LSA as a Measure of Coherence in Second Language Natural Discourse

Scott A. Crossley  
(ph544@msstate.edu)  
Department of English, Mississippi State University  
Mississippi State University, MS 39762

Philip M. McCarthy  
(ppmmccrth@memphis.edu)  
Department of English, University of Memphis  
Memphis, TN 38152

Thomas Salsbury  
(tsalbury@wsu.edu)  
Department of Education, Washington State University  
Pullman, WA 99164

Danielle S. McNamara  
(dmcmamara@mail.psyc.memphis.edu)  
Department of Psychology, University of Memphis  
Memphis, TN 38152

Abstract

This study explores Latent Semantic Analysis as a model of coherence relations in spoken discourse. Two studies were conducted on language data from six adult learners of English observed longitudinally for one year. The studies investigated whether LSA values increase as a function of time spent learning English and if there were links between negotiations for meaning and LSA values. Results show that LSA values increase significantly and that negotiations for meaning decrease significantly over time. Negotiations for meaning are also negatively correlated to LSA values. These findings provide evidence of co-referential semantic development in L2 learners. The findings also demonstrate that a lack of semantic similarity in natural spoken discourse is sufficient to trigger negotiations for meaning.

Keywords: Coherence; Cohesion; Corpus Linguistics; Second Language Processing; Latent Semantic Analysis

Introduction

Latent Semantic Analysis (LSA) is a theory and method for extracting and representing the meaning of words using statistical computations applied to a large corpus of text (Landauer & Dumais, 1997). LSA has the ability to model human conceptual knowledge in word sorting evaluations (Landauer & Dumais, 1997), relatedness judgments (Landauer et al., 1998) word synonymy judgments, and vocabulary learning (Landauer & Dumais, 1997). LSA is also an important model of coherence relations in both text (Foltz, Kintsch, & Landauer, 1998) and in some categories of spoken discourse (Elvevag et al., 2006; Gorman et al., 2003). However, the potential for LSA to evaluate coherence patterns other than topic shifts in extended, natural spoken discourse has not been examined. Thus, in this analysis, we are interested in demonstrating how LSA functions as a model of semantic co-referentiality in natural speech. To analyze coherence in natural speech, we use a longitudinal corpus of speech data taken from a group of English as a second language (L2) learners. The theoretical assumption is that as L2 learners progress in their language learning (i.e. interlanguage), their discourse will become more coherent and LSA can be used to analyze and monitor this development.

We examine this hypothesis by first investigating language data from English L2 learners to establish whether LSA values increase as a function of time spent studying English, providing evidence that language becomes more semantically similar over time. Second, we hand code the language data of the L2 learners for coherence breaks. We then analyze the frequency of these coherence breaks to determine if the breaks decrease as a function of time; thus supporting the notion that L2 language data becomes more coherent. We also compare the frequency of these coherence breaks to the LSA values for the language data to examine the relationship between coherence breaks and LSA values.

Such an analysis is relevant to cognitive science for two reasons. First, it examines whether LSA can be used to evaluate the coherence of natural spoken discourse based on semantic co-referentiality. Second, it could further our understanding of L2 acquisition by examining L2 semantic growth and provide an automated means to measure L2 semantic development.

Coherence

Coherence is generally associated with the interpretability of discourse (Anderson, 1995). Graesser, McNamara, and Louwerse (2003) further specify that coherence refers to the representational relationships of a text in the mind of a reader or listener, whereas cohesion refers to the cues in the text that help the reader to build a coherent representation (Foltz, 2007).

There are many types of cohesive devices available to a proficient speaker. These include ellipsis, conjunction, anaphora (Halliday & Hasan’s, 1976), causal verbs and particles (McNamara, Cai, & Louwerse, 2007) and semantic similarity (Foltz, 2007). Cohesive devices are important in text for connecting ideas with topics (Graesser et al., 2004) and in speech for allowing indications of coherence in a message, and providing interlocutors with a means to interpret messages (Tanskanen, 2006). Cohesive devices also allow interlocutors to make links between pieces of discourse and transition information from one section of discourse to another. Gaps in cohesion force participants to either make inferences to complete the gaps (McNamara et al., 1996) or, if inferences are not possible, to negotiate meaning.

Coherence Breaks

In both first language (L1) studies and L2 studies, coherence breaks in natural speech often lead to negotiations for
meaning to repair misunderstandings (Bazzanella & Damiano, 1999). If successful, the negotiation re-establishes shared coherence (Bazzanella & Damiano, 1999). In both L1 and L2 discourse, negotiations for meaning are signaled by requests for elaborations and clarifications asking for coherent input. However, in L1 discourse, coherence breaks are usually the result of topic shifts (Bublitz & Lenk, 1999).

Unlike L1 discourse, negotiations in L2 discourse mostly occur when L1 speakers or L2 speakers signal that they do not understand an utterance based on semantic, phonological, or morphosyntactic features. These features are considered the triggers for the negotiations for meanings. The triggers and corresponding signals lead to an opportunity to negotiate for meaning and the resulting interaction allows participants to understand and reuse the language that caused the initial loss of meaning. In L2 studies, negotiations for meaning assist L2 learners in recognizing and adjusting incomprehensible input. These negotiations for meanings include clarification requests, comprehension checks, and confirmation checks (Gass & Varonis, 1994). The negotiation and its interaction can lead to the introduction of new and varied linguistic input as well as new and varied output (Swain, 1985). According to Swain, negotiated meanings push L2 learners to modify their own output and make it more comprehensible and natural.

**Latent Semantic Analysis**

Latent Semantic Analysis (LSA) determines the similarity of passage meaning through the analysis of large corpora. However, LSA does not depend on word frequency, syntax, or pragmatics to measure semantic similarity between text samples. To determine the similarity of passage meaning, LSA employs a mathematical formula known as singular value decomposition (SVD), which reduces thousands of dimensions and relationships between words to hundreds of dimensions and relationships (Landauer, 2007). SVD also reduces computational complexity and smooths data by filling in missing data points. In this way, SVD is similar to a factor analysis. The dimensions used in LSA simply represent how often a word or words occur within a document (defined at the word, sentence, paragraph, or text level) and these documents become weighted vectors. Text selections are typically matched by comparing the cosine between two sets of vectors (the words, or bag of words included within the document) and receive values between -1 and 1. This cosine relates to the similarity or dissimilarity between documents (Landauer & Dumais, 1997).

The theory behind LSA is that the similarity of words is based on topical and referential meaning and that these meanings come from a large domain of knowledge where there are many direct and indirect relationships. Thus, if two words appear in the same context, and every other word in that context appears in many other contexts without them, the two will acquire semantic similarity to each other but not to the rest (Landauer & Dumais, 1997; Landauer, 2007). In this way, connections between related words develop. As an example, all component features related to legs, tails, ears, and fur are related to each other not only because of the occasions when they occur together, but, importantly, as the indirect result of the occasions when they occur with other elements (such as animals). In essence, LSA measures how likely two words will appear in similar discourse settings and then relates this inversely to their semantic distance, thus making word associations based on semantic similarity. Therefore, how often words co-occur relates to their psychological similarity. This co-occurrence also allows for the plotting of their semantic similarity and output distance (Landauer & Dumais, 1997).

As a model of discourse coherence, LSA tracks the overlap and transitions of meaning as they move across discourse by computing the semantic similarity of text segments. As discourse moves from one utterance to another, an LSA cosine value can be assigned to each section of text. Discourse that is highly coherent is represented by higher values as compared to less coherent discourse, which has lower values.

LSA has been used to effectively measure semantic coreferentiality coherence in text and predict the effect of text coherence on comprehension. In a replication of two classic coherence studies (Britton & Gulgoz, 1991; McNamara et al., 1996), Foltz, W. Kintseh, & Landauer (1998) demonstrated that LSA was able to predict the readability and cohesiveness of texts that had been modified to make them more or less coherent. In replicating the Britton and Gulgoz (1991) study, LSA cosines were compared to three comprehension measures (number of propositions, efficiency, and inference multiple choice) and demonstrated significant correlations with the measures. In analyzing the coherence of the texts used in the McNamara et al. (1996) study, LSA was able to predict which texts had low, medium, and high coherence. Foltz et al. (1998) contended that these findings support the notion that LSA provides an accurate measure of text comprehensibility by measuring the semantic relatedness of adjacent text and topics (see also, McNamara et al., 2007).

While LSA has been applied to models of spoken discourse such as register classification (Louwerse et al., 2004), dialog act classification (Serafin & di Eugenio, 2004), and information retrieval (Kurimo, 2000), we are interested in LSA as a model of spoken coherence. In measuring the coherence of spoken discourse, it is argued that LSA allows both the measurement of internal coherence of an utterance and how coherent that utterance is to another person (Foltz, 2007). However, in practice, past studies have mostly concentrated on how LSA can measure topic shifts and not focused on how LSA relates to coherence breaks that lead to negotiations for meaning. For instance, Gorman et al. (2003) found promising results when comparing LSA values with topic shifts in aircraft simulation missions. Their conclusions, while exploratory, indicated that topic shifting as measured by LSA values appeared more common with low performing missions, thus suggesting lower coherence. In a second study, looking at the
coherence patterns of schizophrenic patients, LSA values were significant predictors of topic shifting (Elvevag et al., 2006). However, neither of these studies looked at extended discourse in a natural spoken environment. The Gorman et al. (2003) study looked at a specific, controlled register (flight simulations) and the Elvevag et al. (2006) study analyzed schizophrenic discourse using either word lists or moving windows of two to eight word lengths. Neither of these studies examined extended natural discourse and how spoken coherence related to semantic co-referentiality.

Study One

We collected a corpus of L2 spoken dialogs to investigate whether or not LSA values increase as a function of time spent learning English. This corpus was longitudinal and was composed of spoken language samples taken from six L2 learners over the course of the year. The L2 learners were all enrolled in an intensive English program at a large university in the United States. They were interviewed approximately every 2 weeks (not including program and university breaks) over the year period by at least four different interviewers (about one interviewer every four months). Learners’ proficiency levels were tested upon arrival to the program, and all participants in the study tested into the lowest proficiency level (i.e., Level 1 of a 6-level program). The learners ranged in age from 18 to 29 years old and were from diverse language backgrounds. Language data was collected in naturalistic settings; that is, while interviewers came to the sessions prepared with various topics and tasks (picture descriptions, questions) from which the learners could choose, the sessions were generally characterized by spontaneous, naturally occurring discourse. The sessions were tape recorded and later transcribed. The transcripts were modified by excising interjections and non-words.

Coh-Metrix

To collect LSA measurements, each text file was analyzed using the computational tool Coh-Metrix, which measures cohesion and text difficulty at various levels of language, discourse, and conceptual analysis (Graesser et al., 2004; McNamara et al., 2006). LSA values from Coh-Metrix were used to measure semantic development because they operate at the text level and not the word level. Thus, they provide measurements that examine similarity in meaning and conceptual relatedness between text segments. In Coh-Metrix, sentences, paragraphs, and texts are measured as weighted vectors and LSA values taken from the TASA college level space are computed as geometric cosines between these vectors. Because the data used in this study was based on spoken utterances and not written text, only LSA paragraph to paragraph values were analyzed. This is because sentence punctuation for the spoken utterances would be artificial and many spoken utterances are too short to provide proper semantic coverage. However, complete propositions were easily broken up based on conversational turns. These propositions are analyzed at the paragraph level by Coh-Metrix.

Analysis

We conducted a repeated measures Analysis of Variance (ANOVA) using the LSA results from Coh-Metrix to track the development of LSA values over the increasing temporal intervals. The ANOVA was used to test the assumption that as the learners study English, semantic co-referentiality develops as well. Such a finding would support the hypothesis that L2 speech patterns become more coherent. Because participants did not share all the same temporal data points, the ANOVA examined development on a trimester basis. This allowed for breaks in the data related to winter and spring breaks to be considered as well as missing data points resulting from participant absences. Because data was available for the second and fourth week and the 50th and 52nd weeks for all six learners, that data was included. These data points were analyzed with data from the 16th week and the 32nd week as well. We supplemented the ANOVA with a post-hoc test of within-subjects contrasts to examine where in the temporal progression significant differences in output could be identified.

Results

ANOVA results comparing the first, second, and third trimester of learning to the LSA values show that similarity between utterances as measured by LSA significantly increased as time increased, $F(5, 25) = 3.95, p < .01$ (see Table 1). A test of within-subjects contrasts demonstrated that the LSA values from the last meeting on the 52nd week were significantly higher than the first meeting on the second week, $F(5, 25) = 10.878, p = .02$. Additionally, significant increases in LSA values were found between the second week and the 16th week, $F(5, 25) = 36.68, p < .01$, and the second week and the 50th week, $F(5, 25) = 20.29, p < .01$. These findings provide evidence that LSA semantic similarity values increase with time spent studying English and supports the hypothesis that L2 learners’ speech exhibits more semantic co-referentiality as a function of time spent learning a language.

Study Two

Our second study assessed whether there was a link between negotiations for meaning and LSA values and whether or not there was a decrease in L2 negotiations for meaning as a function of time spent studying English. To evaluate this relationship, we examined the frequency of L2 utterances in each transcript that triggered negotiations for meaning. The transcripts used for this study were taken from Study one; however, unlike Study one where missing data points excluded the analysis of some transcripts, all available transcripts were examined (N=99). The frequency of the negotiations for meaning found in each transcript was then compared against the LSA values for the transcript to establish whether LSA can assess semantic coherence in
Table 1: Mean and Standard Deviations (SD) for LSA Values and Negotiation for Meaning (NfM) Counts

<table>
<thead>
<tr>
<th>Week</th>
<th>Mean LSA Value</th>
<th>SD LSA Value</th>
<th>Mean NfM</th>
<th>SD NfM</th>
<th>Week Comparison</th>
<th>F</th>
<th>p</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.16</td>
<td>0.01</td>
<td>0.241</td>
<td>0.015</td>
<td>LSA</td>
<td>4.85</td>
<td>0.008</td>
<td>6.52</td>
<td>0.05</td>
</tr>
<tr>
<td>4</td>
<td>0.20</td>
<td>0.04</td>
<td>0.012</td>
<td>0.004</td>
<td>2 to 4</td>
<td>36.68</td>
<td>&lt;.01</td>
<td>10.50</td>
<td>0.02</td>
</tr>
<tr>
<td>16</td>
<td>0.23</td>
<td>0.04</td>
<td>0.010</td>
<td>0.004</td>
<td>2 to 16</td>
<td>3.99</td>
<td>0.01</td>
<td>12.17</td>
<td>0.02</td>
</tr>
<tr>
<td>32</td>
<td>0.30</td>
<td>0.17</td>
<td>0.006</td>
<td>0.004</td>
<td>2 to 32</td>
<td>20.29</td>
<td>&lt;.01</td>
<td>12.50</td>
<td>0.02</td>
</tr>
<tr>
<td>50</td>
<td>0.26</td>
<td>0.06</td>
<td>0.005</td>
<td>0.003</td>
<td>2 to 50</td>
<td>10.88</td>
<td>0.02</td>
<td>12.02</td>
<td>0.02</td>
</tr>
<tr>
<td>52</td>
<td>0.32</td>
<td>0.12</td>
<td>0.004</td>
<td>0.002</td>
<td>2 to 52</td>
<td>10.88</td>
<td>0.02</td>
<td>12.02</td>
<td>0.02</td>
</tr>
</tbody>
</table>

extended, natural, spoken discourse. The frequency of negotiation for meaning counts were also compared against the temporal data points to examine whether negotiations decreased as L2 learners studied English.

Negotiation for Meaning Frequency Counts

To test the hypothesis that negotiations for meanings were directly related to coherence, we conducted a frequency count for the number of negotiations for meaning in the transcripts. Two expert raters specializing in second language acquisition (SLA) analyzed each of the 99 transcripts from the L2 learner corpus. Based on Nakahama et al. (2001), the two raters located signals of negotiations for meaning (see Table 2 for examples) and then worked backwards from the signal to locate the trigger in the proceeding L2 learner utterance or utterances.

Table 2: Examples of Negotiations for Meaning

<table>
<thead>
<tr>
<th>Clarification</th>
<th>L2 Speaker</th>
<th>I feel you are not American.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confirmation</td>
<td>L2 Speaker</td>
<td>Like white clothes. I don’t know the name.</td>
</tr>
<tr>
<td>Elaboration</td>
<td>L2 Speaker</td>
<td>Something happened… About what you talked about…summer, hot, raining…?</td>
</tr>
<tr>
<td></td>
<td>Interviewer</td>
<td>Like a robe?</td>
</tr>
</tbody>
</table>

In this study, the trigger was coded as producing a negotiation for meaning only in those instances in which there was a clear loss of meaning. For our definition of loss of meaning, we borrow from Gass and Selinker’s (2001) model in which loss of meaning is triggered by interlocutors interrupting the flow of conversation because there is a lack of understanding about the conversational meaning. In this study, if multiple signals resulted from the same trigger, only the original signal was included in the frequency count. Also, if the L2 learner self-corrected without being prompted by the interlocutor (Foster & Ohta, 2005), no loss of meaning was coded. In addition, if the interlocutor provided the L2 speaker with continuers (Foster & Ohta, 2005) in an effort to show interest in what the L2 learner was saying, no loss of meaning was coded. Lastly, if the L2 learner asked for assistance in the form of a comprehension check (e.g., How do I say?) and was thus responsible for prompting the interlocutor to interrupt, no loss of meaning was coded. Because this study was only concerned with measuring semantic markers of coherence, we only included semantic triggers in the frequency count.

The inter-rater reliability of the two raters for the 99 transcripts in the corpus for coding the number of triggers, the trigger types, and the negotiation cycles was 𝑟 = .98. Those triggers agreed upon were given a raw frequency score that was normalized by the number of words in the transcript. This normalized frequency was used as an assessment of the number of negotiations for meaning per transcript. During the hand coding of the transcripts, we noticed that task seemed to play an important role in the use of negotiations for meaning. Transcripts that involved visual tasks such as picture descriptions had few negotiations for meaning because participants were able to reference the pictures. This finding suggested that two analyses should be conducted: one with all transcripts and a second with only those transcripts that did not contain picture description tasks.

Analysis

An ANOVA was conducted using the negotiation for meaning count and the temporal data points to test the assumption that coherence increased as the learner acquired a second language. This ANOVA was used to track the development of the frequency count over the increasing temporal intervals and to examine whether learners’ time spent studying English contrasted with their need to negotiate meaning. This ANOVA used the same temporal data points as in the first study and was supplemented with a post-hoc test of within-subjects contrasts in order to examine where in the temporal progression significant differences in output occurred.

We also conducted a Pearson’s correlation to calculate the relationship between LSA values and coherence in spoken discourse. An additional correlation was conducted that included only the transcripts that did not contain picture description tasks.

Results

ANOVA results comparing the first-, second-, and third-, trimesters of learning to the negotiation for meaning counts suggest that negotiation for meaning counts significantly
decreased as time increased, \( F(5, 25) = 10.69, p < .001 \) (see Table 1). Within-subjects contrasts demonstrated that the frequency of negotiations for meaning from the last meeting on the 52nd week significantly decreased from the first meeting on the second week, \( F(5, 25) = 12.02, p = .018 \). Additionally, significant decreases in negotiation for meaning counts were found between the second week and the 16th week, \( F(5, 25) = 10.50, p = .023 \), the second week and the 32nd week, \( F(5, 25) = 12.17, p = .018 \), and between the second week and the 50th week, \( F(5, 25) = 12.50, p = .017 \). These findings provide evidence that the number of negotiations for meaning decrease with time spent learning English and support the findings from the first study showing that coherence increases in L2 learners’ speech as a function of time.

When comparing LSA values to the negotiation for meaning frequency counts for all transcripts, significant negative correlations were reported (\( N=99, r = -.26, p < .01 \)). When comparing LSA values to the negotiation for meaning frequency counts for the transcripts that did not include visual stimuli, negative significant correlations were also reported (\( N = 59, r = -.30, p < .01 \)). This finding provides evidence that LSA values correlate to breaks in coherence and support the notion that LSA can be used to measure coherence in natural, spoken discourse.

**Discussion**

In this study, we compared LSA values to the frequency of negotiations for meaning in a natural language corpus of spoken L2 English. The study demonstrates a link between LSA values and coherence in spoken language. The study shows that LSA values are significantly correlated to the frequency of negotiations for meaning in the corpus. This finding provides evidence that LSA can be used as a measure of semantic coherence in extended, natural language, spoken discourse. This study also demonstrates that L2 discourse exhibits a significant decrease in frequency of negotiations for meaning as a function of time. This finding suggests that the speech of L2 learners becomes more coherent as they study a language. This finding is supported by results showing that as L2 learners’ interlanguages develop, their use of semantic co-referentiality, in the form of LSA values, increases.

In linking LSA to semantic coherence in extended, natural language discourse, the statistical analyses in this study have shown that there is a negative correlation between the LSA values of L2 learners’ spoken discourse and the frequency of negotiations for meaning. This appears to occur more often when discourse does not use visual cues. This finding demonstrates that a lack of semantic similarity in the natural spoken utterances of L2 learners can likely trigger negotiations for meaning. Thus, semantic coherence is linked to signals from the L1 speakers. When there is a loss of semantic coherence, the L1 speaker begins the task of negotiating for meaning. We argue that LSA can be used to measure the coherence of extended, natural, spoken discourse based on semantic co-referentiality because the absence of semantic co-referentiality appears to affect comprehensibility and thus coherence.

Additionally, the findings of this study demonstrate that L2 learners actively engaged in learning English develop more coherent speech patterns over the course of a year. In fact, significant differences in coherence patterns as measured by the frequency of negotiations for meaning triggered by L1 speakers were noted within four months of the study. These significant trends continued until the end of the year long study with L1/L2 transcripts demonstrating fewer negotiations for meaning as the year closed. Given that the study controlled for any familiarity effects between the L2 learners and their interviewers, this result is not the effect of interviewers becoming familiar with the speech patterns and discourse strategies of the L2 learners, but likely the result of L2 learners developing coherence patterns. Additional evidence for the development of more coherent speech is apparent in the ANOVA results, which demonstrate that LSA values of L2 learners’ natural discourse increase as a function of time over the course of the year. This finding demonstrates that as learners develop their lexicon, their adjacent utterances become more semantically similar. This semantic similarity is important in maintaining semantic links between utterances and ensuring that interlocutors are able to co-reference new information with past information.

**Conclusion**

We do not argue that semantic similarity alone is responsible for coherence. This notion is overly simplistic. While semantic similarity is an important feature that assists in creating coherent utterances in natural discourse (Bublitz & Lenk, 1999) and listeners depend on semantic co-referentiality to recognize and interpret coherent utterances (Tyler, 1994), semantic similarity is but one of many lexical cues a listener can access. Additional features such as phonology (McCarthy, 1988), prosody, other aspects of physical context, paralinguistic cues, grammar, or structural relations (Gumperz et al., 1986) are also likely responsible for creating coherent discourse. Future studies should include other cohesive devices as well as LSA values to determine to what degree cohesive devices can help predict the coherence of a natural spoken discourse. This study is an important step toward this goal as it demonstrates that semantic similarity as measured by LSA can contribute to an explanation of why negotiations for meaning are triggered. From a practical perspective, it provides evidence for the growth of semantic co-referentiality in L2 learners and a computational method to measure such development.

**Acknowledgments**

The research was supported in part by the Institute for Education Sciences (IES R305G020018-02). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IES.
References


