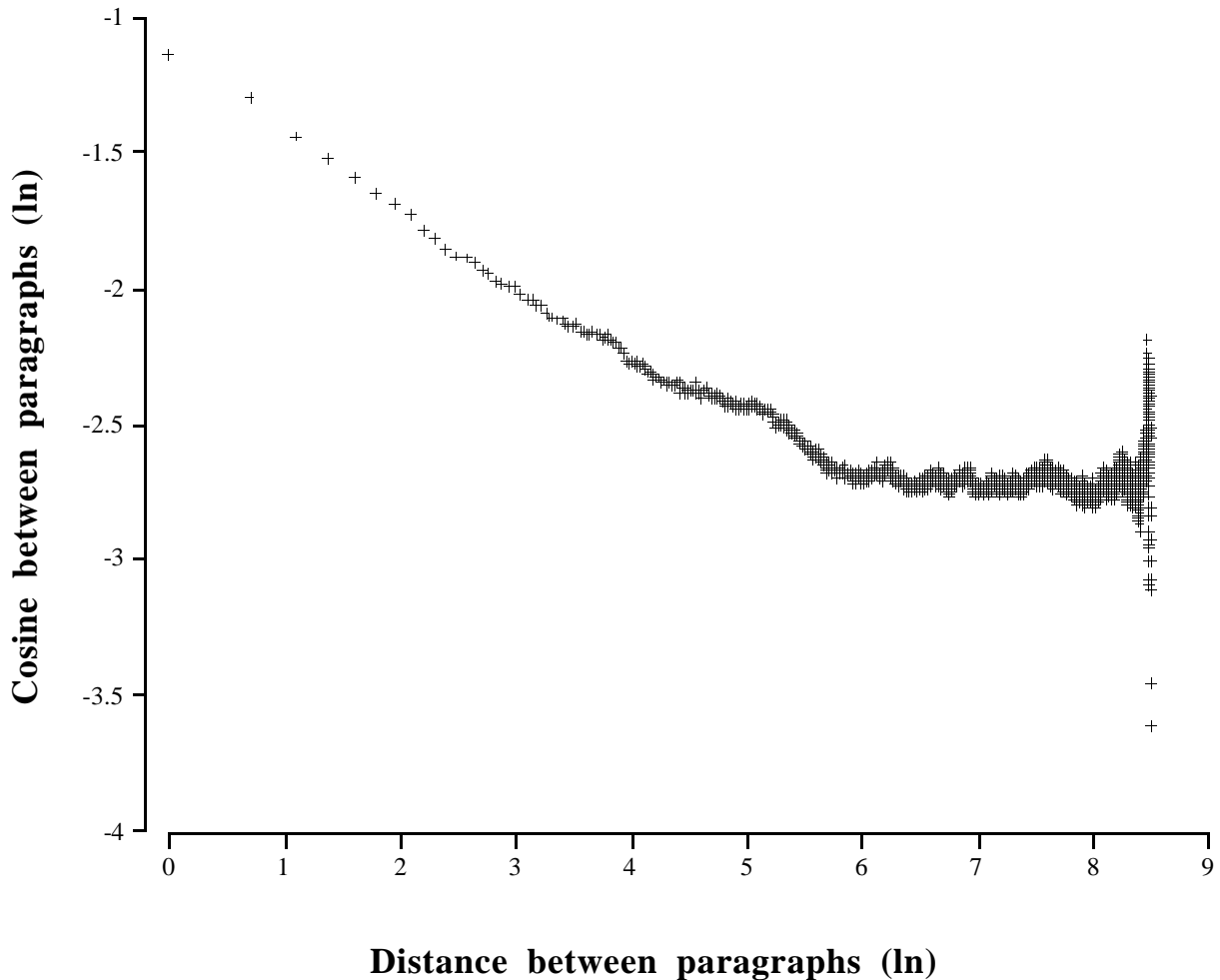
<u>Figure 3</u>. Log cosine as a function of log distance between paragraphs for the Myers textbook.

The exceedingly regular initial portion of the function warrants further discussion. In Figure 4 we show the inter-paragraph distances <u>1-10</u> versus inter-paragraph cosines plotted on log-log coordinates for the Myers text and for a second introductory psychology textbook (Sternberg, 1995). The Sternberg LSA analysis was based on the text's 3911 paragraphs by 15549 unique words, using 300 dimensions. The straight line fits corresponding to the power functions are virtually perfect. The smoothness of the functions are attributable to the very large number of observations at each point. In addition, the parameters of the fitted functions are of interest. It depends on the average change in vector position in the LSA semantic space between one paragraph and the next, on the dimensionality of the subspace in

which the centroids of sets of <u>k</u> successive paragraphs are embedded, and on the trajectories of paths taken through the space.   The fits for the Myers and Sternberg texts are highly similar. This indicates that, on average, the amount of change in semantic information from one paragraph to the next is almost equivalent between the two texts. It is interesting to notice, though, that the average cosine for the Sternberg text for adjacent paragraphs is slightly greater than that for the Myers text (Cosine Sternberg=0.35, Cosine Myers=0.32), although the difference becomes much smaller at greater distances between paragraphs. This seems to indicate that, although the Sternberg text is slightly less locally coherent than the Myers text, they cover approximately equivalent amounts of information over larger numbers of paragraphs.

Figure 4. Log cosine as a function of log distance between paragraphs for the first 10 paragraphs in the Myers and Sternberg textbooks.

We currently lack a rigorous theory with which to model this process in more detail, but we conjecture that the flattening of the curve (decrease in the exponent) with longer distances between paragraphs reflects movement in higher dimensionalities, that is, over a greater number of abstract features. The idea is that there are more <u>ways</u> in which meaning changes over large distances than over

small distances. This makes good sense, of course, and the fact that LSA captures the phenomenon may suggest other useful applications of this sort of analysis, such as to characterize the structure and information content of large bodies of discourse. For example, texts with a greater slope would indicate less overall coherence between paragraphs indicating that the text covers a more diverse set of topics.

Discussion

The results of the analyses of the Britton and Gulgoz and McNamara et al. texts indicate that LSA can provide an accurate model of coherence of the texts. In addition, these coherence predictions correspond well to the comprehension of low-knowledge readers of those texts. It is important then to understand what aspects of coherence are captured by the LSA analysis that permit these predictions.

<u>What discourse features are used for computing coherence?</u>

An LSA coherence analysis determines coherence entirely based on the derived semantic relatedness of one text unit to the next. Thus, it is making a coherence judgment based on the extent to which two text units are discussing a semantically related topic or have words which directly overlap. The method, though, is not performing any syntactic processing or parsing of the text.   Within any unit of text, it does not take into account the order of the words.   It further does not take into account some of the features typically analyzed in cohesion (e.g., Halliday & Hasan, 1976), such as pronominal reference, substitution or ellipsis. It also ignores linking clauses and signals (e.g., <u>therefore</u>, <u>since</u>) and does not detect originality. Repeating the same sentence would result in a text that would be judged highly coherent (although, to a human not very interesting.) Nevertheless, this is the same prediction that would be made by propositional modeling of such a text. Therefore, although readers need coherence in a text, for much learning to occur there must be at least some change in the semantic content across the text sections.

Despite not taking into account syntactic features, the analysis of the semantic features provide considerable strength in prediction. LSA captures Halliday and Hasan's notion of cohesion through lexical reiteration, synonymy and hyponymy. In addition, it goes beyond this level in determining coherence based on semantic relatedness due to terms tending to occur in similar contexts. Thus, a sentence using the term <u>birth</u> will tend to be judged as coherent with a sentence using a term such as <u>baby</u>. Even when a sentence uses syntactic signals of coherence, it is likely that there will be semantic signals of coherence in the sentences as well. With this method of analyzing the semantic features, LSA's coherence predictions are similar to those made by propositionally modeling a text.   The primary linking in a

propositional model is based on argument overlap, but unlike LSA, it is not capable of providing links based on overall semantic relatedness. It should be noted however, that syntactic features can be encoded into a propositional model (e.g., Kintsch, 1992) as well as linking of other propositional elements such as pronouns, which would not be automatically performed in an LSA analysis, at least as currently constituted.

What is the appropriate unit of analysis?

One difference between a propositional and an LSA analysis of coherence is the unit of analysis. In the previously described modeling of readers' comprehension, the coherence was computed between sentences. This is a larger unit than most propositions. In models of text comprehension, such as the Construction-Integration (CI) model (Kintsch, 1988, In press), text processing does not always occur in a manner such that an entire sentence is processed into working memory in one cycle. Instead, for sentences with more propositions than can be held in working memory, the propositions within a sentence are formed separately, then linked over several cycles. Thus, in the CI model, coherence also involves linking propositions within sentences and not just between sentences. While it would be possible to perform similar analyses with LSA at a clause level, which would be closer to the size of the CI model's propositions, LSA's coherence predictions may be more effective at the sentence level. Some clauses may be very short, containing little or no semantic information that is relevant to the topic and thus may not provide any of the semantic information needed for LSA to make accurate predictions. Therefore, there would tend to be more variability in the judged coherence of a text analyzed at the clause level. Moreover, since coherence breaks tend to occur more frequently between sentence breaks than within sentences, analyzing coherence at the sentence level seems appropriate. In addition, using sentences finesses the difficulties of actually parsing text into sub-sentence phrases.

By comparing individual sentences, LSA is capturing primarily effects of local coherence. However, LSA coherence measures can also be used with much larger units of analysis, such as paragraphs or multiple paragraph sections of text. By representing two paragraphs as vectors, the cosine indicates the degree to which they are on the same topic. This approach permits a characterization of the macrocoherence between two paragraphs. An alternate approach to using this method would be to use a sliding window in which comparisons are made between a vector composed up of the first N sentences of a text and a vector of the next N

sentences, then moving ahead one sentence and making another comparison.   The advantage of the sliding window is that it tends to smooth the coherence predictions, although a large drop in coherence still would indicate that there is a marked change in the general semantic content of the text at a particular point. A second advantage of the sliding window technique is that it captures, to some degree, the fact that some propositions are held over in working memory for several sentences.

Is LSA an expert or novice model of text knowledge?

One question raised about the representation of textual information that is generated by LSA is whether it is closer to that of a novice or an expert of a domain. The McNamara et al. study found an interaction between the reader's knowledge and the coherence of the text.   The LSA predictions matched best the comprehension scores of the novice reader. Based on the initial encyclopedia articles that were used to develop the LSA space, the vectors for terms such as birth and baby have a high cosine with each other. These types of commonly used terms are likely to occur frequently enough across a large number of similar articles that they would tend to be represented as being related in LSA. This would also be consistent with the general knowledge of a low-knowledge reader who should still be able to make an inference between two sentences that uses those two terms, but would perhaps not be able to infer connections between less familiar technical terms.   Thus, LSA's representation, based on a comparatively small text corpus during its learning phase, may be more similar to that of a novice in the domain. This would also be consistent with findings that LSA best approximates a novice model discussed in other papers in this issue (see Wolfe, Schriener, Rehder, Laham, Foltz, Kintsch & Landauer, this issue)

In addition, however, the LSA representation depends on the particular texts upon which LSA is trained. For example, in the LSA analysis of the heart encyclopedia articles, the cosine between the vectors for congenital and birth was also fairly high, indicating that the model would predict a reader's tendency to find a sentence with the word congenital coherent with a sentence with the word birth. One would not expect low-knowledge readers to be able to make this inference, since they likely would not have encountered the term congenital enough times to associate it with birth. It is possible that training LSA on highly technical texts would result in a much more elaborated representation of the semantics of the topic and would better capture the effects of coherence for expert readers of the text.

Additional applications for coherence analysis

The accuracy of the coherence predictions made by LSA suggests other areas to which LSA can be applied. Since there is a strong relationship between coreference and causal coherence, (e.g., Trabasso, Secco & van den Broek, 1984; Fletcher et al. 1995), LSA could be used to predict causal chains in text. Although two sentences are not adjacent, if they have a high cosine between them, they may be causally linked. By computing the cosines of all possible pairs and retaining those above a certain threshold, the method could locate chains of related events mentioned in sentences that occur across the text. Preliminary research using LSA analyses of history texts shows that in texts that have multiple causal threads, the method is able to identify causally related sentences, even if they occur in texts written by different authors. (see Foltz, 1996; Foltz, Britt & Perfetti, 1996 for related research on history texts).

Another application of the method is as a writing critic. Britton and Gulgoz (1988) revised their text based on a propositional analysis that repaired breaks in argument overlap. LSA could automatically compute the sentence-to-sentence coherence and then mark places in the text where it predicts that the coherence is lower than average. These places may indicate areas of the text in which readers, particularly low-knowledge readers, may have more difficulties. A writer could then use this information to decide whether sentences in those places in the text should be revised.   In addition, such a critic could provide some overall measure of the text's global coherence. However, the main studies described in this paper involve the re-analysis of texts in which the coherence was deliberately varied by manipulating linguistic features of the text. For determining the overall coherence of any single text, it may be more difficult to set a criterion for the coherence value, since it may vary with a variety of factors such as the choice of the original texts used for the LSA scaling, the size of the unit chosen for comparison, and the style and purpose of the author's writing. For these reasons, this approach may be more appropriate for comparisons of different versions of texts as well as for critiquing texts.

Is LSA a model or method of text coherence?

As pointed out in the paper in this issue by Landauer et al., LSA can be viewed as both a model of the underlying representation of knowledge and its acquisition or as a practical method for estimating aspects of similarities in meaning. As a model of knowledge, the coherence predictions are similar to those of propositional modeling (e.g., Kintsch, 1988, In press).   One can think of our experience of coherence as being an effect of computing semantic relationships

between pieces of textual information. These semantic relationships are based on our exposure to this information in the past. Through our experiences of words co-occurring, or occurring in similar contexts, we develop knowledge structures which capture these relationships. The LSA coherence predictions model both the effects of coreference and also the semantic relatedness as measured by the analysis of contextual occurrences in the past. In this case, the past experiences are based on a set of initial training texts.

Although LSA lacks certain components of a cognitive architecture, such as word order, syntax, or morphology, the representation it produces is highly similar to that of humans (see Landauer & Dumais, 1997). As a model of text comprehension, it approximates some of the same features found in propositional models of text comprehension. For analyzing coherence, LSA links textual information similarly to the way the Construction-Integration (CI) model links propositions through argument overlap, elaboration, and inferencing. In both cases, linking is based on using the same terms, or on semantic relatedness between terms that would be consistent with simple bridging inferences made by the reader (e.g., baby, birth). In LSA, the strength of connections between textual items is also based on degree of semantic similarity, while also taking into account the information value of the textual content. This is similar to the use in the CI model of connection strengths between propositions[5]. In the same manner as in the CI model, the meaning of a concept is therefore situation specific, depending on its relationship to the other terms around it. Thus, LSA's induction of meaning similarities produces a representation that is similar to other modeling approaches to text comprehension.

As a practical method, LSA produces a useful representation for text research. The ability to measure text-to-text relationships permits predictions of human judgments of similarity. These judgments are based not only direct term co-occurrence but also a deeper measure of inferred semantic relationships based on past contextual experiences. The results from the analyses described in this paper indicate that LSA captures to a large degree the variable coherence of texts which correlate highly with readers' actual comprehension of the texts. Since the method is automatic, it permits rapid analyses of texts, thereby avoiding some of the effort involved in performing propositional analyses and allowing analyses that could not have been performed previously.

In summary, LSA can be conceived as being both a model of the representation of knowledge and a practical method. LSA provides a powerful

measure of the representation of meaning derived from a text, and this representation corresponds well to that of a reader. It further permits a characterization of how the semantic content changes over a text. This provides a measure of the text's coherence and can be used to predict measures of a reader's comprehension.

References

Britton, B. K., & Gulgoz, S. (1991). Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. Journal of Educational Psychology, 83, 329-345.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391-407.

Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology, 32, 221-233.

Fletcher, C. R., Chrysler, S. T., van den Broek, P., Deaton, J. A., & Bloom, C. P. (1995). The role of co-occurrence, coreference, and causality in coherence of conjoined sentences. In R. F. Lorch & E. J. O'Brien (Eds.), Sources of coherence in reading. Hillsdale, NJ: Erlbaum.

Foltz, P. W. (1996). Latent Semantic Analysis for text-based research. Behavior Research Methods, Instruments and Computers, 28(2), 197-202.

Foltz, P. W., Britt, M. A., & Perfetti, C. A. (1996). Reasoning from multiple texts: An automatic analysis of readers' situation models.   In G. W. Cottrell (Ed.) Proceedings of the 18th Annual Cognitive Science Conference (pp. 110-115). Hillsdale, NJ: Erlbaum.

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

Klare, G. R. (1963). The measurement of readability, Ames: Iowa State University Press.

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

Hearst, M. A., & Plaut, C. (1993). Subtopic Structuring for Full-Length Document Access. In Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, 59-68.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. Psychological Review, 2, 95, 163-182.

Kintsch, W. (1992). A cognitive architecture for comprehension. In H. L Pick, Jr., P. van den Broek, & D. C. Knoll (Eds.) Cognition: Conceptual and methodological issues (pp. 143-164). Washington, DC: American Psychological Association.

Kintsch, W. (in press). Comprehension: A paradigm for cognition. New York: Cambridge University Press.

Kintsch, W. & van Dijk, T. (1978). Toward a model of text comprehension and production. Psychological Review, 85, 363-394.

Kintsch, W. & Keenan, J. M. (1973). Reading rate and retention as a function of the number of the propositions in the base structure of sentences. Cognitive Psychology, 5, 257-274.

Kintsch, W., & Vipond, D. (1979). Reading comprehension and readability in educational practice and psychological theory. In L. G. Nilsson (Eds.), Perspectives on Memory Research. Hillsdale, NJ: Erlbaum.

Landauer, T. K., & Dumais, S. T. (1997) A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (this issue). An introduction to Latent Semantic Analysis. Discourse Processes.

Lorch, R. F. & O'Brien, E. J. (Eds.). (1995). Sources of coherence in reading. Hillsdale, NJ: Erlbaum.

McNamara, D. Kintsch, E., Butler Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. Cognition and Instruction, 14(1), 1-43.

Miller, J. R., & Kintsch, W. (1980). Readability and recall of short prose passages: A theoretical analysis. Journal of Experimental Psychology: Human Learning and Memory, 6, 335-454.

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.

Salton, G., & McGill, M. J. (1983). Introduction to modern information retrieval. NY: McGraw-Hill.

Trabasso, T., Secco, T., & van den Broek, P. (1984). Causal cohesion and story coherence. In H. Mandl, N. Stein, & T. Trabasso (Eds.). Learning and comprehension of text. Hillsdale, NJ: Erlbaum.

Turner, A. A. (1987). The propositional analysis system (Tech. Rep. No. 87-2). Boulder: University of Colorado, Institute of Cognitive Science.

van Dijk, T. A., & Kintsch, W. (1983). Strategies of discourse comprehension. New York: Academic Press.

Wolfe, M. B., Schreiner, M. E., Rehder, R., Laham, D., Foltz, P. W., Kintsch, W., & Landauer, T. K. (this issue). Learning from text: Matching readers and text by Latent Semantic Analysis. Discourse Processes.

Author Note

Footnote

---

[1]It should be noted that because of the term weighting function used, these high frequency terms are typically given very little weight and therefore contribute very little to the analysis. They therefore could be included in the analyses without having much effect.

[2] Because cosines are closely related to correlations (only the normalization is different), it is appropriate to apply Fisher's r-to-z transforms on the cosines before the ANOVA. However, for both the Britton and McNamara analyses, the r-to-z transform did not change the results in a meaningful way, so the raw cosines were used for both analyses.

[3] Although direct term overlap could be used without applying term weighting, term weighting helps account for the actual information value of that term within the text. This approach is commonly used in information retrieval in which the overlap between terms in a query and terms used in documents is weighted based on some transformation of the word frequency.

[4] Because of the nested property of words, sentences and larger segments and the linear vector combination in LSA, results at any higher level of granularity are equivalent to results at every lower level averaged over smaller, naturally varying, sample sizes. The assortment of the words in paragraphs into their contained sentences would cause imputed functions on average distances between sentences or words (as compared to directly calculated ones) to differ only by implicitly weighting words in short sentences less heavily than those in longer sentences.

[5] It should be noted that LSA's similarity ratings are almost all positive, while the CI model can have inhibitory connections between nodes. However, this could be adjusted as a matter of scale, in which low cosines below some threshold could represented as inhibitory connections.