A Study of the Relationships between Lexical Richness and Writing Quality: Taking the English Majors at Guangxi University as an Example

Y.J. XIE
Foreign Languages College, Guangxi University, Guangxi, China

Y.SHEN
Foreign Languages College, Guangxi University, Guangxi, China; Institute of Intelligent Systems, The University of Memphis, Tennessee, USA

ABSTRACT: The present paper aims to investigate the correlation of lexical richness with writing quality by measuring lexical richness in EFL English writing. The research was performed by measuring timed compositions of 56 senior English majors with four indices of lexical richness, including text length, high frequency words, lexical density, and lexical sophistication. The finding of this study is that lexical richness is indeed taken into consideration in scoring the quality of English writing. The measure of the four indices of lexical richness can distinguish between the higher-scoring compositions and lower-scoring compositions, which shows, to a statistically significant extent, that lexical richness correlates with the quality of English writing. It is hoped that the findings of the present study will be useful for improving the teaching and learning of English vocabulary.
KEYWORD: lexical richness; writing Quality; English majors

1 GENERAL INTRODUCTION

Learning to speak and understand English is very fashionable in China. Indeed, as China has increased exchanges with the outside world in recent years, students in China find it increasingly necessary and important to learn English, the world language. However, one problem students face in learning English is the difficulty of expressing themselves in English writing. As a way to convey ideas, thoughts, or suggestions to the reader, writing is one of the most important means of communication, and is one of the basic skills a language learner must grasp. However, it also proves to be a very challenging skill to acquire. Many researchers have explored the reasons for inferior written work on the part of EFL learners, and they have found that a critical factor is a lack of vocabulary, which hinders EFL learners in producing high quality written pieces, in spite of the clarity of their thoughts and ideas.

This phenomenon is concerning, due to the fact that it is largely ignored as an area of research. In its place, applied linguistics research has largely emphasized the grammatical, phonological and orthographic aspects of English. Some research has investigated EFL learners' vocabulary, most of which focuses on the relationship between vocabulary and reading rather than lexical competence and writing quality. This present study focuses on lexical richness, an important indicator of language learners' active vocabulary, and will investigate the correlation between lexical richness and the quality of English writing.

Due to the critical role that lexical richness plays in learners' language production, it is worth substantial effort to shed light on how lexical richness correlates with the quality of English writing. Based on the corpus of 56 English majors' timed compositions, the present thesis adopts a quantitative analytic approach. The results show that lexical richness significantly correlates with the quality of English writing. Another, more concerning issue is also made clear in that college students' vocabulary size is found, generally, to be far from adequate. The present study has been performed in the hope of providing some insight into the teaching of English writing and the acquisition of English vocabulary for EFL learning. The study aims especially at especially drawing college students' attention to how critical vocabulary building is.

2 SOME KEY TERMS ABOUT LEXICAL RICHNESS IN ENGLISH WRITING

Based on the implications of previous studies, this thesis aims to further investigate the correlation between lexical richness and the quality of English writing. In order to ensure clarity within the thesis, it is necessary to define various terms used.

2.1 *The Notion of Lexical Richness*

2.1.1 *Word Tokens*

Word tokens refer to the total number of word forms, which means that every word is counted in a spoken or written form if an individual word occurs one time or more. Take "desk" as an example. If both singular and plural form of this word (i.e. *desk* and *desks*) appear in the text, they are counted as two items. Word token can indicate text length.

2.1.2 *Word Types*

The number of types in written production is the total number of the different word forms, which means that the repeated word forms are not counted. To be more specific, in calculating the number of word types, inflected forms are counted once. That is, *desk* and *desks* are deemed only one item. Word types corresponds approximately to dictionary entries, and reflect the difficulty of a text.

2.1.3 *Lexemes*

In vocabulary studies, the base and inflected form of a word are collectively known as a lemma. When researchers conduct a study involving counting the number of word types in a spoken or written texts, one step is to lemmatize tokens, so that the inflected forms are counted as instances of the same lemma as the base form (Read, 2000:18).

2.1.4 *Word Families*

Another important notion is that of a "word family." This is composed of a headword, all of its inflected forms and its closely related derived forms. Therefore, the word *work*, its inflected forms *works*, *working* and *worked*, as well as its derivative *worker*, all belong to a word family. According to Laufer (1991, 1992 & 2003), a lexical threshold of 3,000 word families (or 5,000 lexical items), which provides a coverage of between 90% and 95% of any text, is necessary for non-native speakers to master minimal reading comprehension.

2.2 *The Measurement of Lexical Richness*

2.2.1 *Lexical Density*

Lexical density is defined as the percentage of lexical or content words within a text. Lexical words include nouns, verbs, adjectives and adverbs, which primarily convey information. A text is considered "dense" if it contains many lexical words relative to the total number of words. In other words, the higher the number of content words that are used in the text, the larger the amount of information it can convey. Previous studies on lexical density show two methods used to measure the percentage of lexical words in the texts. One measurement, put forward by Laufer, takes lexical density as the ratio of the number of lexical words to the total number of words in a text. Another method, shared by Engber and Read, defines the accounting of lexical density as the ratio of total number of lexical words to the total number of words in the essay. In the present research, the measurement given by Laufer is deployed.

2.2.2 *Lexical Sophistication*

Lexical sophistication is defined as the percentage of "advanced words" in a text. It is calculated by dividing the number of sophisticated words by the total number of words in the text. Some consider sophisticated words to be the words that the subject is not required to know, while others define them as low frequency words. Therefore, when labeling "advanced" words, it is necessary to take the learner's level of training into consideration. There is one reliable and objective measure for this dimension, known as the Lexical Frequency Profile (LFP). Proposed by Laufer and Nation in 1995, the LFP is a measure of lexical diversity in writing that categorizes words in a learner's text according to the frequency band to which each word belongs. There are four frequency levels, which include the 1,000 highest frequency words, the second most common 1,000 words, words judged as being academic words, and off-list words, which denotes all words that do not appear in the other three categories. The LFP focuses solely on the lexis aspect and is therefore suitable for assessing lexical richness. Furthermore, the definition of advanced words is based on word frequency and is able to discriminate between subjects who use higher or lower frequency vocabulary. In the present study, two indices of lexical richness are based on the LFP. Vocabulary words from the 56 compositions that fall in the 2,000 most frequent words are considered high frequency words, while academic and off-list words are adopted to signify lexical sophistication.

2.2.3 *Lexical Variation*

This is widely defined as the type-token ratio. Specifically, this is the ratio between the different word forms in the text and the total number of word forms. It aims to demonstrate how inclined the learner is to repeat the same words in written production. However, the type-token ratio is criticized for its sensitivity to text length. It is found that the greater the number of tokens found in a composition, the lower the type-token ratio will be. This therefore renders the test an unfair method of assessing lexical variation of compositions of varying length. The compositions used for the present study were required to be written in about 400 words, yet inevitably they vary in text length. Therefore, instead of using the lexical variation as

one indice to measure lexical richness, text length is deployed, which makes a difference in the quality of English writing.

The above measurement mentioned provides a great insight, for the present investigation, of the correlation between lexical richness and writing quality and how the four indices are used to test the correlation will be discussed in detail in the research part of this thesis.

3 RESEARCH PROCEDURE

3.1 *Research Questions*

The present study is an empirical investigation, attempting to test the correlation between lexical richness and writing quality. Four variables are included, and they are the four indices of lexical richness, specifically, lexical density, lexical sophistication, high frequency words and text length. The study aims to answer the following questions:

(1) Is there any significant correlation between lexical richness and the quality of EFL writing in terms of the four indices? If so, how is this achieved?

(2) Which indice in lexical diversity has the most significant impact on the quality of English writing?

3.2 *Subjects and Research Time*

The participants in the study are senior English majors who have taken part in the TEM8 mock examination. The students were asked to write their thoughts about the value of college and university education in China, and whether this has decreased with reports of unskilled migrant workers often earning more than recent graduates. The students were limited to 40 minutes of writing time, with no dictionary or other references allowed. This was to ensure that students put their mind on writing to exert their lexical competence with their vocabulary base.

All compositions written by the students were graded according to the TEM8 scoring standard. Fifty-six compositions were chosen randomly as samples for the present study. The following is a sample of a student's writing.

3.3 *Data Collection and Research Methods*

All 56 compositions were input into a computer and carefully checked to ensure the input was identical to each student's composition. Spelling errors were corrected if the intended words were discernible. Finally, each composition was saved and labeled with its corresponding grade. Based on the scores, the samples were divided equally into two groups, a high-scoring group and a low-scoring group.

All the compositions were coded, tagged and counted for selected measures of lexical richness by utilizing computing instruments and utilities such as Vocab-profiler, Excel, and SPSS 19.0.

In order to collect all the data of the writers, an Excel spreadsheet based on the results of each writer was created. Excel was also used to establish descriptive statistics. SPSS 19.0 was applied to see if there was any significance of these variables, and the correlation between lexical richness and the quality of English writing was further verified with the help of SPSS 19.0.

4 RESEARCH RESULTS AND DISCUSSIONS

4.1 *Text length*

Text length is determined by the number of word tokens, and therefore the total number of words forms in each composition. Table 1 shows that there is substantial difference between the high-scoring group and the low-scoring group. The former used a total number of 9980, while the latter group used only 9013 words tokens. In addition, the significance of the word tokens is 0.013 with a mark. This shows that text length has a positive significance in the quality of English writing.

4.2 *High Frequency Words*

In this research, the term "high frequency words" refers to the first and second thousand highest frequency words used in senior students' compositions. In order to demonstrate the high frequency words more clearly, both components' statistics are presented. Unlike the text length, which shows a clear difference between the two groups, the averages within the first and second thousand words in the two groups are very similar, regardless of the different level of writing quality within the samples and their grade. From the Pearson Correlation, we see that high frequency words have a negative significance for writing quality. Therefore, the more high frequency words used in the composition, the lower of the score will be.

4.3 *Lexical Density*

The two groups make up a relatively low percentage compared with the percentage of high frequency words in the compositions, with 50% in the high-scoring group and 53% in the low-scoring group. According to the Pearson Correlation, lexical density has a negative significance with writing quality.

4.4 *Lexical Sophistication*

According to the results of descriptive statistics, it is apparent that the academic words and off-list words

occupy a very low percentage of use in both the groups of students. However, a more substantial difference exists between the high-score group and low-score group, with results of 9.20% and 8.13% in terms of lexical sophistication, respectively. What is more, lexical sophistication shows a positive and obvious significant correlation with the quality of English writing with two marks. All the data reveal that lexical sophistication plays a vital role in the scores of compositions.

5 CONCLUSIONS

This study investigated 56 compositions by English majors in their senior year to find whether a correlation exists between lexical richness and writing quality. The measurement of lexical richness in the compositions was realized by analyzing four lexical indices including text length, high frequency words, lexical density and lexical sophistication. As the data have shown, a correlation does indeed exist between lexical richness and the quality of EFL writing. Text length and lexical sophistication show a positive and significant relationship with writing quality. On the contrary, high frequency words correlate negatively with writing scores, but this category of words makes up a large part of the compositions of both groups. No significant relationship is found between lexical density and writing quality.

It is hoped that the findings of this research will shed light on vocabulary learning and teaching. As the results reveal, senior students' written vocabulary size is far below the expected amount. Adequate vocabulary size is essential to express ideas and enhance writing quality, so one major pedagogical implication of this study is that vocabulary size needs to be enlarged and enriched. Specifically, while practicing writing, students should be encouraged to use sophisticated words to impress raters. They should also vary alternate words to convey the same idea, which will significantly level up their writing quality. At the same time, teachers should pay attention to students' vocabulary knowledge. In addition to exposing learners to a massive amount of vocabulary, teachers should also help students to discover and acquire collocations and lexical bundles, and use them in different topics and different registers.

Though it was carefully designed, the study may have some limitations. Writing is a very complicated process. There are many factors that could affect the quality of writing, like organization, punctuation, rhetorical devices, syntactic complexity, etc. Lexical richness is only one of these factors. Furthermore, holistic rating involves many subjective factors. If a rater is incompetent, or not familiar with learners' writings, the validity and reliability of their judgment may be open to question. The present samples were scored by several graduate students at Guangxi University, which cannot completely ensure the final scores of students' compositions reflect their language proficiency. As the analysis of the compositions has shown, some writings with a high level of lexical sophistication obtained a low writing score. This may somewhat distort the research findings. Thus this study is still preliminary and subject to revision in light of the results of future studies.

ACKNOWLEDGEMENTS

This is research is financially supported by Innovative Research Foundation for National University Students in 2014 (project number is 141059349).

REFERENCES

[1] Laufer, B.1991. The development of lexis in the production of advanced L2 learners. *The Modern Language Journal* 75: 440-448.

[2] Laufer, B.1992. Reading in a foreign language: How does L2 lexical knowledge interact with the learner's general academic ability? *Journal of Research in Reading* 15: 95-103.

[3] Laufer, B. & P. Nation. 1995. Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*16 (3):307-322.

[4] Laufer, B. 2003. The influence of L2 on L1 collocational knowledge and on L1 lexical diversity in free written expression. In Cook V(ed.). *Effects of the Second Language on the First*. Clevedon: Multilingual Matters LTD. 19-31.

[5] Read, J. 2000. *Assessing Vocabulary*. Cambridge: Cambridge University Press.