PEOPLE: International Journal of Social Sciences
ISSN 2454-5899

Bomi Lee, 2019

Volume 5 Issue 3, pp. 120-132

Date of Publication: 21st November 2019

DOI- https://dx.doi.org/10.20319/pijss.2019.53.120132

This paper can be cited as: Lee, B. (2019). Analysis on Lexical Density and Sophistication in Korean Learners' Spoken Language. PEOPLE: International Journal of Social Sciences, 5(3), 120-132.

ANALYSIS ON LEXICAL DENSITY AND SOPHISTICATION IN KOREAN LEARNERS' SPOKEN LANGUAGE

Bomi Lee
Department of Korean Language and Literature, Yonsei University, Seoul, Republic of Korea
bomi.spring6642@gmail.com

Abstract

The proficiency level of second language development can be judged by the learner's output and be used to estimate the learner's stage of learning. The linguistic production can be used to see how a learner's language knowledge has been internalized. In this case, factors such as accuracy, fluency, and complexity can be considered. In this study, lexical density and sophistication were analyzed in Korean learners' spoken production as an aspect of measuring lexical complexity. The lexical density and sophistication scores were obtained from 44 students' spoken data whose language proficiency varied from intermediate (level 3 and 4) to advanced (level 5 and 6). Also, using one-way ANOVA, statistical analysis was conducted on the differences according to the proficiency of each level. As a result, the lexical density score showed little difference according to the proficiency level, however the difference was statistically not significant. However, in the case of lexical sophistication score, the scores of advanced students were higher the intermediate students. The level of lexical sophistication corresponded with the proficiency level increased, and the differences were statistically significant.

Keywords

Korean Language Education, Interlanguage, Lexical Density, Lexical Sophistication

1. Introduction

Second language learners make various mistakes and errors in the language learning process. During the process, they get close to the target language by going through the interlanguage stage. The learners' output (such as writing and speaking) in their second language shows various information: The learner's language level, in which stage of learning the learner is placed at, where the errors occur the most, and which among them are fossilized. In other words, analyzing and studying the learner's language production is the best way to describe learners' language development (Wolfe-Quintero, Inagaki, and Kim, 1998; Ortega, 2003; Nam, 2015).

Accuracy, fluency, and complexity are considered as factors to determine when analyzing the language productions to assess to which degree the language has been internalized. Assessment on how accurately the language can be used without errors, and how fluently it can be used without hesitation or pause, are important evaluation factors. At the same time, also, complexity is an important factor in measuring learners' language development. The complexity is the ability to use refined and complex language structures like a native speaker.

The complexity not only includes lexical complexity but also syntactic complexity in order to see how complicated the sentence is structured. *Lexical diversity*, *lexical density*, and *lexical sophistication* are widely studied as a measure for lexical complexity. According to Read (2000), this lexical richness (lexical complexity in this paper) demonstrates the effective use of vocabulary in the learner's output. To explain the concept of each term, *lexical diversity* is a method of analyzing how diverse and rich vocabulary is being used through the learner's output. It is a measure of how many new lexical types are produced out of the total lexical tokens. Therefore, the higher the learner's use of various vocabulary, the higher the measure. *Lexical density* is a method of measuring the ratio of how much content words are used in the overall vocabulary produced by the learner. In other words, lexical density measures how well learners can convey practical meanings. If a learner uses more content words than functional words, which have a grammatical role, it is considered to use a richer vocabulary.

Finally, *lexical sophistication* is a way of analyzing how learners use 'advanced' vocabulary. Read (2000) defined lexical sophistication as 'the proportion of lexical that is relatively less common and higher in the learner's language'. In other words, it is to see how much more professional and less frequent lexical is available in learners' production. If a learner produces a lot of these low frequency vocabulary, it can be said that language proficiency is high. These three measurement methods allow you to see the lexical complexity of the learner.

Many previous studies analyzing lexical complexity focus on lexical diversity. In the case of lexical diversity, the formulas have also been developed in various ways. As well as the most commonly used Token-Type Ratio (TTR), there are various such as mean segmental TTR, G-Guiraud Index, and D optimum average. Basically, however, all the measurement methods mentioned above look at the ratio with the number of lexical type and token. Most of the studies in KFL, which analyzed vocabulary diversity, were conducted with students' writing data (An, 2003; Bae, 2012; Won et al, 2017; Lee, 2017 and many more). There was also a study of learners' spoken data, and Kim (2012) observed the use of vocabulary over six months for women's marriage immigrants. Nam and Kim (2014) conducted some task to Korean learners and calculated the lexical diversity. The results of the lexical diversity data were compared with results of native speaker's lexical diversity. As a result, it was found that in the upper intermediate level, it developed to a degree similar to that of the Korean native speaker.

Unlike lexical diversity research, which has been widely applied and studied, lexical density and lexical sophistication have not been studied so many yet. There are few studies on lexical density and sophistication in Korean. First, in the case of Ahn (2017), the output of the learner was not analyzed, but the lexical diversity and density were measured using the Korean text and the native Korean speaker. As a result, it was found that the lexical density all genre of spoken data was lower than the average of written text. Lecture was only 50.5, which was similar to the average written data of 52.1. In addition, except for purchase conversation and class conversation, all other genres had a more than 40% of the lexical density. Although the result is relatively lower than the written data, it is showed that the density is somewhat higher in the spoken language unlike the English.

Won et al. (2017) analyzed the lexical diversity, density, and sophistication for intermediate and advanced Korean learners, and Kang and Jin (2017) analyzed sophistication for beginner Korean learners to study whether the learner's lexical richness can be used as an indicator for writing ability. As a result of Won et al. (2017) analyzing how the lexical richness changes as learners' proficiency is increased, it was confirmed that the vocabulary diversity does not change significantly even if the proficiency is increased. On the other hand, there was a significant difference in lexical density between learners' grades. Also, the result of analyzing the correlation between lexical richness and writing score revealed that diversity and sophistication were not variables that greatly influence writing score, but only vocabulary density affected to the writing score. Lastly, Nam (2015) is the only study that analyzes the lexical density and sophistication of

Korean learners using spoken data. In order to study the spoken lexical complexity of Korean learners, Nam (2015) analyzed the lexical sophistication. As a result, it was found to increase significantly from the beginner to the intermediate and advanced level. Therefore, this study suggests that it can be used as a useful indicator to distinguish proficiency of foreign learners.

As mentioned above, the study of lexical complexity of Korean language education is mainly focused on lexical diversity, principally using learners' written texts. In recent years, researches on lexical diversity in learners' spoken language data have been conducted, but studies on lexical density and lexical sophistication are still insufficient. Consequently, in this study, efforts were made to measure the learners' complexity using lexical density and lexical sophistication through learners' spoken data. Detailed research questions are as follows.

- Does the lexical density differ by the level of the learners?
- Does the lexical sophistication differ by the level of the learners?

2. Methodology of the Study

2.1 The Objective of the Study and its Procedures

In order to analyze the lexical complexity of the spoken output of Korean learners, this paper used some of the learners' corpus provided by *Korean Learners' Corpus Search Engine* – National Institute of Korean Language (https://kcorpus.korean.go.kr/index/goMain.do). This corpus includes 81 L1 learners, such as English, Chinese, Japanese, Spanish, etc., and the number of learners according to L1 is very different from learners' level. It is because the majority of Korean learners' first language is Chinese, and learners whose first language is Arabic, Hebrew, Norwegian, Polish, etc. are a few. On account of imbalance of L1, in this study, the sample was limited to learners whose L1 is Chinese. In addition, this corpus includes various learners' level from level 1 (beginner) to level 6 (advanced), or higher (highest level). Level 1 and level 2 are beginner levels, and their spoken data length is very short. It is composed of very simple and easy vocabulary, which makes it difficult to analyze lexical sophistication. Therefore, this study was aimed to analyze intermediate (level 3, 4) and advanced (level 5, 6) learners excluding the beginners.

The learners' spoken corpus constructed by 2017 includes about 700 samples, of which about 160 samples are Chinese intermediate and advanced level learners. However, within the 160 samples, a significant amount of data were conversations among multiple speakers such as an

interview or debate, typically including teachers. In order to analyze a learner's lexical complexity, data consisting of only one speaker (i.e. learner) rather than a sample composed of several speakers was needed. Thus, only the 'monologue' genre was extracted. As a result, analyzed data from the subjected corpus is as followed in Table 1.

Table 1: *Information of the Extracted Corpus*

Level of Learner	Number of Samples	Total Number of Words	Average Number of Words
Level 3	30	6,953	231.8
Level 4	18	4,287	238.2
Level 4	17	12,740	749.4
Level 6	11	5,525	502.3
Total	76	29,505	388.2

Even within the extracted data, the difference in length of the words was very large. The shortest sample from level 3 contains 98 words, and the longest sample from level 5 contains 1,735 words. Since the denominator is the number of word types in the measurement of lexical density and sophistication, the measured value got smaller as the length of the word got longer. Hence, considering that most of the intermediate samples are around 250 words, the corpus is reconstructed again with 200-300 words size samples. For level 5 and 6, the samples were randomly cut and adjusted to 200-300 words size. Cutting the size of samples, some of samples were cut in the front part, other some of them were cut in the middle, and the others were cut in latter part, so that the introduction, middle, and conclusion parts of the presentation were all included in the corpus. In addition, samples were cut at the part where the utterance was completed, without cutting it in the middle of the utterance. Since the number of samples for level 6 was the least by 11 samples, the number of samples for all the series was standardized to 11.

Table 2: *The Number of Words in each Sample*

Level 3		Level 4		Level 5		Level 6	
Name of Sample	Number of words	Name of Sample	Number of words	Name of Sample	Number of words	Name of Sample	Number of words

11530	219	8197	238	4977	295	8293	259
13664	295	8198	293	4988	252	11534	262
13851	280	8313	221	4989	299	11603	304
13856	225	8315	252	4991	276	11614	270
13858	271	8316	262	5006	285	13484	284
13860	283	8317	247	5114	287	13635	267
13862	285	8318	252	5115	265	13638	262
13874	286	11539	229	8183	215	13644	234
13876	280	13588	318	11532	255	13646	241
13888	221	13592	335	13640	261	13864	247
13895	226	13880	220	14051	256	13869	270
Average	**216.0**	**Average**	**260.6**	**Average**	**267.8**	**Average**	**263.6**

The selected samples were analyzed according to the following procedures. First, the samples were annotated using the Korean POS tagger program, *Utagger*, which is developed by the Korean Language Processing Laboratory of Ulsan University. For the learners' spoken production, there were a lot of errors. Thus, for accurate analysis, I manually checked the results, tagged them individually to make sure that they were done correctly. Then, the number of word types and tokens, the number of content words, and the number of low-frequency words for each sample were counted. Afterward, lexical density and sophistication value for each sample were calculated. In order to examine developmental patterns by learners' proficiency, the differences among level groups were statistically analyzed using the one-way ANOVA. I, also, analyzed the correlation between the learners' level and the lexical complexity.

2.2 Set Criteria for Lexical Density and Lexical Sophistication

Several criteria are needed to measure lexical density and sophistication. First, what kind of words are being analyzed in this study? Second, how will the learner's errors be handled? Third, what are the content words? Fourth, what are the low-frequency words?

In Korean, one word is composed of content words, such as nouns, verbs, and adjectives, and functional words like *eo-mi* (ending), and *jijeongsa* (copula). These functional words have no meaning in the sentences, but only have grammatical functions. Thus, they were excluded from the lexical analysis in this study.

Second, the learner's errors were reprocessed. If the learner's errors are analyzed as one individual form, the total number of word types will increase since the mistakes have been treated as a separate form. For this reason, the learner's error must be taken care of. For example, in the case of a mistake that pronounced 'Chingu (friend)' as 'Jingu', it was analyzed as the original form 'Chingu'. However, in the case of substitution errors rather than utterance errors, the words spoken by the learner were analyzed. Because of the nature of the spoken language, many hesitations or stuttering are found. When the learner stuttered and the word was not fully uttered, I did not analyze it as a portion of a whole word. For example, if the learner uttered only 'Wo', which is originally 'Wo-ri (we)', it was not recorded, because it was impossible to treat it as a word. In contrast, when a part of a content word was uttered and a part of a functional word was failed to be uttered (e.g. 'hakkyo (school)' instead of 'hakkyo-e (to school)'), I included it in the lexical analysis.

Third, there still remains a question to which degree are words content words in Korean? Content words are elements that represent the actual meaning such as object, action, event in a sentence or utterances. In English, there are nouns, verbs, adjectives, adverbs, quantifiers, interrogatives, and negatives. There is a difference in opinion among scholars about the border between content words and functional words in Korean. In this paper, I referred to the previous studies and set nouns, pronouns, numerals, determiners, verbs, adjectives, adverbs as a content word.

Finally, it is a discussion of what a low-frequency word is. It is a word that is not normally used in everyday life and it usually includes words such as academic words and terminologies. However, for Korean learners, the criterion of the low-frequency words may be different from the Korean L1 learners. Read (2000) also noted that low-frequency words can be viewed as a word, excluding high-frequency words, and is often seen as a word not included in 2,000 high-frequency words. Since Nam (2015) targeted Korean L2 learners, in this study, the low-frequency words were defined as words except 1,700 basic words. Won et al. (2017), also, used a vocabulary list of the National Institute Korean Language (2015), and analyzed the low-frequency words except for 1,800 beginner vocabulary words. Hence, in this paper, I followed the criteria of the preceding researches written above. The low-frequency words in this paper refers to words excluding the 1,836 beginner level vocabularies based on the list provided by the National Institute of Korean Language (2015).

2.3 Lexical Density and Lexical Sophistication Measurement Formula

The formula to measure lexical density and lexical sophistication based on the criteria presented in section 2.2 is as follows.

$$Lexical\ Density\ (LD) = \frac{the\ number\ of\ contents\ words}{the\ number\ of\ words} \quad (1)$$

$$Lexical\ Sophistication\ (LS) = \frac{the\ number\ of\ low-frequency\ words}{the\ number\ of\ words} \quad (2)$$

After calculating both lexical density and lexical sophistication in the learners' spoken production with the above formula, the results were arranged in an Excel file.

3. Findings

3.1 Lexical Density

First, the number of word types and tokens, and average number of content words are as shown in Table 3 below. Both types and tokens are considered in determining the lexical density.

Table 3: *Lexical Density Average according to Learners' Level*

	N	Number of total word tokens	Number of total word types	Number of content word tokens	Number of content word types	Lexical density A	Lexical density B	Lexical density C
Level 3	11	326.36	130.73	200.09	104.00	79.47	32.05	61.12
Level 4	11	310.36	114.36	186.18	89.00	77.70	28.23	59.75
Level 5	11	213.27	117.45	213.27	117.45	78.92	33.68	61.20
Level 6	11	326.64	135.91	201.64	107.00	78.02	32.41	61.55
Total	44	328.09	132.34	200.30	104.36	78.53	31.72	60.91

Lexical density A is the number of content word tokens divided by the number of total word tokens. Lexical density B is the number of content word types divided by the number of total word tokens. Lastly, lexical density C is the number of content word types divided by the number of total word types. Although there is no significant difference among the three methods, the lexical density of level 4 is lower than that of the other levels. Also, as the proficiency increased, the lexical density did not show a constant increase, but it decreased, then increased again, and then decreased back. ANOVA analysis was conducted to determine whether these differences were

statistically significant, but it was not statistically significant, as there was an insignificant difference in scores. As a result of the analysis, lexical density A and C were not statistically significant at p = .721, F =.444 and p = .888, F = .212, respectively. The lexical density B, also, showed a significance of .081, which was not a statistically significant difference. However, considering that the significance was decided with the standard of .05, the value of lexical density B was more significant as it was closer to .05 in comparison to the value of lexical density A and C. If the size of the corpus is increased, the assumption could be made that the lexical density would vary according to the learners' proficiency as it has been observed in lexical density B.

The analysis of the lexical density was slightly different from the results of the previous studies. In the case of Nam (2015) and Won et al. (2017), the lexical density was steadily increased as the learners' proficiency increases, and the difference was statistically significant. However, unlike the previous studies, there were no significant differences between lexical density as the proficiency grew.

3.2 Lexical Sophistication

Next, I measured how lexical sophistication changes as learners' proficiency increases. Like the lexical density measurement, lexical sophistication was analyzed by considering both the number of word types and tokens. Lexical sophistication is also divided into A, B, and C likewise the lexical density A, B and C. The lexical sophistication A is the number of low-frequency word tokens divided by the number of total word tokens, the lexical sophistication B is the number of low-frequency word types divided by the number of total word tokens, and the lexical sophistication C is the number of low-frequency word types divided by the number of total word types. Table 6 represents the average results of each levels' score.

Table 4: *Lexical Sophistication Average According to Learners' Level*

	N	Number of total word tokens	Number of total word types	Number of low-frequency word tokens	Number of low-frequency word types	Lexical sophistication A	Lexical sophistication B	Lexical sophistication C
Level 3	11	222.45	112.36	33.91	20.27	17.77	9.10	15.06
Level 4	11	201.91	96.18	43.91	24.45	25.60	12.33	22.00
Level 5	11	234.09	126.64	117.45	75.09	37.74	20.90	31.84
Level 6	11	218.91	113.73	56.45	35.82	30.38	15.99	25.32
Total	44	219.34	112.23	52.34	32.45	27.87	14.58	23.55

In lexical sophistication A, B, and C, it was found that the learners were using low-frequency words more often as they moved from the intermediate level to advanced level. However, the lexical sophistication scores of level 6 are slightly lower than those of level 5. This tendency can be seen in Figure 1 below.

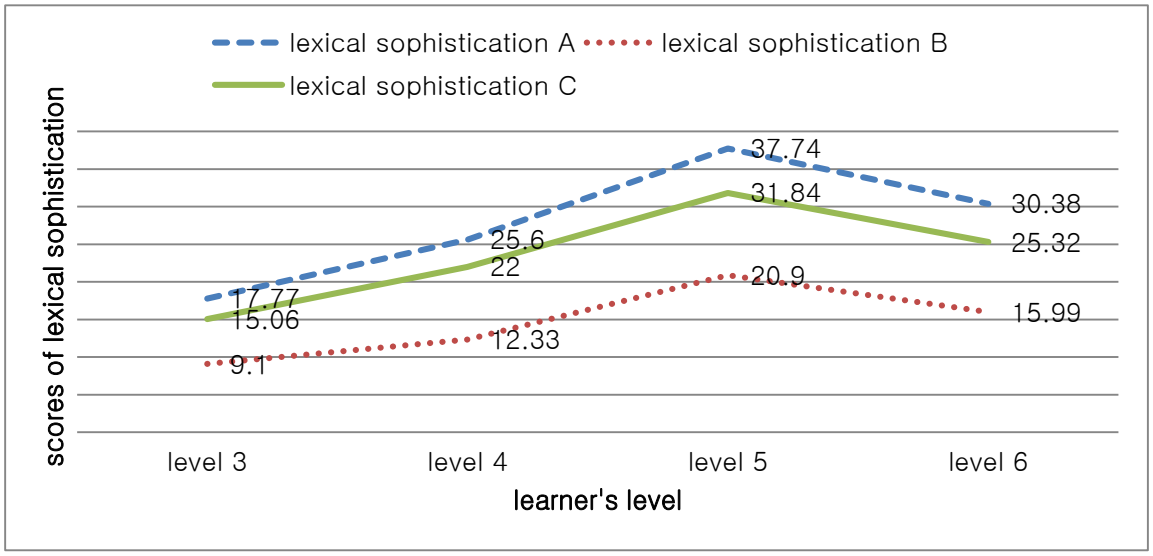

Figure 1: *Changes in Lexical Sophistication According to Learners' Level*

Later, one-way ANOVA analysis was conducted to determine whether there is a statistically significant difference in lexical sophistication according to learners' proficiency. As a result, the difference of sophistication was statistically significant in groups A, B, and C. The difference in group A was most statistically significant as p = .000 and F = 7.758. As a result of post-hoc analysis, the mean value of level 3 had a significant difference compared to those of level 4, 5 and 6. The result of B was p =.001 and F = 7.094, and the post-hoc result showed that the mean value of level 5 showed statistical difference from the results of level 3 and 4. Finally, the value of C was also statistically significant at p = .007 and F = 4.707. In the post-hoc analysis, there were statistical differences only between level 5 and 3. In order to find out whether the lexical sophistication A, B, and C are correlated with the learners' proficiency, I conducted an additional correlation analysis.

The analysis showed that lexical sophistication A was correlated with the learners' proficiency the most. In general, a correlation of .9 or higher is highly correlated, a correlation of .7 to .9 indicates a high correlation, and a correlation of .4 to .7 indicates a somewhat higher correlation. The p value of A, which is .467, showed a somewhat higher correlation. Furthermore,

I analyzed correlation between learners' proficiency level and lexical density B and lexical sophistication A. Lexical sophistication showed a slightly higher correlation, but lexical density was .144, showing little correlation with learners' proficiency level. This result was different from the previous studies. The correlation between the learners' writing level, lexical diversity, lexical density, and lexical sophistication in Won et al. (2017) showed that the lexical density was most correlated with the writing proficiency at p = .548. In this study, however, there was no correlation between learners' proficiency and lexical density, whereas lexical sophistication was correlated more. This difference may have shown due to the differences in registers. Won et al. (2017) analyzed the learner's written data, and this paper analyzed the learner's spoken data, so the difference in result might have occurred. In future studies, it is necessary to elucidate why contradictory results have come out.

4. Conclusion

In this paper, I analyzed the difference of the lexical complexity in learners' spoken production by using statistical methods. Lexical complexity is a way of showing how a learner's vocabulary knowledge is internalized. To analyze lexical complexity, lexical diversity, lexical density, and lexical sophistication are widely used. However, research on lexical density and sophistication has not been done much compared to diversity. Especially, it is very rare that studies have been done with learners' spoken language. Hence, I measured lexical density and sophistication of learners' spoken data. As a result, learners' lexical sophistication varied according to learners' proficiency, but density showed little difference. Unlike the previous studies, in which lexical density differs by learners' proficiency, it appeared that there is no difference between lexical density scores of intermediate learners and those of advanced learners. On the other hand, in the case of lexical sophistication, there was a difference between intermediate and advanced learners, and this difference was statistically significant. Also, lexical sophistication was analyzed to be somewhat highly correlated with learners' proficiency.

In this study, I analyzed about 11,000 words in 44 Korean learners' spoken data and found out that it is difficult to generalize the lexical complexity of learners' spoken production. In the intermediate level, most of the sample data only contained about 200-300 words, therefore the data of advanced learners had to be cut to a similar size. In the case of beginners' samples, there was almost no use of low-frequency words. Thus, the limitation of this study is that the beginners were

excluded from the analysis. For further research, a larger amount of data should be gathered, and lexical complexity of the learners' spoken language must be studied continuously.

References

Ahn, E. (2017). Analyzing lexical diversity and lexical density in Korean texts. *Language Facts and Perspectives*, *41*, 349-365. https://doi.org/10.20988/lfp.2017.41..349

An, G. H. (2003). Interlanguage lexicology: Issues and implications. *Bilingual Research*, *23*, 167-186.

Bae, D. Y. (2012). A Study on the lexical variation and lexical density shown in writing of KFL learners. *Journal of Language Sciences*, *19*(1), 99-117.

Kang, G. M., & Jin, J. R. (2017). A study on measurement method of lexical sophistication for Korean learners. *In proceeding 24th Korean Language and Culture Education Society Conference*, 141-149.

Kim, S. (2012). The characteristics of Korean utterance of Korean language learners: Focusing on MLU and TTR. *The Language and Culture*, *8*(1), 1-17. https://doi.org/10.18842/klaces.2017.13.1.1 https://doi.org/10.18842/klaces.2019.15.1.1

Lee, J. (2017). A study on lexical diversities in writing of university students – Focusing on the comparison of Koreans and foreign students. *The Korean Journal of Literacy Research*, *19*, 61-91.

National Institute of the Korean Language (2015). *Research on Korean Language Education Vocabulary Content Development (Stage 4)*. Seoul

Nam, J., & Kim, Y. (2014). Measuring lexical diversity in Korean learners' speech production focusing on D value. *Korean Semantics*, *45*, 69-97.

Nam, J. (2015). *A Study of oral complexity of Korean learner: Syntactic complexity and lexical complexity*. (Unpublished doctor thesis). Kyung Hee University, Seoul, Korea.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied linguistics*, *24*(4), 492-518. https://doi.org/10.1093/applin/24.4.492

Read, J. (2000). *Assessing Vocabulary*. Cambridge: Cambridge university press. https://doi.org/10.1017/CBO9780511732942

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity* (No. 17). University of Hawaii Press.

Won, M., Wang, Y., Zhu, Y., & Wang, H. (2017). A study of lexical richness of Korean learners' writing: the possibility of using lexical richness to measure language level. *Korean Languaeg and Literature, 71*, 33-55.