Modern English Education

http://dx.doi.org/10.18095/meeso.2018.19.4.17

A Coh-Metrix Analysis of Lexical, Syntactic and Discourse Aspects in the Newspaper Articles of Korean and British University Students*

Jiyoung Lee**

Pusan National University

ARTICLE INFO

Received 15 September 2018
Revised 24 October 2018
Accepted 10 November 2018

Examples in: English
Applicable Languages: English
Applicable Levels: Tertiary

KEYWORD

Coh-Metrix analysis/
corpus analysis/
English university newspaper
코메트릭스/
코퍼스 분석/
대학영자신문 분석

ABSTRACT

Lee, Jiyoung. (2018). A Coh-Metrix analysis of lexical, syntactic and discourse aspects in the newspaper articles of Korean and British university students. *Modern English Education, 19*(4), 17-26.

This study aims to identify differences between English newspaper articles of Korean and British university students in order to give implications for writing pedagogy. English newspaper articles from 60 Korean university students and 60 British students are analyzed. The analysis was conducted with Coh-Metrix, which is a computational textual assessment tool. The measured variables in texts were classified into lexical, syntactic and discourse aspects. The lexical aspects included lexical diversity and characteristics. The syntactic aspects refered to syntactic complexity. The discourse aspects included referential and semantic cohesion. The results are as follows. First, as for the lexical aspect, the two groups showed significant differences in lexical diversity, word frequency, familiarity, meaningfulness and polysemy. Korean students used less content words, high frequency words and familiar words, and polysemy, but more meaningful vocabulary. Second, as for the syntactic aspect, Korean students used significantly more syntactically complex patterns. Third, as for the discourse aspect, the two groups showed significant differences in content words overlap, LSA all sentences, and LSA givenness. Korean students repeat content words more than native English students. Korean students used similar and associated words in a text. These results indicate that instructors need to prepare schema activations and feedback sessions.

I. INTRODUCTION

As the importance of communicative competence on the Internet has increased, written communication in English is a prerequisite for the global competence. Whereas it is demanding for students, McNamara, Crossley and McCarthy (2010) mentioned that writing in English is critically important to succeed in various situations and professions. Emphasis has been put on teaching writing in English education in Korea in order to develop students' capability to enhance national competitiveness (E. Yoon & S. Bae, 2013).

However, teaching writing in English has been challenging to English language teachers in Korea. This is because teachers have to consider the quality of texts written by native writers in addition to appropriate linguistic knowledge. Non-native speakers need to use not only general linguistic knowledge but also specific rhetorical characteristics corresponding to the genre in order to write in the target language (H. C. Min & McCarthy, 2010).

* This paper has been developed from the presentation delivered at the 2018 MEESo International Conference.
** Author: Jiyoung Lee (Pusan National University, PhD Student)

Jiyoung Lee
Department of English Language Education, Pusan National University, 2 Pusandaehak-ro, Geumjeong-gu, Busan 46241, Korea
Tel: (051) 510-1612 / Email: jylee1915@gmail.com

Brown (2007) also mentioned that L1 and L2 writing differs in appropriate grammatical and rhetorical conventions and lexical use. In order to know appropriate grammatical and rhetorical conventions of the target language, research on comparison between native English speakers and Korean EFL learners has been conducted (Y. Choi & J. Lee, 2006; M. Jeong, 2015; M. Jeong & N. Kim, 2014; H. C. Min & McCarthy, 2010). These contrastive analyses may shed light on identifying not only the differences between L1 and EFL writing but also the general linguistic characteristics of English native writers and Korean EFL learners. These findings may help Korean EFL students to write more similar to native writers.

When contrastive analyses are conducted, the genres of texts are also considered in that linguistic characteristics of texts vary across genres. Narrative texts contain easier words, but more difficult grammatical patterns compared to social studies and science texts (McNamara, Graesser, & Louwerse, 2012). Korean EFL learners also showed different linguistic features in different genres of writing. Korean university students show more lexical and syntactic complexity, correctiveness, and fluency in narrative writings compared to argumentative writings (E. Hwang, 2013). S. Ahn (2018) also mentioned the differences between Korean students' expository and argumentative texts in terms of lexical diversity, syntactic complexity, and cohesion. Researchers have mainly focused on argumentative essays (Crossley & McNamara, 2011a; Crossley & McNamara, 2014; M. Jeong, 2015; M. Jeong & N. Kim, 2014), English language textbooks (M. Jeon, 2011; M. Jeon & I. Lim, 2009), and abstracts of journals as research materials (McCarthy, Lehenbauer, Hall, Fujiwara, & McNamara, 2007). However, English university newspapers have been rarely adopted as a research material in Korea. Analyses of English university newspapers may show distinctive findings from other Coh-Metrix analyses on narrative, expository and argumentative texts. Thus, an analysis on English university newspaper is needed.

The current study analyzes English university newspaper articles. The computational tool, Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004) is used with 13 selected indices based on a quantitative analysis. Lexical sophistication is measured with lexical diversity, age of acquisition, word frequency, familiarity, concreteness, meaningfulness, and polysemy. Syntactic complexity is measured with the mean numbers of words before the main verb and modifiers for NP. Cohesion is measured with argument overlap, content word overlap, Latent Semantic Analysis (LSA) for all sentence pairs and givenness.

The purpose of this study is to evaluate the degree to which Korean-English texts differ from the text of their British counterparts. The aim in identifying such differences is to provide lexical, syntactic, and discourse information that may facilitate non-native English speakers' writing to be closer to the models of native English speakers. To accomplish this purpose, two corpora of texts written by British university students and Korean university students are analyzed.

The specific research questions are as follows:

1) Are there any differences in descriptive statistics in the newspaper articles of Korean and British university students?
2) Are there any different lexical, syntactic, and discourse aspects in the newspaper articles of Korean and British university students?
3) Do lexical, syntactic and discourse aspects of Korean and British university students have correlations?

II. THEORETICAL FRAMEWORK AND LITERATURE REVIEW

1. Coh-Metrix Measures

Coh-Metrix is an automated tool that analyzes and assesses the characteristics of texts (Graesser et al., 2004; M. Jeong & N. Kim, 2014; M. Jeong, 2015; McNamara, Graesser, McCarthy, & Cai, 2014). The on-line version of Coh-Metrix is freely available on-line (http://tool.cohmetrix.com/). The tool provides 106 indices including text cohesion, difficulty and syntactic features and lexical characteristics (M. Jeong & N. Kim, 2014; McNamara et al., 2014). The attributes of indices that Coh-Metrix provides are explained below.

1) Descriptive Statistics

Coh-Metrix basically provides the number of paragraphs, sentences, and words that occur in a text (M. Jeong & N. Kim, 2014). It also gives information about the mean length of paragraphs and sentences. These descriptive indices help the user to interpret patterns of data (McNamara et al., 2014). If the measured value is 1, it means that the number of types is equal to tokens (M. Jeong, 2015). In other words, none of words in a text are overlapped.

2) Lexical Diversity

Type-token ratio (TTR) is a Coh-Metrix index of lexical diversity (M. Jeong & N. Kim, 2014). The term type refers to the variety of unique words, and the term token refers to the total number of words (Crossley & McNamara, 2011a; McNamara et al., 2014). Overall, the lexical diversity index considers the diversity of unique words in a text related to the total number of words.

3) Word Frequency

Word frequency indices measure how frequently words in the text are used in English language (McNamara et al., 2014). The measured values are calculated based on the CELEX corpus (Baayen, Piepenbrock, & Gulikers, 1995).

The higher the word frequency is, the more familiar and easier the word is (M. Jeong, 2015; M. Jeon & I. Lim, 2009).

4) Word Information

Coh-Metrix provides measured values to analyze age of acquisition, concreteness, familiarity, and, imagability based on MRC Psycholinguistics Database (Coltheart, 1981). These indices are crucial for L2 lexical networks (Crossley & McNamara, 2009). Age of acquisition refers to how early particular words appear in Children's language (McNamara et al., 2014). Word imagability and concreteness refer to word associations and word abstractness respectively (Crossley & McNamara, 2011a). Word familiarity indices measure how familiar words in a text are to adults. 7-point scale is utilized. 1 means the least familiar and 7 means the most familiar. The measured values are multiplied by 100 (McNamara et al., 2014). Word polysemy refers to the number of unique senses or meanings that a word has (Parker & Riley, 2010). This value is used to demonstrate the relative ambiguity of words (Crossley & McNamara, 2011a; M. Jeong, 2015).

5) Syntactic Complexity

Coh-Metrix provides words before the main verb and modifiers of NP as syntactic complexity indices (Crossley & McNamara, 2011a; M. Jeong & N. Kim, 2014; McNamara et al., 2014). Words before the main verb refers to the mean number of words on the left of the main verb. Modifiers of NP refers to the mean number of modifiers per noun phrase (McNamara et al., 2014). The higher the measured value is, the more difficult the text is (M. Jeon & I. Lim, 2009; M. Jeong, 2015).

6) Referential Cohesion

Referential cohesion indices measure cohesiveness within adjacent sentences or in all sentence pairs (McNamara et al., 2014). The all sentence pairs index includes the overlap between each sentence and all other sentences in a whole text. The former one is a local index, but the latter one shows a more global index (McNamara, Graesser, & Louwerse, 2012). Argument overlap considers the overlap of nouns, pronouns or noun phrases (McNamara et al., 2014). Content word overlap measures the proportion of content words that are overlapped (McNamara et al., 2014).

7) Latent Semantic Analysis (LSA)

Latent Semantic Analysis is a mathematical statistical technique which gives measures of semantic coreferentiality (Crossley & McNamara, 2011a). The system is based on a large corpus which represents world knowledge (McNamara et al., 2014). McNamara and his colleagues (2014) also mentioned that LSA in Coh-Metrix indicates text coherence and difficulty. However, it has been argued that LSA measures text cohesion not coherence in that LSA investigates similarities between adjacent sentences or all possible pairs of sentences in the text by finding explicit words of the text base.

2. Research on Coh-Metrix Analysis

Coh-Metrix has been utilized as a textual assessment tool in the English education field (Crossley & McNamara, 2014; M. Jeon, 2011; M. Jeon & I. Lim, 2009; S. Kim & M. Jeon, 2016; McNamara et al., 2010). By analyzing texts with Coh-Metrix, accuracy and reliability may be highly improved. The researchers selected Coh-Metrix indices based on their own purposes. The selected indices could measure either one linguistic aspect or more than two aspects.

There are relatively fewer numbers of Coh-Metrix analyses on one linguistic aspect. Crossley and McNamara (2014) examined syntactic features of L2 writers such as syntactic complexity and syntactic density using Coh-Metirx. The research aimed to demonstrate development of English writing. As a result of the study, L2 writers showed more noun and phrasal complexity. Crossley and McNamara (2009) investigated lexical differences in L1 and L2 argumentative essays. L2 writers tend to be lexically less proficient than L1 writers with higher values of word frequency, and age of acquisition. However, we cannot know other linguistic aspects of text written by those subjects.

If materials are analyzed with indices across various linguistic features, more comprehensive findings from Coh-Metrix analyses would be identified. Coh-Metrix inherently offers multiple levels of language and discourse as it follows the nature of language (Graesser et al., 2004). If we know linguistic factors of L2 writers including lexical sophistication, syntactic complexity, and cohesive devices, we may deeply understand the unique characteristics of L2 writing (Cumming, 2001). L2 writers need to utilize L2 lexical knowledge and syntactic structures, and apply these linguistic elements to a coherent text in order to write in English successfully (Crossley & McNamara, 2011a). The current study selects 13 indices of lexical, syntactic and discourse features to get comprehensive linguistic characteristics of two groups.

3. Research on the Differences Between L1 and L2 Writing

L1 writers and L2 or EFL writers may have clear differences in the convention of written forms and rhetorical choice. Peregoy, Boyle, and Kaplan (2013) mentioned that nonnative writers may have some limitation in expressive abilities in terms of sentence structure and vocabulary, but they may not have the sense of conventional written form in English. In order to explore these differences between native and nonnative writers, contrastive rhetoric and con-

trastive corpus analysis have been used.

Contrastive rhetoric laid the foundation for perspectives on second language or EFL writing and writing pedagogy. Kaplan (1966) suggested that different language families have conventionally different structures of texts. For example, English was described as a straight line to indicate its straightforward rhetorical patterns. Unlike English texts, Korean texts are indirect because Koreans put main points at the end of texts. K. Kim (1997) compared the rhetorical patterns of Korean-Korean, Korean-English and American-English editorials in campus newspapers. The study showed that Korean students transferred their L1 rhetorical style into English writing. On the other hand, Korean college students did not show the typical Korean rhetoric patterns in English composition (H. Ryu, 2006). As K. Kim (1997) and H. Ryu (2006) mentioned, they adopted human rating which has reliability and objectivity problems.

A lot of L2 and EFL writing analyses have been conducted, compared to native English speakers' writing (Crossley & McNamara, 2009; Crossley & McNamara, 2011b; M. Jeong, 2015; M. Jeong & N. Kim, 2014; McCarthy et al., 2007). Contrastive corpus analyses have generally adopted argumentative essays (Crossley & McNamara, 2011b; M. Jeong, 2015; M. Jeong & N. Kim, 2014). Crossley and McNamara (2011b) analyzed argumentative essays written by both L1 and L2 writers from earlier studies. The study suggested the relationship between writing quality and linguistic aspects with human ratings. As for lexical sophistication, more lexical diversity is significantly related to increasing scores while decreasing word frequency, familiarity and meaningfulness are negatively correlated to essay scores. Syntactic complexity is positively correlated to essay scores. More cohesive texts are assessed as lower quality.

Researchers also adopted abstracts of scientific journals as research materials. McCarthy et al. (2007) examined journal abstracts of Japanese, American, and British scientists using Coh-Metrix. As a result of the analysis, Japanese scientists significantly used more locational items, temporal items, high frequency words, and high familiarity words than native English speakers. H. C. Min and McCarthy (2010) analyzed abstracts from experimental scientific journals in order to find the differences between American and Korean scientists in terms of discourse styles. To analyze textual patterns of abstracts, they adopted the Gramulator, which is a textual assessment tool to identify underlying textual patterns. As a result, Korean scientists appear to use acceptable but non proto-typical discourse style.

L2 language learners' L1 can influence their L2 written production (Crossley & McNamara, 2011a). University students tend to depend on L1 in English writing (Y. Choi & J. Lee, 2006). It shows that the linguistic characteristics affect English writing. In addition, quantitative research has shown remarkable differences among different genres or registers in usage of vocabulary and grammar (Kennedy, 2014). Y. Ryu (2010) also mentioned that English university newspapers have rarely been adopted as a research material. Thus, English university newspapers need to be analyzed. This study analyzes English newspaper articles written by Korean and British university students.

III. METHOD

1. Corpus Collection

Two corpora were selected: a corpus of newspaper articles written in English by native Korean speakers (KE) and a matching L1 corpus written in English by native English speakers (BE). The Korean students The KE corpus (n = 60) was taken from three universities in Busan, Korea: The Hyowon Herald, The Pukyong Herald, and Aranuri. The BE corpus (n = 60) was collected from three university web pages on the Tab (https://thetab.com/): Cambridge, Oxford, and University of Portsmouth. The website provides English university newspapers of over 80 universities in Britain and America. All the newspaper articles in this study were written by different students. The two corpora consist of news articles excluding editorials. All published since 2015. Table 1 presents brief information of the two corpora.

TABLE 1
Information of Corpora

Variable	Number of articles	Number of words
Korean university students	60	28,662
British university students	60	28,543
Total	120	57,205

2. Selected Coh-Metrix Indices

From 106 indices, 13 indices are extracted and then the results are restored with their serial numbers. Table 2 shows linguistic aspects, measured variables, and Coh-Metrix indices.

TABLE 2
Coh-Metrix Analysis Items

	Measured variable	Coh-Metrix index
Lexical aspects	Lexical diversity	Type-token ratio
	Word information	Frequency
		Age of acquisition
		Familiarity
		Concreteness
		Meaningfulness
		Polysemy
Syntactic aspects	Syntactic complexity	Words before the main verb
		Modifiers for NP
Discourse aspects	Referential cohesion	Argument overlap in the whole text
		Content word overlap in the whole text
	Semantic cohesion	LSA for all sentence pairs
		LSA given-new

This study aims at lexical, syntactic, and discourse aspects of newspaper articles. For measuring lexical aspects, lexical diversity and word information are considered as measured variables. The Coh-Metrix index of lexical diversity refers to type-token ratio. Word information includes six Coh-Metrix indices such as word frequency, age of acquisition, word familiarity, concreteness, meaningfulness, and polysemy in this study. For measuring syntactic aspects, syntactic complexity is considered as a measured variable. Syntactic complexity includes two Coh-Metrix indices such as words before the main verb and modifiers for NP in this study. For measuring discourse aspects, referential and semantic cohesion are considered as measured variables. Referential cohesion includes two Coh-Metrix indices such as argument overlap in the whole text and content word overlap in the whole text in this study. Semantic cohesion is measured by Latent Semantic Analysis (LSA). Semantic cohesion includes two indices such as LSA for all sentence pairs and LSA given-new in this study.

3. Statistical Analysis

The stored results are analyzed using an open-source software R 3.4.4. An independent two sample *t*-test was conducted using the selected 13 Coh-Metrix indices as the dependent variables and the Korean and British articles as the independent variables. To demonstrate correlations among three linguistic aspects, Pearson's product moment correlation was used.

IV. RESULTS AND DISCUSSION

1. Differences in Descriptive Statistics

Research question 1 considers whether differences in descriptive statistics in newspaper articles of Korean and British university students exist. In order to investigate differences in descriptive statistics between two groups, independent two sample *t*-test was conducted. Table 3 shows the descriptive statistics for 120 newspaper articles of Korean and British university students.

TABLE 3
Descriptive Statistics

Variable	Korean (n = 60) *M (SD)*	British (n = 60) *M (SD)*	*t*	*df*	*p*
Number of words	477.70 (208.11)	475.71 (278.82)	.04	109.17	.965
Number of sentences	28.15 (14.73)	23.76 (13.86)	1.68	117.57	.096
Number of paragraphs	5.52 (2.45)	8.92 (3.97)	-5.64	98.30	.000
Number of words per sentence	17.93 (3.18)	20.37 (4.08)	-3.64	111.39	.000
Number of sentences per paragraph	5.18 (2.08)	2.66 (1.27)	7.99	97.69	.000

The number of words and sentences did not show significant differences. The results indicate that text length of Korean and British newspaper articles is similar. However, the average number of paragraphs (t = -5.64, p < .001), words per sentence (t = -3.64, p < .001), and sentences per paragraph (t = 7.99, p < .001) is significantly different between two groups.

The two groups do not show significant differences in the text difficulty in that M. Joen and I. Lim (2009) suggested that the text that contains more words and sentences is more difficult. The results suggest that Korean students use more sentences in a paragraph than British students. That is why texts written by Korean students contain fewer paragraphs. The sentence length of British students is longer than one of Korean students.

2. Differences in Lexical, Syntactic and Discourse Aspects

Research question 2 investigates whether differences in lexical, syntactic, and discourse aspects in newspaper articles of Korean and British university students exist. In order to examine differences in three linguistic aspects between two groups, independent two sample *t*-test was conducted.

1) Lexical Aspects

Table 4 presents the results of the *t*-test for lexical aspects of Korean and British newspaper articles. There were statistically significant differences between two groups for lexical diversity (t = -6.05, p < .001), word frequency (t = -2.58, p < .01), word familiarity (t = -3.90, p < .001), word meaningfulness (t = 2.23, p < .05), and word polysemy (t = -3.49, p < .01).

TABLE 4
The *t*-Test Results for Lexical Aspects of Korean and British Articles

Variable	Korean (n = 60) *M (SD)*	British (n = 60) *M (SD)*	*t*	*df*	*p*
Lexical diversity	.69 (.07)	.77 (.07)	-6.05	118	.000
Word frequency	1.04 (.46)	1.26 (.46)	-2.58	118	.011
Age of acquisition	370.28 (31.39)	360.66 (35.46)	1.57	116	.118
Word familiarity	568.52 (9.60)	574.42 (6.72)	-3.90	106	.000
Word concreteness	389.71 (20.81)	381.95 (22.27)	1.97	117	.051
Word meaningfulness	435.18 (14.59)	429.71 (12.12)	2.23	114	.027
Word polysemy	3.70 (.37)	3.94 (.39)	-3.49	118	.001

The Korean university students used significantly less numbers of words in their newspaper articles. It means that the Korean students repeated the same words in the text more than the British students. Crossley and McNamara (2011b) showed that decreasing lexical diversity is associated with increasing quality of writing in English.

M. Jeong and N. Kim (2014) also suggested that Korean EFL learners need to learn how to utilize a variety of expressions.

The Korean students used less high frequency words and familiar words. Korean students (568.52) are lower than the average of word familiarity index (569). In other words, the Korean students used more difficult words and less everyday English expressions. The results contradict the Coh-Metrix analyses on argumentative essays (M. Jeong, 2015; M. Jeong & N. Kim, 2014). Korean university students use more high frequency words and Everyday English expressions in English argumentative essays. It indicates that the different linguistic characteristics exist among different genres of texts. The Korean students could use professional and difficult words in English newspaper articles because they cover and study specific stories with enough time.

The Korean students use more meaningful words and less polysemy. Two indices indicate clarity and ambiguity of lexical meanings. In that regard, the newspaper articles of the Korean students are easier to understand.

2) Syntactic Aspects

Table 5 demonstrates the results of the *t*-test for syntactic aspects of Korean and British newspaper articles. There were statistically significant differences between two groups for both words before the main verb ($t = 2.75$, $p < .01$) and modifiers for NP ($t = 2.66$, $p < .01$).

TABLE 5

The *t*-Test Results for Syntactic Aspects of Korean and British Articles

Variable	Korean ($n = 60$) *M (SD)*	British ($n = 60$) *M (SD)*	*t*	*df*	*p*
Words before the main verb	5.00 (1.34)	4.21(1.80)	2.75	108.87	.006
Modifiers for NP	1.01 (0.20)	.92 (0.17)	2.66	116	.008

The Korean newspaper articles contain more words before the main verb, and more modifiers for NP. Example sentences in Extract (1) are extracted from newspaper articles of Korean university students.

Extract (1)

(a) Professor Kwon Han-sang, who looks at technology prospects positively said, "Technology is a living organism. (The Pukyong Herald 2017, July)

(b) On November 11th, Ms. Peru beauty contest, which originally only estimated women's appearance, has reported femicide in Peru. (The Hyowon Herald 2017, December 1)

(c) The emergence of Time Commerce apps, the growth of social activity platforms, and content creators are examples of YOLO being on the rise for consumers. (The Hyowon Herald 2018, March 3)

In Extract (1), Korean students use complex noun phrases such as relative clauses and the conjunction and to give exact information about people or events. Example sentences in Extract (2) are extracted from newspaper articles of British university students.

Extract (2)

(a) Musty Kamal, a fresher, had been elected to the committee with 153 votes – 40 more than his next closest rival – making him the most popular candidate amongst the 17 that stood, and 11 that were elected. (Oxford University 2017, December 12)

(b) Bryan Cranston does not have time for the rampant misogyny of the film industry – he wants, in short, a fresh start for Hollywood and show business. (Cambridge University 2018, January 21)

(c) A Facebook event was created by King's theology student, Harriet Fisher, and quickly gained momentum. (Cambridge University 2018, January 20)

In Extract (2), unlike Korean students, British students use simple noun phrases by using proper nouns, pronouns and ellipsis. Crossley and McNamara (2011b) demonstrated that syntactically complex texts have better quality. The results contradict the Coh-Metrix analyses on argumentative essays (M. Jeong, 2015). The native speakers' essays contain more words before the main verb and modifiers for NP. The Korean students could use relatively complex structures in English newspaper articles since they can get feedback from their peers and native English professors.

3) Discourse Aspects

Table 6 shows the results of the *t*-test for discourse aspects of Korean and British newspaper articles. There were statistically significant differences between two groups for content overlap ($t = -6.05$, $p < .001$), LSA all sentences ($t = -2.58$, $p < .01$), and LSA givenness ($t = -3.49$, $p < .01$).

TABLE 6

The *t*-Test Results for Discourse Aspects of Korean and British Articles

Variable	Korean ($n = 60$) *M (SD)*	British ($n = 60$) *M (SD)*	*t*	*df*	*p*
Argument overlap	.45 (.14)	.42 (.15)	1.01	118	.314
Content word overlap	.08 (.03)	.06 (.03)	2.65	118	.009
LSA all sentences	.19 (.06)	.12 (.08)	5.02	118	.000
LSA givenness	.31 (.04)	.26 (.04)	6.91	111	.000

Overall, the Korean students' newspaper articles are more cohesive. The results are similar with the results in argumentative essays. M. Jeong (2015) showed that significantly higher values results in content word overlap

and LSA all sentences of the Korean students. Crossley and McNamara (2011b) indicated that the less cohesive a text is, the better scores the student get.

3. Correlations among Lexical, Syntactic, and Discourse Aspects

Research question 3 asks whether lexical, syntactic, and discourse aspects of Korean and British university students have correlations. In order to investigate correlations among three linguistic aspects, Pearson's product moment correlation was conducted.

1) Correlations Between the Lexical Indices and the Syntactic Indices

Correlations were investigated between the lexical indices and the syntactic indices for the 120 newspaper articles. Nine indices demonstrated significant correlations between the lexical and the syntactic indices (see Table 7).

TABLE 7
Correlations Between Lexical Aspects and Syntactic Aspects (n = 120)

Variable	Words before the main verb	Modifiers for NP
Lexical diversity	.002	.011
Word frequency	-.289**	-.300***
Age of acquisition	.242**	.172
Word familiarity	-.175	-.396***
Word concreteness	.177	.330***
Word meaningfulness	.273**	.241**
Word polysemy	-.345***	-.229*

*p < .05, **p < .01, ***p < .001

Word frequency has significant negative correlations with words before the verbs (r = -.289, p < .01) and modifiers for NP (r = -.300, p < .001). A negative correlation exists between word familiarity and modifiers for NP (r = -.396, p < .001). It means that the higher frequency words students use and the more familiar words, the less complex syntactic patterns the students use. High frequency and familiar words are easier to use. In that regard, less high frequency words and more complex syntax indicate one's linguistic ability.

Word meaningfulness has significant correlations with words before the verbs (r = .273, p < .01) and modifiers for NP (r = -.241, p < .01). Word polysemy has significant correlations with words before the verbs (r = -.345, p < .001) and modifiers for NP (r = -.229, p < .05). A negative correlation exists between word concreteness and modifiers for NP (r = -.330, p < .001).

Korean students tend to use words that are higher age of acquisition. Age of acquisition is significantly correlated to syntactic complexity. M. Jeong (2015) mentioned that age of acquisition is highly related to syntactic aspects that affect text difficulty. However, the result is contradictory to the result from the other research (M. Jeong, 2015) in that Korean university students use words that are lower age of acquisition and simple syntactic structure in their argumentative essays.

Korean students use less familiar words, which are more difficult words. This word characteristic is significantly correlated to the number of modifiers for NP. It can suggest that students who use difficult words can make complex sentence structures.

2) Correlations Between the Syntactic Indices and the Cohesive Indices

Correlations were investigated between the syntactic indices and cohesive indices. Four indices demonstrated significant correlations (see Table 8).

TABLE 8
Correlations Between Syntactic Aspects and Discourse Aspects (n = 120)

Variable	Words before the main verb	Modifiers for NP
Argument overlap	.377***	.128
Content word overlap	.302***	.108
LSA all sentences	.345***	.232*
LSA givenness	.136	.150

*p < .05, ***p < .001

The mean number of words before the main verb has positive correlations with argument overlap (r = .377, p < .001), content word overlap (r = .302, p < .001), and LSA all sentences (r = .345, p < .001). A correlation also exists between modifiers for NP and LSA all sentences (r = .232, p < .05). The results indicate that the more students write words before the main verb, the more cohesive the text is.

Korean students show higher measured values on referential cohesion, which seems to make texts weak. These discourse indices are significantly correlated to words before the main verb. Thus, students could avoid too much overlap by reducing syntactic complexity.

3) Correlations Between the Cohesive Indices and the Lexical Indices

Correlations were investigated between the cohesive indices and the lexical indices for the 120 newspaper articles. Nine indices demonstrated significant correlations between the lexical and the syntactic indices (see Table 9).

Lexical diversity has negative correlations with content words overlap (r = -.461, p < .001), LSA all sentences (r = -.437, p < .001) and LSA givenness (r = -758, p < .001). Words frequency has negative correlations with content word overlap (r = -.296, p < .01) and LSA all sentences (r = -.239, p < .01). Word familiarity has negative correlations with LSA all sentences (r = -.201, p < .05) and LSA givenness (r = -.210, p < .05). It indicates that the more students use unique expressions, high frequency words, and familiar vocabulary, the less cohesive the newspaper articles are.

TABLE 9
Correlations Between Syntactic Aspects and Discourse Aspects (n = 120)

Variable	Argument overlap	Content word overlap	LSA all sentences	LSA givenness
Lexical diversity	-.145	-.461***	-.437***	-.758***
Word frequency	-.283**	-.296**	-.239**	-.169
Age of acquisition	.134	.098	.249**	.146
Word familiarity	-.021	-.063	-.201*	-.219*
Word concreteness	.080	.098	.224*	.025
Word meaningfulness	.152	.074	.301***	.038
Word polysemy	.058	.064	-.138	-.020

*p < .05, **p < .01, ***p < .001

Age of acquisition has a positive correlation with LSA all sentences (r = .249, p < .01). Word concreteness has a positive correlations with LSA all sentences (r = .224, p < .05). Word meaningfulness has a positive correlations with LSA all sentences (r = .301, p < .001).

The remarkable result is that lexical diversity has significant negative correlations with referential and semantic and semantic cohesion, which make readers to think the text does not contain enough information. In this regard, students need to utilize a variety of expressions in order to make newspaper article richer.

V. CONCLUSION

The current study aims to identify different lexical, syntactic, and discourse aspects between native and non-native writers in order to give implications for writing pedagogy. For this study, English newspaper articles of Korean and British university students were analyzed based on 13 selected indices that a computational textual assessment tool, Coh-Metrix, provides.

The first research question is whether differences in descriptive statistics in newspaper between two groups exist. The second research question is whether lexical, syntactic, and discourse differences between two groups exist. The third research question is whether lexical, syntactic, and discourse aspects of Korean and British university students have correlations. The results of this study are as follows.

In the descriptive statistics, the number of words and sentences in a text is similar between two groups, but the sentence and paragraph lengths are significantly different. Korean students use shorter and simple sentences and longer paragraphs.

In the lexical aspects, the two groups show significant differences in lexical diversity. The Korean university students wrote English newspaper articles with relatively less various expressions. The result suggests that the Korean students may have a lack of vocabulary knowledge. The two groups also show significant differences in word frequency and familiarity. Namely, the Korean university students used less high frequency and familiar words. The result suggests distinctive features in the newspaper genre. Newspaper articles need more specific terms to describe special events.

In the syntactic aspects, the two groups demonstrate significant differences in syntactic complexity. The Korean students use more complex syntactic patterns in newspaper articles. The results are contradictory to the results of argumentative essays. It also indicates the distinctive features of this genre. Korean EFL students can get feedback from their peers and professors through editing.

In the discourse aspects, the two groups show significant differences in cohesion. The newspaper articles of Korean students are relatively more cohesive because they often overlap content words and repeat given information. The sentences in a whole text are closely related. However, more cohesive texts seem less fluent and proficient. In that regard, instructors need to inform Korean students of how to avoid excessive repetition of content words.

In the correlations among three linguistic aspects, cohesion has negative correlations with three indices of lexical sophistication such as lexical diversity, word frequency and familiarity. The results indicate that lexical characteristics are significantly associated with discourse aspects. It would be difficult to enable students to understand the notion of cohesiveness. Instructors could explain cohesiveness with the relationship between lexical sophistication and cohesion.

The educational implications are as follows. As for the lexical aspects, it is crucial to spend much time to cover specific stories in order to write better English articles or texts. Korean university students used less various words than British university students. Thus, language teachers should give students enough activities to think about expressions related to their topics.

As for the syntactic aspects, newspaper articles of Korean university students show more complex syntactic patterns unlike other results from Coh-Metrix analyses on other genres such as expository or argumentative essays. When Korean reporters publish English newspaper, they get feedback from other reporters and professors. Thus, it is important to get feedback from peers and instructors in order to use appropriate syntactic structure.

As for the discourse aspects, Korean EFL students need to avoid the repetition of the same vocabulary and content. Using same expressions and expressing the same content over and over again disrupt the text richness. Korean students tend to repeat the same words and content in a text compared to native speakers. Thus, instructors need to prepare various activities in order to enable learners to activate related knowledge and various English synonyms.

The limitation of this study is related to generalizations. First, this study refers to British university students as native writers or native speakers. However, McCarthy and his colleagues (2007) suggested that there are linguistically significant differences between American and British people. Further studies need to compare and contrast campus newspapers among Korean, American, and British students. Second, this study aimed at the published articles of English university newspapers. Therefore, the results

are limited to English university newspaper reporters. Further studies could aim at Korean EFL learners who are not reporters of English newspapers. Third, the newspaper articles of Korean students were selected from one city, Busan. Further studies would select newspaper articles from universities which are located in various areas.

In conclusion, this study seems to have implications for researches on English education in that this study analyzed English university newspaper articles, which have been rarely adopted as a research material. This study also tried to highlight the comprehensive linguistic differences between native and non-native writers.

REFERENCES

Ahn, Soojin. (2018). An analysis of Coh-Metrix on the differences in English expository and argumentative writing of Korean and native English university students. *New Korean Journal of English Language and Literature, 60*(3), 177-205.

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

Brown, H. D. (2007). *Teaching by principles: An interactive approach to language pedagogy* (3rd ed.). White Plains, NY: Pearson Education.

Choi, Yeonhee, & Lee, Jieun. (2006). L1 use in L2 writing process of Korean EFL students. *English Teaching, 61*(1), 205-225.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology, 33*(4), 497-505.

Crossley, S. A., & McNamara, D. S. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing, 18*(2), 119-135.

Crossley, S. A., & McNamara, D. S. (2011a). Shared features of L2 writing: Intergroup homogeneity and text classification. *Journal of Second Language Writing, 20*(4), 271-285.

Crossley, S. A., & McNamara, D. S. (2011b). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life Long Learning, 21*(2-3), 170-191.

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing, 26*, 66-79.

Cumming, A. (2001). Learning to write in a second language: Two decades of research. *International Journal of English Studies, 1*(2), 1-23.

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

Hwang, Eunkyung. (2013). *Korean EFL college learners' linguistic features in narrative and argumentative writings in terms of CAF* (Unpublished doctoral dissertation). Sookmyung Women's University, Seoul.

Jeon, Moongee. (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 Linguistics Science, 56*(1), 201-218.

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.

Jeong, Mikyung. (2015). *A comparative study on the English argumentative essays of Korean and native English university students with Coh-Metrix : Focused on lexical, syntactic, and discourse features* (Unpublished doctoral dissertation). Kangwon National University, Chuncheon.

Jeong, Mikyung, & Kim Namgook. (2014). An analysis of the linguistic features on the corpus of Korean EFL learners and native English speakers with Coh-Metrix. *Studies in Linguistics, 33*, 373-395.

Kaplan, R. B. (1966). Cultural thought patterns in inter-cultural education. *Language Learning, 16*(1-2), 1-20.

Kennedy, G. (2014). *An introduction to corpus linguistics.* London: Routledge.

Kim, Kyeongja. (1997). A comparison of rhetorical styles in Korean and American student writing. *Intercultural Communication Studies, 6*, 115-150.

Kim, Sojung, & Jeon, Moongee. (2016). An analysis study of English writing of elementary school 6th grade English language learners using Coh-Metrix. *Modern English Education, 17*(3), 263-287.

McCarthy, P. M., Lehenbauer, B. M., Hall, C., Fujiwara, Y., & McNamara, D. S. (2007). A Coh-Metrix analysis of discourse variation in the texts of Japanese, American, and British scientists. *Foreign Languages for Specific Purposes, 6*, 46-77.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication, 27*(1), 57-86.

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 21st century: Aligning and applying advances in the reading and measurement sciences* (pp. 89–116). Lanham, MD: R&L Education.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix.* New York: Cambridge Press.

Min, Hyunsoon. C., & McCarthy, P. M. (2010, May). *Identifying varietals in the discourse of American*

and Korean scientists: A contrastive corpus analysis using the gramulator. Paper presented at the Twenty-Third International Florida Artificial Intelligence Research Society Conference, Daytona Beach, FL.

Parker, F., & Riley, K. (2010). *Linguistics for non-linguists: A primer with exercises.* Boston: Pearson.

Peregoy, S. F., Boyle, O., & Cadiero-Kaplan, K. (2013). *Reading, writing, and learning ESL: A resource book for teaching K-12 English learners* (6th ed.). Upper Saddle River, NJ: Pearson.

Ryu, Hoyeol. (2006). Rhetorical patterns in Korean college students' English expository writings. *English Teaching, 61*(3), 273-292.

Ryu, Youngmi. (2010). *A study on frequency of modals based on corpus analysis on English newspapers of Korean universities* (Unpublished master's thesis). Sookmyung Women's University, Seoul.

Yoon, Eugene., & Bae, Sanghoon. (2013). An analysis on variables affecting English writing self-efficacy and ability. *Foreign Languages Education, 20*(2), 135-162.