Modern English Education, Vol. 17, No. 2, Summer 2016
http://dx.doi.org/10.18095/meeso.2016.17.2.01

Exploring the Role of Automatically-derived Text Complexity Features in L2 Reading Test Development

Sumi Han *
Seoul National University
Jeong-Ah Shin
Dongguk University

Han, Sumi & Shin, Jeong-Ah. (2016). Exploring the role of automatically-derived text complexity features in L2 reading test development. *Modern English Education*, *17*(2), 1-19.

Although automatic text analysis tools are available, little research has been conducted on the application of such tools in reading assessments. When the ratio of academic vocabulary and transitions are computed automatically and used in test development, the text selection-revision procedure can be fast and transparent by complementing test developers' expertise. To obtain empirical evidence for the utility of automatic text complexity features, this study attempted to explore the role of automatically-derived text complexity features in an intensive English program (IEP) reading assessment. Based on previous literature and the testing context, a total of 11 text complexity features as lexical, syntactic, and semantic variables were chosen, and their accountability for the IEP reading item difficulty was automatically measured by using three text analysis tools—Lexile, the Compleat Lexical Tutor, and Coh-Metrix. Results showed that seven complexity features significantly correlated with the reading item difficulty. Stepwise multiple regressions showed that a set of four lexical and semantic text complexity features (i.e., word length, total word counts, Lexical Semantic Analysis (LSA), connectives) explained about 45% of the variance in the reading item difficulty. The results and findings of this study are discussed with regard to limitations and implications for both reading assessments and instruction.

**[reading test development/text complexity/automated text analysis/
읽기 시험 개발/텍스트 복잡성 요인/자동 텍스트 분석]**

* First author: Sumi Han, Corresponding author: Jeong-Ah Shin

I. INTRODUCTION

In reading comprehension assessments, texts are the major input that readers need to interpret while answering test questions. Thus, the choice of texts as the very first step of test development requires careful decision-making to meet the test goal and test-takers' reading ability (Alderson, 2000). Typically, it is time-consuming to search for appropriate texts and revise them iteratively to meet pre-determined text selection criteria such as topic coverage, text length, types of vocabulary, and grammatical structures. An important consideration into text selection is test-takers' text readability because too complex and difficult texts would discourage students to read further while too easy texts would cover distinct achievements among students.

With the help of technology, text complexity and reading ability have been examined more automatically. Traditionally, the predictability of reading comprehension by text complexity measures was examined manually, which was more time-consuming and less consistent (e.g., Gray & Leary, 1935). Thus, researchers wanted one comprehensive measure of readability called readability formulas (based on a certain set of weighted features). Currently, there are more than 100 readability formulas. With advances in computer technology, two readability formulas, MetaMetrics' Lexile Analyzer (Lexile, henceforth) and Microsoft's Flesch-Kincaid Grade Level (FKGL), have been most widely used as they are automatically computed (see Fisher, Frey, and Lapp (2012) for a review on various readability formulas). Simply based on sentence length and word length (number of syllables per word) or total word counts, these two formulas automatically retrieve or assess text readability. Recently, online text analyzers have been widely employed in text analysis research because they encompass a variety of features such as the number of academic vocabulary and the number of connectives beyond word- and sentence-level features. Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004) and Cobb's (n.d.) the Compleat Lexical Tutor (http://www.lextutor.ca) are most widely used as research tools. It is because such automatically-derived text features or measures can allow researchers to conduct an extensive and iterative text analysis much faster in a more objective and reliable way.

To date, automatic text analyzers and their features have been widely used in text analysis for developing reading materials in the field of corpus research. Little has been done in the field of reading assessments. When automatic text complexity features are adopted in test development, however, selecting and modifying texts appropriate for the test goal and test-takers' reading ability can be done more efficiently and reliably. Texts can be analyzed in terms of various lexical, syntactic, and semantic characteristics in a few seconds. As a variety of text features influence reading test performance, detailed profiles of texts can be provided; otherwise, test developers would solely rely on their own

experiences and intuition. The appropriateness of texts would vary in the amount of their expertise. In this sense, automatic, quantitative features of text complexity can be guidelines to decipher text characteristics more efficiently and reliably. It is important to note that these automatically-computed text features cannot replace human judgments; rather, it can complement somewhat subjective judgments by test developers and assist in mechanical and tedious frequency analyses. To address the lack of research, this study aimed to explore the relationship between a set of automatically-derived text complexity features and the difficulty of reading items of intensive English program (IEP) placement tests in an English as a Second Language (ESL) context. Practically, a certain set of text complexity features, which have been explicit or implicit criteria of text selection/modification in the IEP reading test development, were targeted for obtaining empirical evidence for justifying the use of such criteria. It is hoped that this investigation will point out the benefits of technologies in developing well-constructed reading tests and teaching materials for our future students.

II. LITERATURE REVIEW

1. Text Complexity and Automatic Text Analysis Tools

The most comprehensive framework of text features and their relation with reading comprehension is the United States' Common Core State Standards (National Governors Center for Best Practices & Council of Chief State School Officers [NGA & CCSSO], 2010). The Standards takes into consideration text complexity because successful reading skills are important for college study and future career, and thus, reading passages should be selected as text complexity increases gradually in successive years of schooling. Text complexity is further classified into qualitative, quantitative, and reader-task dimensions. The first two dimensions focus on the inherent complexity of text, but the last dimension focuses on readers and tasks in reading situations. A brief description is given for each dimension as follows (NGA & CCSSO, 2010, p. 4):

- Qualitative dimensions of text complexity refer to aspects measured qualitatively such as language conventionality and clarity, knowledge demands, levels of knowledge (e.g., background knowledge, prior knowledge, cultural knowledge).
- Quantitative dimensions of text complexity refer to aspects measured in numbers such as word length or counts, sentence length, and text cohesion, which are typically measured by computer software programs.
- Reader-Task dimensions of text complexity refer to aspects related to particular

students (e.g., motivation, knowledge, and experiences) and tasks (e.g., purpose, and the complexity of tasks or questions).

The three dimensions of text complexity are interrelated as the reader and the text interact while reading. They also complement one another as they demonstrate different aspects of text complexity. Among the three dimensions, this study exclusively focuses on the quantitative dimension of text complexity with a focus on the burgeoning automatic text analysis tools.

Quantitative features of text complexity are advantageous because they can be measured with computer software easily and reliably. When we deal with long and various texts, this automatic measurement of text complexity is most beneficial for predicting the difficulty of texts and selecting and revising texts appropriately. Fisher and Frey (2014) point out that the use of quantitative measures facilitates initial searches for target texts. That is to say, "quantitative measures are best used to initially locate a text within a particular grade range. In fact, they are fairly effective at identifying texts with a prescribed band of complexity—an important first step in identifying potential texts" (Fisher & Frey, 2014, p. 238). McNamara, Graesser, and Louwerse (2012) also argued that these automated features are essential for determining texts with appropriate amounts of complexity: Readability formulas can be used in test design "to choose a readability range for the passages and write test items to the passage along some construct model" (p. 89). In formative assessments, automated features showing specific text characteristics can also be used to "identify specific weaknesses and strengths of readers" (McNamara et al., 2012, p. 90). In this sense, automated, quantitative measures of text complexity are worthy of examining to identify what types of text characteristics make reading comprehension easier or more difficult, which can guide future text selection.

To date, four major automatic text analysis tools have been developed and used in developing reading materials by measuring text readability or text complexity. As mentioned before, Lexile and FKGL are simply based on total word counts and sentence length in each text. Advances in computer technology have allowed us to access more detailed quantitative text information. The Compleat Lexical Tutor and Coh-Metrix are more sophisticated tools which quantify lexical, syntactic, or semantic aspects of texts in determining the degree of text complexity. The Compleat Lexical Tutor is a specialized tool of analyzing different types and levels of vocabulary. Coh-Metrix is a more advanced text analyzer developed at the University of Memphis, which automatically measures 80 variables of text complexity features and formulas (M. Jeon, 2015; see Appendices 1, 2, and 3 for the screenshots of the three automatic text analysis tools, Lexile, the Compleat Lexical Tutor, and Coh-Metrix, in turn). Though different sets of text complexity features are employed, these automatic tools of text analysis have been seen useful for deciphering

the characteristics of texts efficiently and reliably.

Previous Korean studies employed automated text analyzers mainly for analyzing text complexity of instructional materials rather than developing test development. J. E. Kim and I. C. Choi (2015), for instance, examined the overall difficulty level of EBS-CSAT prep books and High School Textbooks in terms of vocabulary, syntactic complexity, coherence, and readability using Lexile and Coh-Metrix. They found that the linguistic difficulty levels of EBS-CSAT prep books are much higher than those of High School Textbooks. M. Jeon and I. Lim (2009) used Coh-Metrix to analyze eight popular first grade middle school English textbooks in terms of the basic count, readability, word frequency, lexical diversity, syntactic complexity, co-referential coherence, and semantic coherence measures. They found differences in these measures among the chapters of each textbook. Both studies yielded pedagogical implications of the automatic text analysis tools for materials development.

2. Automatic Text Complexity Features in Reading Test Development

In the literature of reading assessments, a variety of text complexity features have been examined in order to identify which text and text-related lexical, syntactic, and semantic variables together explain a significant amount of the variance in predicting reading item difficulty or comprehension. In early research with little technical help, text complexity features were manually identified and compared with reading comprehension. Gray and Leary (1935) measured the predictability of reading comprehension by average sentence length, number of different hard words, number of personal pronouns, percentage of unique words, and number of prepositional phrases. These features explained about 64.5% of the variance in reading comprehension test scores. Drum, Calfee, and Cook (1981) found that simple characteristics of texts and items, such as word counts in the stem and options and word counts with more than one syllable, accounted for 70% of the variance in reading item difficulty. Focusing on a few variables examined in Drum et al.'s (1981) study, Embretson and Wetzel (1987) and found out that the existence of connectives (e.g., *however*, *thus*) in reading passages significantly predicted reading item difficulty: the more connectives a text contains, the easier its corresponding reading items are. Freedle and Kostin (1993) conducted a comprehensive study on the predictability of reading item difficulty by various text and text-related variables. With 213 multiple-choice reading items of TOEFL (Test of English as a Foreign Language), they found that 33% of the variance in the item difficulty was accounted for by eight variables, such as word counts appearing in the topic sentence and passage with argument structure.

With the development of computer programs, more recent research has identified the role of quantitative features of text complexity in accounting for reading item difficulty. In

general, sentence and word length and word counts have been found significant predictors of item difficulty; however, other quantitative features have also been investigated by employing online text analyzers. Despite a growing interest in computerized text analyses in corpus linguistics research, few studies have been conducted to decipher automatic text complexity features in reading assessments. To date, McNamara and his colleagues have extensively examined measures of cohesion and coherence automatically computed by using Coh-Metrix (e.g., Graesser, McNamara, Louwerse, & Cai, 2004; McNamara & Kintsch, 1996; McNamara et al., 2012). In this article, cohesion refers to cues in the text helping the reader build a coherent representation whereas coherence, usually associated with the interpretability of discourse, refers to the representational relationships of a text in the reader's mind (Graesser et al., 2004).

Among the measures, connectives and Latent Semantic Analysis (LSA) have been frequently investigated. Connectives are one index of text cohesion, which is measured by the use of cohesive devices in a text. LSA is one index of text coherence, which is measured as the degree to which each sentence is similar to every other sentence in the text. Connectives such as transitions can be cues for how much sentences and ideas are organized; LSA, based on a mathematical formula called singular value decomposition (SVD), indicates the degree to which the ideas discussed in the text are semantically similar or connected. McNamara and Kintsch (1996) found that high text cohesion and coherence improved reading comprehension, especially for low proficient readers. When reading more cohesive and coherent texts, less skilled readers could compensate for their lack of word or background knowledge about text messages. On the other hand, McNamara et al. (2012) pointed out that texts in high cohesion and coherence can be relatively difficult to read if they include many difficult words, mostly academic vocabulary. Science texts are relatively more cohesive and coherent than narrative texts but can be more challenging to read due to discipline-specific words expressing scientific concepts and ideas. In conclusion, the findings of studies by McNamara and his colleagues suggest that text cohesion and coherence and their relation with text comprehension should be understood along with other factors such as reading proficiency and the ratio of vocabulary. Crossley, Greenfield, and McNamara (2008) further reported that the combination of three measures—content word overlap, sentence syntax similarity, and word counts—accounts for 86% of the variance in the Japanese students' cloze test performance.

As can be seen above, a wide range of automatically-measured features of text complexity have been examined, but different sets of features have been suggested as significant predictors of reading item difficulty or comprehension. Little empirical research has been carried out, and thus, the influence of such significant features on text comprehension has not been definitive, either. This study attempted to address this lack of

research on automatic text complexity features in the context of Intensive English Program (IEP) reading test development. More specifically, this study aimed to meet a practical need of reliable and evidence-based text selection in the IEP test development by identifying automatic features of text complexity that predict the reading item difficulty. With these aims of research, the following two research questions are addressed in this study:

1) To what degree do text complexity features correlate with the reading item difficulty of the IEP tests?
2) To what degree do text complexity features predict the reading item difficulty of the IEP tests?

The two research questions will be addressed by correlation and multiple regressions analyses, in turn, and it is hoped that results of this study will offer important information for facilitating the IEP text selection process.

III. METHOD

1. Reading Comprehension Test Data

1) Participants and Test Settings

A total of 443 examinees took one of the Fall 2010, 2011, and 2012 IEP placement tests at a Southwestern university in the United States. All of the students just came from either Saudi Arabia or China at the time of test administration. A majority of them were male students. Their levels of English proficiency varied from beginning to advanced levels. These ESL students did not have an official score of any standardized English proficiency test, and the placement test results were used to place them into appropriate levels of class in the IEP. Thus, most examinees were highly motivated to perform well on these high-stakes placement tests.

2) Reading Sections of the Placement Tests

The IEP placement tests assessed the examinees' academic English language proficiency. The reading section of each placement test consisted of five passages. There was one common reading passage *Bioluminescence* used as a linking passage across the three reading sections. Therefore, the pool of the three reading sections from the three placement

tests included a total of 13 reading passages with 110 items. A total of six objectives were targeted in the reading test: vocabulary, details, and main ideas for basic comprehension; and text organization, inference, and purpose/attitude for advanced comprehension. Different reading passages in each reading section represented different levels of reading proficiency.

Reading passages of the IEP placement tests were chosen and modified in both objective and subjective ways; however, no empirical evidence had been sought for the use of the recommended text selection guidelines. From an objective perspective, an in-house guideline of text complexity is mainly used including total word counts and mean sentence length, which are automatically computed and provided in the test specification per passage. With these quantitative, objective criteria, test development staff members searched for authentic reading passages and revised them suited to a target level of reading proficiency. From a subjective perspective, test development staff members employed their testing experiences and intuition to find topically related or level-appropriate passages. Passages were modified by replacing basic vocabulary with academic words or vice versa and rewriting sentences with simpler or more complex structures and in passive or active structures. Cohesive devices were added to passages to demonstrate logical sequences of ideas better. In sum, various features of text complexity, either objectively or subjectively measured, were taken into account in the text selection and modification procedure of the IEP reading test development. Unfortunately, no empirical research had been conducted to justify the application of such objective and subjective features of text complexity to the test development procedure; this study was initiated to improve this development process with the help of automatically-computed text complexity features.

2. Analyses

1) Reading Item Difficulty as Dependent Variable

In the statistical analyses of this study, reading item difficulty was used as the sole dependent variable. As the reading item difficulty values, which are percent correct per item, were based on three different groups of examinees, any influence on item difficulty due to the difference in group ability needed to be eliminated by using delta equating (Freedle & Kostin, 1993; Livingston, 2013). The proportion of correct per item was an average item score (P), with which a delta value (Δ) was computed using the following equation: $\Delta = NORMINV((1-P), 13, 4)$. In this way, the delta values were normally distributed with mean = 13 and standard deviation = 4. Once the delta values were obtained, delta equating was conducted with the Fall 2012 Placement test and its linking passage *Bioluminescence* as a reference group because the placement test had better item

discrimination power and normal distribution than the other two placement tests. In general, the delta equating employed in this study took three steps: First, the delta values of the same passage (i.e., the linking passage) in the three reading sections were equated in pairs: delta equating was done on the items between the Fall 2012 linking passage (equated delta of reference group) and each of the Fall 2010 and Fall 2011 linking passage (observed delta of observed group). Secondly, the means and standard deviations of the delta values for the three linking passages were computed. Lastly, the descriptive statistics of the delta values were used to obtain equated deltas of the other reading passage items in each test. For the rest of this article, the value of equated delta is called reading item difficulty.

2) Text Complexity Features as Independent Variables

Based on the IEP reading test development procedure and previous research findings, a set of 11 independent variables were chosen and categorized into lexical, syntactic, and semantic categories, including five lexical variables (Lex1-Lex5), three syntactic variables (Syn1-Syn3), and three semantic variables (Sem1-Sem3). Table 1 presents three pieces of information about each of the 11 variables: feature, name, and definition. For instance, the very first variable presented is a type of lexical category, its abbreviated and full variable names are Lex1 and word length, and it is defined as syllables per word. Each feature was automatically computed per passage using one of the three automatic text analyzers, Lexile, the Compleat Lexical Tutor, and Coh-Metrix. In particular, word counts, word length, and sentence length were chosen with regard to the IEP guideline of text complexity. The other features were selected because they were either considered in the IEP text selection procedure or shown to correlate with reading item difficulty as found in the literature. Thus,

TABLE 1

A Set of 11 Measures of Text Complexity Used in This Study

Feature	Name	Definition
Lex1	Word length	Syllables per word
Lex2	Word counts	Total word counts
Lex3	Academic words*	Academic word counts/Total word counts
Lex4	Off-list word counts*	Word counts on the off-list/Total word counts
Lex5	Type-Token Ratio	# of unique words/Total word counts (for content words)
Syn1	Sentence length	# of total word counts/# of sentences
Syn2	NP density	# of modifiers/# of noun phrases
Syn3	Passive	# of passive voice structures/Total word counts
Sem1	Co-reference cohesion	% of all sentence pairs sharing a common noun, pronoun or noun phrase
Sem2	Lexical Semantic Analysis (LSA)	an LSA cosine value for adjacent sentences
Sem3	Connectives	# of connectives (e.g., transition words)/Total word counts

Note. *Categories from the Compleat Lexical Tutor; others from Coh-Metrix version 3.0.

other possible features such as subject density (the number of words before the main verb of the main clause in a sentence) were not included in this study. It can also be noted that LSA as a mathematical algorithm is a score of the meaning similarity among sentences in a text, thereby showing how much similar topics and ideas are discussed in the text.

3) Statistical Analyses

Statistical analyses were performed on the reading test data by using correlation and multiple regression techniques. All the dependent and independent variables used in the analyses were continuous variables. SPSS 22.0 software was used to perform the statistical analyses in this study (IBM corp., 2013).

As a correlation technique, Pearson product-moment correlation (usually denoted by *r*) was used to measure the degree and direction of relationship between two variables, that is, the reading item difficulty and each feature. Pearson's *r* coefficients range from -1 (perfect negative relation) to 1 (perfect positive relation), and 0 is no relation between the two variables. When more than three or four pair-wise comparisons are measured simultaneously, it is recommended to use an adjusted level of significance to control for Type I error (Tabachnick & Fidell, 2007). In this study, the Bonferonni procedure was used to determine the critical probability. The adjusted *p* value used to test significance was .005, 1-tailed (.0045 as .05 divided by 11 variables).

Multiple regression was employed to measure the variance in the reading item difficulty as the sole dependent variable explained by a set of text complexity features as independent variables. In so doing, significant predictors of the reading item difficulty were identified. More specifically, stepwise multiple regression analyses were performed. In stepwise multiple regression, independent variables are entered into the equation one at a time, and the order of entry of independent variables is based on the strength of correlation: the independent variable with the strongest (positive or negative) correlation with the dependent variable is entered first. Then, the variable with the strongest partial correlation is added to the equation and tested for significance. Those variables already added are deleted at any step when they do not contribute significantly to regression. Throughout this regression analysis, a parsimonious subset of text complexity features that significantly predicted the item difficulty was identified. The assumptions of multiple regression including normality, linearity, and constant variance were checked and all the assumptions were met. Outliers were also checked and no extreme datapoints were identified (Tabachnick & Fidell, 2007).

IV. RESULTS

This section provides the results of correlation and multiple regression analyses on the reading test data and discussions while addressing each of the two research questions.

1. Research Question 1

The results showed that the reading item difficulty moderately correlated with each independent variable. Table 2 lists correlation coefficients between the reading item difficulty (ID) and the 11 text variables. No noticeable strong correlations were found among the variable categories. Among the 11 variables, the correlations with seven variables were statistically significant at the adjusted p level. The significant correlations show weak or moderate relationships (absolute coefficients ranging from .23 to .53). In particular, the reading item difficulty had a negative moderate relationship with Sem3 (r = -.53) but a positive negative relationship with Lex1 (r = .52). All the lexical variables but Lex5 significantly correlated with the item difficulty. Syn2, Sem1, and Sem2 did not demonstrate significant relationships with the item difficulty.

TABLE 2

Correlation Coefficient Between Reading Item Difficulty (ID) and Each Feature

	Lex1	Lex2	Lex3	Lex4	Lex5	Syn1	Syn2	Syn3	Sem1	Sem2	Sem3
ID	.52*	.46*	.34*	.23*	.23	.40*	.04	.37*	-.19	.15	-.53*

Note. *Significant coefficients at the adjusted p = .005, 1-tailed.

2. Research Question 2

Table 3 illustrates the results of a stepwise multiple regression analysis with the reading item difficulty as the sole dependent variable and the 11 text variables as a set of independent variables. A total of eight regression models with different sets of predictors of the equated delta were found statistically significant at the p level of .000. For each model, R^2, adjusted R^2, SE (Standard Error) of Estimate are provided along with its predictors and partial coefficients for the predictors. Sets of five text variables (Sem3, Lex3, Lex2, Lex1, Sem2) contributed to regression across the models, but Sem3 in Model 4 and Lex3 in Model 7 were not significant predictors, and each of them was deleted in the subsequent model.

TABLE 3

Stepwise Multiple Regression With Reading Item Difficulty as Dependent Variable

Model	R^2	R^2 (Adjusted)	*SE* Estimate	Partial Coefficients (*B*) of Predictors				
				Sem3	Lex3	Lex2	Lex1	Sem2
1	.277	.271	1.896	-15.515*				
2	.385	.373	1.757	-15.216*	38.322*			
3	.411	.395	1.727	-8.054*	44.969*	.004*		
4	.438	.416	1.670	6.762	24.608*	.008*	7.442*	
5	.434	.418	1.670		30.360*	.006*	4.912*	
6	.454	.433	1.671		30.793*	.006*	5.113*	3.951*
7	.478	.453	1.642	18.410*	15.388	.012*	12.123*	6.292*
8	.471	.450	1.645	24.051*		.013*	15.279*	6.971*

Note. *Significant coefficients at the adjusted p = .05, 1-tailed.

Among the eight models, three models of Models 2, 5, and 8 that include only significant predictors provide important results. First, Model 2 is based on two significant predictors, Sem3 and Lex3, explaining about 37% of the variance in the equated delta. Among the 11 variables, Sem3 and Lex3 were first entered into the equation as statistically more significant predictors. The predictive power by Sem3 was about 27% and by Lex3 about 10%. Second, Model 5 has three significant predictors of Lex3, Lex2, and Lex1, which are all lexical variables. These three lexical variables explained about 42% of the variance in the equated delta. Last, Model 8 has four significant predictors of Sem3, Lex2, Lex1, and Sem2, which explained about 45% of the variance, the largest accountability afforded by a set of only significant predictors. Model 7 with an additional but insignificant predictor of Lex3 had the same amount of accountability by its five predictors. Another noteworthy point is that the sign of the partial coefficient of Sem3 that was negative in Models 1, 2, and 3 became positive in Model 8, perhaps because of the added predictors such as Lex1 and Sem2. Thus, Model 8 had the most parsimonious set of significant predictors with the largest predictive power of the variance in the equated delta, with an F (4, 105) = 23.329, p = .000 (R^2 of .471; adjusted R^2 of .450).

V. DISCUSSION

This study explored the relationship between a set of text complexity measures generated by online tools and IEP reading item difficulty to address two research questions. The first research question was on the relationship between the reading item difficulty and each text complexity features. Among the 11 variables, lexical features except Type-Token Ratio were significantly correlated with the reading item difficulty. Among them, word length most significantly correlated with the item difficulty, showing that lengthy words and texts increased the difficulty level of reading comprehension. Types of vocabulary

were also shown to be related with the reading test performance: The more academic words and off-list words are used in a text, the more difficult it is to answer the reading questions. That is, difficult words make reading challenging.

Among the three syntactic features, not only sentence length but also the ratio of passive structures significantly correlated with the item difficulty. The more passive structures were used in a text, the more difficulty the test-takers had in answering the questions about the text. The last significant feature was the ratio of connectives as the sole semantic variable. Unexpectedly, connectives had the strongest but negative relation with the item difficulty among the seven significant features of text complexity. It is because total word counts and mean sentence length are the most widely used features when measuring text complexity (e.g., Drum et al., 1981; Gray & Leary, 1935).

Moreover, the result shows that the more connectives are used, the more difficult they answered the questions of the text. This is consistent with the finding of McNamara et al.'s study (2012) that highly cohesive texts can be relatively difficult to read because they often deliver abstract concepts or ideas with difficult words. As such, the IEP reading tests measured the test-takers' academic reading ability with academic texts. Cohesive devices in academic texts could not be helpful for even less skilled readers. Interestingly, the ratios of academic vocabulary, off-list word counts, passive, and connectives were those features considered in the subjective phase of text selection and modification. In this way, such correlates of the reading item difficulty are empirical evidence for the subjective decision-making process by the IEP test developers.

The second research question was to investigate a parsimonious set of significant predictors of the reading item difficulty. It was turned out that a set of word length, total word counts, LSA, and connectives explained roughly half amount of the variance in the reading item difficulty. First, the ratio of connectives explained the largest amount of the variance among the four features. This is quite a large amount of predictability for the four automatic features when no qualitative features or question items were considered. Along with another semantic predictor of LSA, the reading items were also more difficult when their texts were more cohesive. Second, the reading items were more difficult to answer when the reading texts included more lengthy words and more words as commonly found in previous research (e.g., Drum et al., 1981). Third, it should be noted that the IEP reading tests were developed through three pre-determined quantitative measures of text complexity, including total word counts, mean sentence length, and FKGL. As FKGL is also based on total word counts and sentence length, the criteria are on the two surface-level features of text complexity. As the results show, total word counts was the only significant predictor; sentence length was not included in the regression model. Lastly, the four significant predictors of the reading item difficulty indicate interrelationships among them. Text cohesion needs to be understood along with text length: when texts become

longer, they become more cohesive but become difficult to read. LSA and the ratio of connectives, those features subjectively considered by the test-developers, were quantified in this study and turned out significant predictors of the reading item difficulty but the ratio of academic vocabulary was not. As McNamara et al. (2012) explained, however, highly cohesive texts like the IEP reading passages include a high ratio of academic vocabulary words to deliver abstract and complex concepts and ideas. Again, text cohesion and coherence and the use of academic vocabulary need to be understood as a whole in academic English reading assessments.

VI. CONCLUSION AND IMPLICATIONS

This study investigated a set of automatically-derived measures of text complexity and their relation with the IEP reading item difficulty. The results showed that different lexical, syntactic, and semantic features each significantly correlated with the reading item difficulty; only two lexical and two semantic features together explained the largest amount of variance in the reading item difficulty of the IEP placement tests. As found in previous research, word length and total word counts significantly predicted the reading item difficulty in this study; cohesion and coherence features of the ratio of connectives and LSA, which have been subjectively considered in the IEP test development, also significantly predicted the reading item difficulty. These findings provide not only empirical evidence for the legitimacy of the use of the objective and subjective criteria used in the IEP reading text selection but also useful information for improving the efficiency of text selection (and item writing) in the future test development.

Implications of this study can be discussed for future reading assessments and instruction. Regarding reading assessments, this study points out the importance of applying automatic text complexity features to test development. With the advances in technology, a large amount of text data can be processed efficiently. Automatic text analyzers such as Lexile, the Lexical Tutor, and Coh-Metrix have been developed to speed up the process of text analysis, which will be more transparent, consistent, and time-saving. In reading test development, such free online text analyzers can facilitate text searches and revisions, which otherwise would be solely done by test developers in many cases.

In addition, this study also shows the importance of seeking empirical evidence to justify iterative decision-making steps in test development. In this study, both objective and subjective criteria of text selection used in the IEP test development were examined if they predict reading item difficulty. Evidenced-based text selection procedures can support decision-making to choose and modify texts so that they fit to particular reading proficiency levels and test purposes (Castello, 2008). Text selection and revision will be

more consistent and time-saving, and test developers will be more likely to focus on item writing. Pedagogically, this study shows that the use of academic vocabulary and text cohesion and coherence are important in determining the complexity of academic texts. Finding and revising texts that match to students' reading proficiency are quite challenging to reading teachers; automatic, objective features can thus aid teachers in analyzing textbooks and supplementary reading passages if they provide an appropriate amount of complexity that challenges students but do not overwhelm them. To sum up, the automation of text analysis can not only facilitate reading test development but also develop more level-appropriate reading materials.

The use of such automatic tools can support developing the newly-developed criterion-referenced English test of the College Entrance Exam. Texts used in previous tests can be analyzed along with their item scores, thereby obtaining information about features which increase or decrease text complexity and thus affect reading test performance. Such information about text features and reading comprehension can also be used to select new texts for different thresholds of reading proficiency and assist teachers in selecting level-appropriate texts for mid-term and final reading tests.

This study was initially motivated to examine a certain set of text complexity measures with a practical interest into the objective and subjective criteria used in the IEP for text selection. Therefore, two limitations can be presented, which will be discussed as venues of future research. First, a better picture of text complexity can be obtained when we consider other aspects encompassing qualitative text features and characteristics of readers and test items as well as quantitative text features (Alderson, 2000; NGA & CCSSO, 2010). Second, different testing contexts will result in a different picture. The three placement reading tests of this study were to assess ESL students at varying proficiency levels, and thus it needs caution in interpreting what was found in this study. For Korean students, a different set of automatic text complexity features will determine their reading ability. As Korean reading tests often include short texts, syntactic features such as noun phrases and passives and the ratio of academic vocabulary rather than total word counts and connectives will play a critical role in explaining the amount of reading comprehension. Overall, more research in either reading testing or instruction is called for examining diverse automatic features of text complexity as determinants of reading ability with different reading texts and readers.

Recently, we have seen computer and online programs deal with a bulk of text data efficiently. Using a variety of automatic text complexity features, they can help us understand text characteristics systematically. When these automatic text analyzers are applied to reading assessments, initial text searches, modifications, and revisions will be efficiently done and then, item writing will begin promptly. In text selection, experiences and accurate judgments on students' reading ability are heavily engaged as well. Automatic text analyses cannot be completely replaced with human experts' judgments yet, but it can

complement subjective judgments, thereby improving the efficiency of the test development process. Most importantly, future research on employing text analyzers in various testing and instructional contexts will lead us to a better understanding of the relation between text characteristics and reading ability.

REFERENCES

Alderson, C. C. (2000). *Assessing reading.* Cambridge: Cambridge University Press.

Castello, E. (2008). *Text complexity and reading comprehension tests*. Bern: Peter Lang.

Cobb, T. (n.d.). *The Compleat Lexical Tutor*. Retrieved from the World Wide Web: http://www.lextutor.ca.

Crossley, S. A., Greenfield J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly, 43*(3), 475-492.

Drum, P. A., Calfee, R. C., & Cook, L. K. (1981). The effects of surface structure variables on performance in reading comprehension tests. *Reading Research Quarterly, 16*(4), 486-514.

Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement, 11*(2), 175-193.

Fisher, D., & Frey, N. (2014). Addressing CCSSO anchor standard 10: Text complexity. *Language Arts, 91*(4), 250-250.

Fisher, D., Frey, N., & Lapp, D. (2012). *Text complexity: Raising rigor in reading*. Newark, NJ: International Reading Association.

Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: Implications for construct validity. *Language Testing, 10*(2), 133-170.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*(2), 193–202.

Gray, W. S., & Leary, B. E. (1935). *What makes a book readable?* Chicago, IL: University of Chicago Press.

IBM Corp. (2013). *IBM SPSS Statistics for Windows, Version 22.0.* Armonk, NY: IBM Corp.

Jeon, Moongee. (2015). An analysis of the continuity among middle school English textbooks with an automated language analysis program. *Modern English Education, 16*(1), 195-218.

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

Kim, Jae-Eun, & Choi, Inn-Chull. (2015). A corpus-based comparative analysis of

linguistic difficulty among high school English textbooks, EBS-CSAT prep books, and College Scholastic Ability Test. *Multimedia-Assisted Language Learning, 18*(1), 59-92.

Livingston, S. A. (2013). *Delta equating*. Princeton, NJ: Educational Testing Service.

McNamara, D. S., Graesser, A. C., & Louwerse, M. M. (2012). Sources of text difficulty: Across the ages and genres. In J. P. Sabatini & E. Albro (Eds.), *Assessing reading in the 21th century: Aligning and applying advances in the reading and measurement sciences* (pp. 89-116). Lanham, MD: R & L Education.

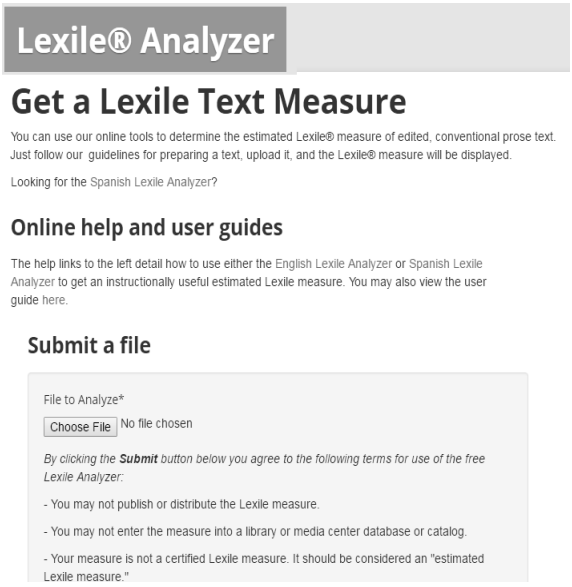
McNamara, D. S., & Kintsch, W. (1996). Learning from text: Effects of prior knowledge and text coherence. *Discourse Processes, 22*(3), 247-287.

National Governors Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects: Appendix A: Research supporting key elements of the standards and glossary of key terms.* Washington, DC: Authors. Retrieved from the World Wide Web: http://www.corestandards.org/assets/Appendix_A.pdf.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson Education, Inc.

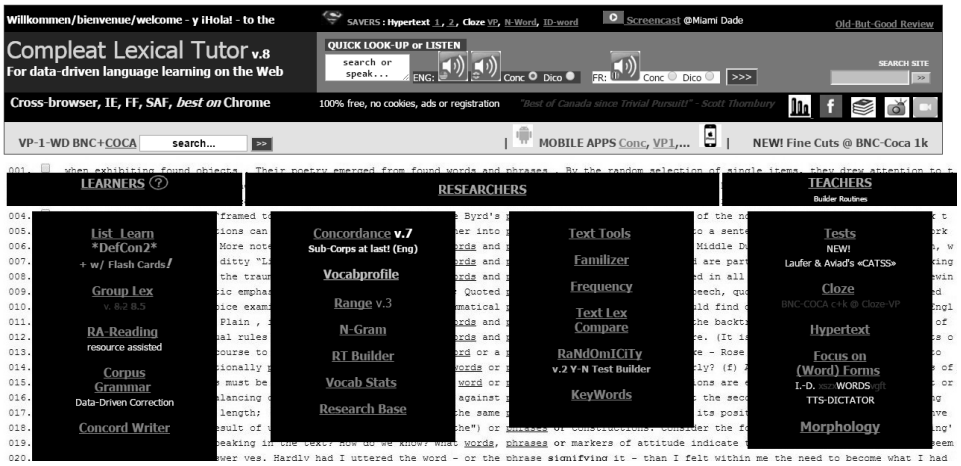
APPENDIX 1

Lexile Analyzer



APPENDIX 2

Compleat Lexical Tutor



APPENDIX 3

Coh-Metrix

예시언어(Examples in): English
적용가능 언어(Applicable Languages): English
적용가능 수준(Applicable Levels): Tertiary

Sumi Han
Department of English Language and Literature
Seoul National University
599 Gwanak-ro Gwanak-gu, Seoul, 08826, Korea
Phone: 02-880-5881
Email: sumihan20@gmail.com

Jeong-Ah Shin
Division of English Language and Literature
Dongguk University
30 Pildong-ro 1-gil Jung-gu, Seoul, 04620, Korea
Phone: 02-2260-3167
Email: jashin@dongguk.edu

Received 28 March 2016
Revised 20 April 2016
Accepted 21 May 2016