# Analyzing the Cohesion of English Text and Discourse with Automated Computer Tools

**Moongee Jeon**
*Konkuk University*

**Jeon, Moongee. (2014). Analyzing the cohesion of English text and discourse with automated computer tools. *Journal of Pan-Pacific Association of Applied Linguistics, 18*(2), 123-133.**

This article investigates the lexical and discourse features of English text and discourse with automated computer technologies. Specifically, this article examines the cohesion of English text and discourse with automated computer tools, Coh-Metrix and TEES. Coh-Metrix is a text analysis computer tool that can analyze English text and discourse on various linguistic and psycholinguistic measures of cohesion, readability, and language. Many researchers in the areas of applied linguistics, English education, and language psychology have now extensively used Coh-Metrix to analyze various English texts and textbooks. Recently, the author of this article has developed a new computer tool, TEES which can be applied to evaluate English texts and essays on various linguistic and psycholinguistic measures such as text readability, text cohesion, sentence structure, vocabulary, and text marker scores. Basically, TEES has been developed to evaluate English texts and essays in terms of a standardized norm. In the TEES system, a huge size of corpus was used to construct the standardized norm. This article introduces Coh-Metrix and TEES, and presents some research findings collected from Coh-Metrix studies.

**Key Words:** English text, English essay, Coh-Metrix, TEES, automated tools

## 1 Introduction

Many language psychologists and applied linguists are interested in cohesion and coherence because they are important factors that influence text comprehension (Graesser, McNamara, & Louwerse, 2003; Halliday & Hasan, 1976). Cohesion reflects pure linguistic features of a text, whereas coherence reflects the psychological characteristics of the mental representations that people actively construct while they are attempting to understand the text (Graesser, Jeon, Yan, & Cai, 2007; Sanders & Maat, 2006). Namely, coherence indicates how people connect text components with their prior background knowledge and cohesion indicates the internal linguistic linking of the text components (Taboada, 2004). So, it is critical to examine the linguistic and psychological features systematically that influence cohesion and coherence to explain the mechanism of text comprehension (Graesser et al., 2007).

Behavioral science studies showed that text cohesion and coherence played an important role for investigating the effect of knowledge-based inferences on the integration of text components, for combining pure text features with people's background knowledge, and for constructing the mental representations of texts (Graesser, Singer, & Trabasso, 1994; Kintsch, 1988, 1998; Long, Wilson, Hurley, & Prat, 2006). The mental representations ultimately reflect deeper understanding, thereby indicating the successful integration of linguistic text-based features and background knowledge (Graesser et al., 2003).

From this perspective, many researchers analyzed the characteristics of cohesion and coherence over the past three decades (McNamara, Kintsch, Songer, & Kintsch, 1996; Sanders & Noordman, 2000; Sanders, Spooren, & Noordman, 1992). For example, McNamara et al. (1996) investigated the interaction effect between cohesion and people's background knowledge. They used various experimental tasks such as a background questionnaire, a reading time and recall task, a post-test task (i.e., text-based questions, elaborative-inference questions, bridging-inference questions, and problem-solving questions), and a sorting task. They manipulated four different experimental conditions to examine the interaction effect between coherence (i.e., a high coherence text condition vs. a low coherence text condition) and background knowledge (i.e., a high-knowledge student condition vs. a low-knowledge student condition). McNamara et al. found that high-knowledge students showed better performance when they read low coherence texts, whereas low-knowledge students showed better performance when they read high coherence texts. The findings of McNamara et al. suggest that text coherence interacts with background knowledge. Simply put, cohesion and coherence are essential components that are required to explain text comprehension (Graesser et al., 2003).

With the help of recent advanced computational linguistic technologies (Jurafsky & Martin, 2008) and corpus linguistic methodologies (Lindquist, 2009; Meyer, 2002), researchers in the Institute for Intelligent Systems (IIS) at the University of Memphis in recent years developed an automated computer system, Coh-Metrix that can computerize various text-based features of cohesion (Graesser, McNamara, Louwerse, & Cai, 2004) and the author of this article recently developed a new computer tool, TEES (an acronym for Text & Essay Evaluation System) that can evaluate English essays and texts based on a standardized norm. The standardized norm was created by a huge size of corpus.

The main purpose of this article is to introduce two automated language analysis tools, Coh-Metrix and TEES that can be used to analyze and evaluate various texts and essays based on many linguistic and psycholinguistic measures. This article also presents some research findings collected from Coh-Metrix studies.

## 2 Coh-Metrix

Coh-Metrix is an automated computer system that was developed by IIS (Institute for Intelligent Systems) researchers at the University of Memphis to analyze English texts and textbooks based on many linguistic and psycholinguistic features on cohesion (Graesser et al., 2007; Graesser et al., 2004).

Coh-Metrix is composed of several computational modules. In detail, the Coh-Metrix system contains a parser and a tagger (Brill, 1995) for parsing and tagging sentences automatically. The Coh-Metrix tool contains several corpus norms to analyze narrative or scientific texts based on different corpus norms. Coh-Metrix uses a mathematical formula, LSA (Landauer, Foltz, & Laham, 1998) to computerize the semantic cohesion for adjacent sentences. Basically, Coh-Metrix consists of various computational algorithms developed by computer scientists (Jurafsky & Martin, 2008).

With these advanced computational systems, Coh-Metrix provides a wide range of linguistic and psycholinguistic measures that reflect the characteristics of cohesion (Graesser et al., 2007). Specifically, the measures of Coh-Metrix include basic counts (the number of words, the number of sentences, the number of paragraphs, average sentence length), syntactic complexity (subject density, noun density), co-referential cohesion (argument overlap for adjacent sentences), semantic cohesion (LSA cosine for adjacent sentences), standard readability scores (Flesch Reading Ease score, Flesch-Kincaid Grade Level), connectives, and lexical diversity (type-token ratio) scores.

### 2.1 Basic counts

Coh-Metrix provides the number of words, the number of sentences, the number of paragraphs, and average sentence length scores. People tend to read longer sentences slowly, thereby indicating that those sentences are difficult to read (Graesser et al., 2004, 2007).

### 2.2 Syntactic complexity

Coh-Metrix provides two syntactic complexity scores, including subject density and noun phrase density scores. The subject density score indicates the mean number of words before the main verb of the main clause in a sentence (Graesser et al., 2004). The noun phrase density indicates the mean number of modifiers per noun phrase. The modifiers contain adverbs, adjectives, and determiners that qualify head nouns in a sentence (Graesser et al., 2004). Readers are inclined to feel difficult to read sentences with complex syntactic structures (Graesser et al., 2004).

### 2.3 Co-referential cohesion

The co-reference cohesion between two adjacent sentences is constructed when a noun in the first sentence appears again in the second sentence or a pronoun in the second sentence indicates another constituent in the first sentence (Graesser et al., 2004). Many behavior science studies showed that co-reference cohesion influenced text comprehension (Cirilo, 1981; Haviland & Clark, 1974; Manelis & Yekovich, 1976). Coh-Metrix uses argument (i.e., nouns, pronouns) overlap scores for adjacent sentences to measure the co-referential cohesion for those sentences (Graesser et al., 2004). Readers tend to feel easy to read sentences when arguments are overlapped in those sentences (Graesser et al., 2004).

### 2.4 Semantic cohesion

Coh-Metrix uses LSA to measure the semantic cohesion for adjacent sentences. LSA is a mathematical computer algorithm that is used for measuring semantic similarity between two text components (i.e., words, sentences, paragraphs, texts) based on a huge size of corpus (Landauer et al., 1998). The semantic cohesion score of Coh-Metrix indicates a LSA cosine value for adjacent sentences (Graesser et al., 2007). In general, people feel difficult to read sentences when the LSA cosine score for those sentences is low (Graesser et al., 2004).

### 2.5 Standard readability scores

The standard readability scores provided by Coh-Metrix are the Flesch Reading Ease score and the Flesch-Kincaid Grade Level score (Graesser et al., 2004). The Flesch Reading Ease score indicates a number between 0 to 100. In general, readers feel easy to read a text when the Flesch Reading Ease score of the text is high. The Flesch-Kincaid Grade Level score refers to a number between 0 to 12, indicating that each number represents a U.S. grade-school level (Graesser et al., 2004). Readers tend to feel difficult to read a text when the Flesch-Kincaid Grade Level score of the text is high. So, these standard readability can be index scores for measuring the level of difficulty of a text (Graesser et al., 2007).

### 2.6 Connectives

Many language psychologists demonstrated that connectives are important text markers that influenced text comprehension (Caron & Thuring, 1988;

Segal, Duchan, Scott, 1991; Millis & Just, 1994; Murray, 1997). Specifically, connectives can facilitate text comprehension (Millis & Just, 1994; Murray, 1997).

Millis and Just (1994) showed that the causal connective (i.e., because) could facilitate the causal relatedness of sentences, thereby indicating that connectives are important text markers that can influence text comprehension. The connective measures of Coh-Metrix consist of positive additive connectives (e.g., also, and, moreover), positive temporal connectives (e.g., after, before, when), positive causal connectives (e.g., because, so, therefore), negative additive connectives (e.g., however, but), negative temporal connectives (e.g., until, by), and negative causal connectives (e.g., although, albeit) for researchers who are interested in examining the effect of connectives on text comprehension (Graesser et al., 2004).

### 2.7 Lexical diversity

The lexical diversity score of Coh-Metrix is a type-token ratio. The type indicates an individual word in a text and the token indicates how many times the word appears in the text (Graesser et al., 2004). Readers are inclined to feel difficult to read a text when the type-token ratio of the text is high, because the readers should process many words in the working memory (Graesser et al, 2004).

## 3 Coh-Metrix based studies

Many researchers in the world have widely used the Coh-Metrix tool to analyze various texts and textbooks (Graesser et al., 2007; Jeon, 2011; Jeon & Lim, 2009; Kim & Jeon, 2013).

Graesser et al. (2007) compared a textbook for Newtonian physics, text materials created by language psychologists, tutorial dialogues between human tutors and college students, and tutorial dialogues between a computer tutor and college students using Coh-Metrix. They found that the physics textbook was similar to the text materials and the human tutor-student interaction tutorial dialogues were similar to the computer tutor-student interaction tutorial dialogue, indicating that the physics and experimental texts reflect the characteristics of written texts and the two types of tutorial dialogues reflect the characteristics of spoken texts.

Graesser, Jeon, McNamara, and Cai (2008) applied Coh-Metrix to analyze Einstein’s Dreams, a novel written by a physicist, Alan Lightman. They investigated whether the novel is more similar to narrative texts or to scientific texts. They collected narrative and scientific text from the TASA (Touchstone Applied Science Associates) corpus. The findings of Graesser et al. showed that the novel, Einstein’s Dreams was more similar to narrative texts than to scientific texts for many Coh-Metrix Measures.

Jeon (2011) examined the continuity of Korean middle school English textbooks using Coh-Metrix. Specifically, Jeon compared the reading materials in the Korean middle school English 1 textbook with those in the Korean middle school English 2 textbook. Jeon found that the continuity between the Korean middle school English 1 textbook and the Korean middle school English 2 textbook was not controlled appropriately.

These Coh-Metrix based studies imply that the Coh-Metrix tool can be effectively used to analyze various texts and textbooks.

## 4 TEES

The author of this article has recently developed a new computer tool that can be used to evaluate (or analyze) English essays and textbooks using various linguistic and psycholinguistic measures. The new computer tool is called TEES (an acronym of Text & Essay Evaluation System). Basically, TEES was developed to evaluate English essays based on a standardized norm. In the TEES system, the TASA corpus was used to create the standardized norm. The TEES system contains a variety of computational algorithms (Jurafsky & Martin, 2008), and uses the Stanford parser to parse and tag sentences.

### 4.1 The interface of TEES

The TEES system was developed by Java programming language in the Microsoft Windows platform. Figure 1 presents the TEES interface.

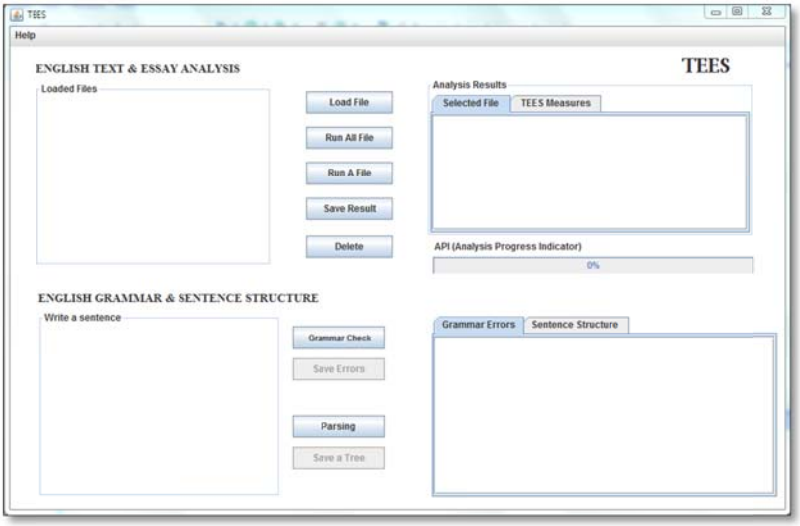
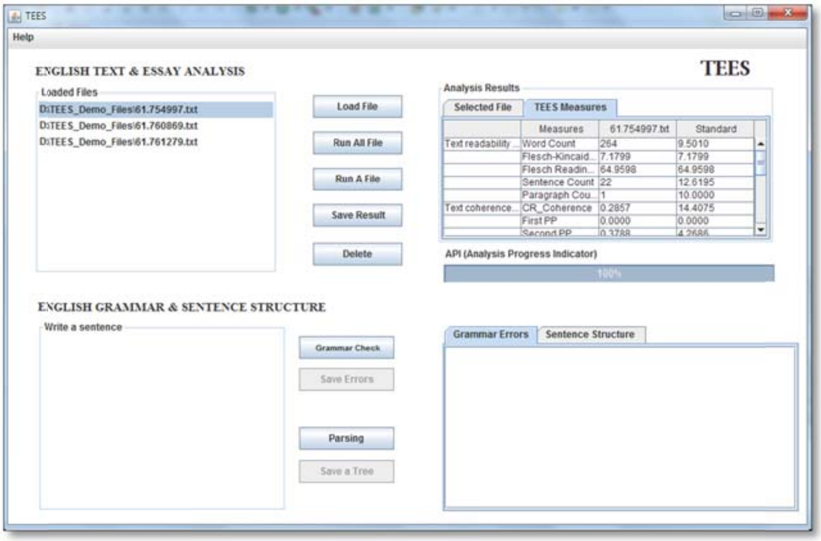


Figure 1. TEES interface

As presented in Figure 1, TEES contains four main modules. The “Loaded Files” module (see upper left in Figure 1) is used for loading essay or text files to be analyzed. The “Analysis Results” module (see upper right) shows the contents of a selected essay or text file, or shows the measures of TEES. The “Write a sentence” module (see bottom left of Figure 1) indicates a space into which user can type sentences directly to analyze the syntactic structures of the sentences or to find English grammar errors. The “Grammar Errors/Sentence Structure” module (see bottom right of Figure 1) presents the results of grammar error and sentence structure analyses.

### 4.2 The main functions of TEES

The main functions of TEES are text analysis, essay evaluation, sentence structure analysis, and English grammar error analysis. Specifically, the TEES system analyzes various types of texts with text readability, text coherence, sentence structure, vocabulary analysis, and text marker measures (see Figure 2).



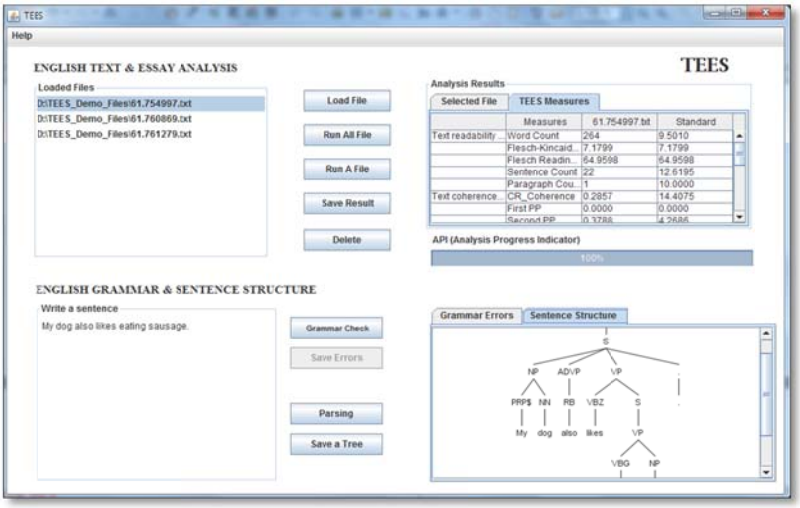
D:\TEES_Demo_Files\61.754997.txt
D:\TEES_Demo_Files\61.760869.txt
D:\TEES_Demo_Files\61.761279.txt

| | Measures | 61.754997.txt | Standard |
|---|---|---|---|
| Text readability ... | Word Count | 264 | 9.5010 |
| | Flesch-Kincaid... | 7.1799 | 7.1799 |
| | Flesch Readin... | 64.9598 | 64.9598 |
| | Sentence Count | 22 | 12.6195 |
| | Paragraph Cou... | 1 | 10.0000 |
| Text coherence... | CR_Coherence | 0.2857 | 14.4075 |
| | First PP | 0.0000 | 0.0000 |
| | Second PP | 0.3788 | 4.2686 |

Figure 2. TEES measures

As presented in Figure 2, the TEES system provides a variety of linguistic and psycholinguistic measures. The TEES system also provides standardized measures on those linguistic and psycholinguistic measures that can be used to evaluate English essays objectively. TEES uses the TASA corpus to create the standard norm.

The TEES system also can be used to analyze the syntactic structures of sentences. As presented in Figure 3, the TEES system uses the Stanford parser to analyze the syntactic structure of a sentence. TEES can be effectively used to scaffold students to learn syntactically complex sentences.



D:\TEES_Demo_Files\61.754997.txt
D:\TEES_Demo_Files\61.760869.txt
D:\TEES_Demo_Files\61.761279.txt

| | Measures | 61.754997.txt | Standard |
|---|---|---|---|
| Text readability ... | Word Count | 264 | 9.5010 |
| | Flesch-Kincaid... | 7.1799 | 7.1799 |
| | Flesch Readin... | 64.9598 | 64.9598 |
| | Sentence Count | 22 | 12.6195 |
| | Paragraph Cou... | 1 | 10.0000 |
| Text coherence... | CR_Coherence | 0.2857 | 14.4075 |
| | First PP | 0.0000 | 0.0000 |
| | Second PP | 0.3788 | 4.2686 |

My dog also likes eating sausage.

Figure 3. The sentence structure analysis of TEES

The TEES system can analyze English grammar errors made by second language learners of English automatically. The TEES system can analyze singular and plural noun-verb agreement errors, article errors, verb usage errors, and so on in the essays written by students. TEES provides the students with correct forms on the grammar errors. So, they can use TEES to learn English grammar for themselves.

## 5 Conclusion

The most advanced computer technologies (Jurafsky & Martin, 2008) and corpus linguistic methodologies (Lindquist, 2009; Meyer, 2002) have made it possible to enable the computer tools such as Coh-Metrix and TEES to analyze (or evaluate) various texts and essays automatically. Coh-Metrix and TEES can be widely applied to investigate the explicit and implicit features of texts based on a variety of linguistic and psycholinguistic measures. Hopefully, Coh-Metrix and TEES will be actively used by many researchers in the areas of applied linguistics, English education, corpus linguistics, language psychology, and computational linguistics to explore the nature of language.

## References

Baayen, R. H., Piepenbrock, R., & Gulikers., L. (1995). *The CELEX lxical dtabase (CD-ROM)*. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, *21*, 543-566.

Caron, J., Micko, H., & Thuring, M.(1988). Conjunctions and the recall of composite sentence. *Journal of Memory and Language*, *27*, 309-323.

Cirilo, R. (1981). Referential coherence and text structure in story comprehension. *Journal of Verbal Learning and Verbal Behavior, 20,* 358-367.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology, 33,* 497-505.

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

Graesser, A. C., Jeon, M., Cai, Z., & McNamara, D. S. (2008). Automatic analyses of language, discourse, and situation models. In J. Auracher & W. van Peer (Eds.), *New beginnings in literary studies* (pp. 72-88). Cambridge Scholars Publishing.

Graesser, A. C., Jeon, M., Yan, Y., & Cai, Z. (2007). Discourse cohesion in text and tutorial dialogue. *Information Design Journal*, *15*(3), 199-213.

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text. In A.P. Sweet & C.E. Snow (Eds.), *Rethinking reading comprehension* (pp. 82-98). New York, NY: Guilford Publications.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers*, *36*, 193-202.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*, 371-395.

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

Haviland, S., & Clark, H. (1974). What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior*, *13,* 512-521.

Jeon, M. (2011). A Corpus-based analysis of the continuity of the reading materials in middle school English 1 and 2 textbooks with Coh-Metrix. *The Journal of Linguistic Science, 56*, 201-218.

Jeon, M., & Lim, I. (2009). A Corpus-based analysis of middle school English 1 textbooks with Coh-Metrix. *English Language Teaching, 21*(4), 265-292.

Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice-Hall.

Kim, J., & Jeon, M. (2013). A corpus-based analysis of the continuity of the listening materials in middle school English textbooks. *Journal of the Korean Data Analysis Society*, *15*(4), 1987-2000.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363-394.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, *25*, 259-284.

Lindquist, H. (2009). *Corpus linguistics and the description of English*. Edinburgh: Edinburgh University Press.

Long, D. L., Wilson, J., Hurley, R., & Prat, C. S. (2006). Assessing text representations with recognition: The interaction of domain knowledge and text coherence. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *32*, 816-827.

Manelis, L., & Yekovich, F. (1976). Repetitions of propositional argument in sentence. *Journal of Verbal Learning and Verbal Behavior, 15*, 301-312.

McEnery, T., & Wilson, A. (2007). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

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

Meyer, C. F. (2002). *English corpus linguistics: An introduction*. Cambridge: Cambridge University Press.

Millis, K., & Just, M.(1994). The influence of connectives on sentence comprehension. *Journal of Memory and Language*, *33*, 128-147.

Murray, J. (1997) Connective and narrative text: The role of continuity. *Memory & Cognition*, *25*(2), 227-236.

Sanders, T. J. M., & Maat, H. P. (2006). Cohesion and Coherence: Linguistic approaches. In Brown, K. et al. (Eds.), *Encyclopedia of language and linguistics* (pp. 591-595). London: Elsevier.

Sanders, T. J. M., & Noordman, L. G. M. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse Processes, 29,* 37-60.

Sanders, T. J. M., Spooren, W., & Noordman, L. G. M. (1992). Toward a taxonomy of coherence relations. *Discourse Processes*, *15*(1), 1-35.

Segal, E., Duchan, J., & Scott, P.(1991). The role of interclausal connectives in narrative structuring: Evidence form adults' interpretations of simple stories. *Discourse processes*, *14*, 27-54.

Taboada, M. (2004). *Building coherence and cohesion*. Amsterdam and Philadelphia: John Benjamins.

Moongee Jeon
Department of English
Konkuk University
120 Neungdong-ro, Gwangjin-gu, Seoul 143-701
Email: mjeon1@kokuk.ac.kr

Received: October 28, 2014
Revised: December 10, 2014
Accepted: December 20, 2014