Insights from Data-driven Learning on Vocabulary Use in the TEM-8 and Students’ Compositions

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

M.Z. XIAO
Foreign Languages College, Guangxi University, Guangxi, China

ABSTRACT: This paper analyzes the features of vocabulary use of TEM-8 compositions (Test of English for English Majors Band 8) and third-year English majors’ compositions at Guangxi University based on the calculations of RANGE, SPSS 19, and the Jukupigai automated scoring system. This research aims to identify weaknesses, problems, and confusions in the writing of English majors and to provide strategies for correcting these. The results indicate that TEM-8 compositions display a variety of vocabulary use; most topics in the TEM-8 are debatable issues in social and cultural areas; most third-year English majors use fixed adverbs, conjunctions, and adjectives in their writing; the accuracy and appropriateness of personal pronouns and modal auxiliary verbs is problematic in students’ writing; and the numbers of conjunctions and verb phrases in students’ compositions have a moderate influence on their English-language writing scores.

KEYWORDS: data-driven learning; TEM-8 compositions; Jukupigai automated scoring system

1 GENERAL INTRODUCTION

As a high-level test for English majors that aims to examine candidates’ integrated English-language skills, the TEM-8 is one of the most essential exams for students majoring in English and for teachers of English. Therefore, most senior English-major students and their teachers are highly motivated to learn strategies for improving English-language skills so that students can pass this exam successfully. The TEM-8 evaluates learners’ English-language skills in the following four areas: listening, reading, writing, and translating. Of these areas, writing, which accounts for 20 percent of the total score of the test, is undoubtedly a critical part of this exam. It is well known that writing skills are closely associated with the ability to use vocabulary accurately and effectively. The richness of one’s vocabulary has a strong impact on one’s writing. Using a variety of carefully chosen words can significantly improve a composition. Therefore, an analysis of vocabulary use for the writing portion of the TEM-8 can not only improve a learner’s writing ability but also provide useful suggestions for English-language teachers.

Scholars have conducted research on vocabulary for many years. Gui (1985) conducted empirical research on English majors’ vocabulary in China. Liu (2003) analyzed the influence of productive vocabulary use on writing quality and developed research focusing on students’ productive vocabulary. These studies have provided some reference for us, and data-driven learning is one of the benefits resulting from the efforts of these scholars. Data-driven learning, proposed by Tim Johns in 1991, is a form of classroom presentation or activity that has been associated with the lexical syllabus approach and communicative methodology. This approach regards students as researchers and teachers as guides, and it has aroused scholars’ interest since it came into being. For example, Wang (2010) researched the mode and method of using corpora to conduct data-driven learning. Suo (2012) discussed the autonomic learning mode of college English from the perspective of data-driven learning.

In this paper, the following questions will be addressed:

Can we find any apparent features in vocabulary uses based on the analysis of TEM-8 compositions?

Are there any evident characteristics in the use of vocabulary concerning third-year English major compositions?

What is the obvious difference in the lexical uses with respect to TEM-8 and third-year English major compositions?

What are the implications for taking the TEM-8 composition exams in the future?

2 RESEARCH OBJECTIVES AND METHODS OF THE STUDY

2.1 *Objectives*

This article is based on a study of third-year English majors’ compositions at Guangxi University and TEM-8 compositions from 2000 to 2013 using the statistical tools RANGE, SPSS 19, and Jukupigai automated scoring system. The work focuses mainly on vocabulary use in the TEM-8 compositions and third-year English majors’ compositions from the perspective of data-driven learning. By the comparisons between TEM-8 and third-year English majors’ compositions, this research aims to identify weaknesses, problems, and confusions in English majors’ writing and provide strategies to help them pass the TEM-8 writing test. In addition, this paper aims to offer advice to teachers of English for their writing instruction.

2.2 *Methodology*

This research used three methods to analyze the compositions. First, we used the data analysis software RANGE to analyze the collected 15 TEM-8 compositions used for the TEM-8 examinations in 2000 through 2013.

RANGE was designed based on word frequency by New Zealand linguists Paul Nation and Averil Coxhead, professors at Victoria University in 2002. At present, RANGE is widely used in linguistic study; for instance, Bao (2005) evaluated productive words using RANGE. Cheng (2009) explored the stylistic use of RANGE corpus analysis software by taking an inaugural address as an example. RANGE has three authoritative word lists: Baseword One, Baseword Two, and Baseword Three. When we use RANGE to process a statistical analysis of the target text, the software refers to these three word lists and the output is the word frequency results. Baseword One includes approximately 1000 of the most frequently used word families, and Baseword Two comprises approximately 1000 word families that are less frequently used. Baseword Three has 570 of the most frequently used academic words appearing in senior high school and college textbooks.

Second, we chose six writing topics from the TEM-8 as writing tasks and asked the 40 third-year English majors to submit their essays on these six topics via the Jukupigai automated scoring system, where it is possible to upload compositions repeatedly. The Jukupigai automated scoring system can analyze vocabulary, such as the distribution of lexical categories and statistics about word frequency.

Third, we used data analysis software SPSS 19 to analyze the uploaded compositions to the Jukupigai automated scoring system by 40 third-year English majors at Guangxi University; Pearson correlation of the numbers of conjunctions, verb phrases, and scores was tested.

3 ANALYSIS AND DISCUSSION OF THE TEM-8 COMPOSITIONS BY RANGE

3.1 *The statistical result of tokens and types in TEM-8 compositions*

There are 5439 tokens and 1385 types in TEM-8 compositions. Of these 1385 words, 818 words belong to Baseword One and account for 59.06 percent; 268 words are contained in Baseword Two and account for 19.35 percent; and 99 words are contained in Baseword Three and account for 7.15 percent. In addition, there are 200 words not included in on the lists, and these words account for 14.44 percent. In addition, it is easy to find that the lexis seen in the TEM-8 writing section comprise the most frequently used words, but several less frequently used words and academic words also appear in modal compositions. Therefore, in order to enrich vocabulary in compositions, students should include less frequently used words and academic words. In addition, English majors should use different words to express the same meaning; a good composition should avoid using particular words or expressions repeatedly.

3.2 *Word frequency of nouns in TEM-8 compositions*

From the data we can see nouns include the following words: *people*, *students*, *education*, *life*, *university*, *society*, *dialects*, *ambition*, *Internet*, and *friendship*, and that these appear in modal compositions of the TEM-8 numerous times. It can also be seen that the words such like *student*, *society*, *university*, and *life* are the most common nouns used in TEM-8 writing. This indicates that most writing topics in the TEM-8 are based on issues within the social and cultural fields. Therefore, according to the results of data analysis, English majors taking the TEM-8 examination should focus on relevant, current social issues and use vocabulary related to these. Moreover, practice plays an important role.

3.3 *Word frequency of conjunctions and adverbs in TEM-8 compositions*

From the data we found that eight conjunctions and eight adverbs are included in the top 100 words used in TEM-8 compositions. From this we consider that conjunctions and adverbs are of great importance in the TEM-8 writing. The eight conjunctions are *and, that, as, if, or, but, however, because* and the eight adverbs are *not, so, when, also, there, even, however,* and *only.*

4 ANALYSIS AND DISCUSSION OF STUDENTS' COMPOSITIONS BY THE JUKUPIGAI AUTOMATED SCORING SYSTEM

4.1 *Lexical category distribution in students' writings*

According to the statistical results of the Jukupigai automated scoring system, the lexical category distribution of Topics 1–6 is similar; nouns, adjectives, and verbs make up the largest part of the composition.

4.2 *Error distribution in students' writings*

The finding indicates that there are two obvious problems in the writings of third-year English majors at Guangxi University. One is the use of "Chinglish" expressions, and the other is a lack of knowledge about basic English grammar. These two problems interrupt the flow of students' writing and prevent them from writing smoothly.

4.3 *Adverb, conjunction, and adjective frequency in students' writings*

(1) The words *not, so, also, much, there, when,* and *however* are the most frequently used adverbs in students' writings.

(2) *And, that, but, or, if, while* and *because* always appear in students' compositions, especially the coordinating conjunction *and.* As the statistical results show, the word *and* appeared 2112 times in all of the compositions written for Topics 1–6; that is an average of nine times in each composition.

(3) The words *good, many, some,* and *this* proved to be students' first choices to use in their writing.

(4) The data demonstrate that most students choose the same adverbs, conjunctions, and adjectives to use in their writing, which has the effect of making their compositions sound familiar. Using less common adverbs, conjunctions, and adjectives in TEM-8 writing instead of commonly used words is the key to making one's writing have a positive impression on exam valuations.

5 DATA COMPARISON BETWEEN TEM-8 AND STUDENTS' COMPOSITIONS

5.1 *Comparison of personal pronoun use*

(1) In TEM-8 compositions, third-person pronouns are the most widely used personal pronouns, followed by first-person pronouns. Second-person pronouns do not appear in the top five existing pronouns.

(2) We know that, first, students show completely different selections in personal pronoun use. In details, students choose first-person pronouns primarily, followed by third-person pronouns. Next, second-person pronouns are in the top five. We believe that students' choices of personal pronouns are connected to their typical manner of Chinese expression. In Chinese writings, people prefer to use first-person pronouns because it is easier to shorten the distance between the author and the readers. Influenced by this writing habit, most students automatically choose first-person pronouns in their compositions.

(3) Students' compositions differ greatly from the model compositions in regard to the use of personal pronouns. The accuracy and appropriateness of personal pronouns is still problematic in students' writings. Readers interact with the writer through the discourse (Wei, 45–51). We consider that the third-person pronouns are more suitable for the TEM-8 writing because the writing in the TEM-8 focuses mainly on argumentation, and using third-person pronouns to state the thesis makes an argumentative composition objective and persuasive. In contrast, a first-person pronoun makes an argumentative composition subjective and a second-person pronoun makes the tone of an argumentative composition too aggressive.

5.2 *Comparison of modal verb use*

Modal verbs are extremely important in any kind of discourse. As Michal and Ronald proposed (2005), modal forms, from a broad perspective, refer to the attitude of the speaker or writer towards information or ideas expressed. Fowler (1979) also points out that modality can reflect writers' or speakers' opinions and attitudes. Students' compositions and model compositions have different preferences in regard to modal verb use. Model compositions use more euphemistic modal verbs, such as *may,* while several students chose modal verbs with strong modal meanings, such as *must.* According to Halliday (2000), *must* is a high-value modal auxiliary with a very strong modal meaning. It should be used sparingly in TEM-8 compositions, especially in argumentation, as this word may make readers uncomfortable. We can infer from the data that the accuracy and appropriateness of modal verbs is still problematic in students' writings.

6 ANALYSIS AND DISCUSSION OF THE STUDENTS' COMPOSITIONS BY SPSS 19

The number of conjunctions in Topic 1 does have a correlation to the scores (r=0.430, p>0.05). The number of conjunctions in Topic 2 has a significant correlation to the scores (r=0.573, p>0.01). But for Topics 3, 4, 5, and 6, the number of conjunctions used does correlate to the scores, but the correlations are not significant. These results demonstrate that the number of conjunctions used in different topic compositions has a certain but not necessarily a decisive influence on the scores they receive.

For Topic 1, the number of verb phrases has a correlation with the scores, but the correlation is not significant (r=0.289 (*p*>0.05). For Topic 2, the Pearson correlation coefficient is 0.626 (*p*<0.01) bilateral, which means that the number of verb phrases has a significant correlation to the scores. In the same way, for Topic 3, the number of verb phrases does have a correlation to the scores, but the correlation is not significant. For Topic 4, the number of verb phrases has a significant correlation with the scores. For Topics 5 and 6, the numbers of verb phrases have a correlation to the scores, but the correlation is not significant.

7 CONCLUSIONS

Through the analysis on the basis of the calculations of RANGE, SPSS 19, and the Jukupigai automated scoring system we know that: (1) the vocabulary in model compositions consists mainly of the most frequently used words, while less frequently used words and academic words also have a certain proportion; and (2) most topics in TEM-8 writing are about social and cultural fields and are closely associated with campus life. According to the statistics of the Jukupigai automated scoring system, we found that nouns, verbs, and adjectives take the largest proportion of the vocabulary in students' compositions; most students choose similar expressions that lead to a lack of richness in their vocabulary use. The comparisons show that the use of modal verbs and personal pronouns are still significant problems in students' compositions. We can infer from the results of SPSS 19 that the numbers of conjunctions and verb phrases used in students' compositions have certain correlations to their scores. Basing on the above results we suggest that teachers should focus on the use of conjunctions and verb phrases during the instructional process and provide instruction that focuses on addressing and eliminating the writing problems described above. Furthermore, when students write, they should pay close attention to their overall use of vocabulary when preparing for their TEM-8 examinations.

ACKNOWLEDGEMENTS

This research is supported by the core research project (XBS100008) of Guangxi University in 2010 and Innovative Research Project for National University Students in 2014 (project number is 141059349).

REFERENCES

[1] Bao Gui & Wang Xia. 2005. An application of *Range* in the evaluation of second language productive vocabulary use. *Media in Foreign Language Instruction*, 54-58.

[2] Cheng Shi. 2009. An application of corpus analysis software *Range* in the research of literary forms. *Chinese Journal of Learning and Teaching of Foreign Languages*, 9, 42-46.

[3] Fowler, R. 1979. *Language and Control*. London: Routledge & Kegan Paul.

[4] Gui Shicun. 1985. The investigation and analysis of the English vocabulary usage for the English-major students in China. *Modern Foreign Language*,1-6.

[5] Halliday, M.A.K. 2000. *An Introduction to Functional Grammar*. Beijing: Foreign Language Teaching and Research Press.

[6] Johns, T. 1991. Should you be persuaded: two samples of data-driven learning materials. *English Language Research Journal*, 4, 1-16.

[7] Liu Donghong. 2003. The functions of vocabulary in English-language writing. *Modern Foreign Language*, 2, 180-187.

[8] Michal McCathy & Ronald Carter. 2005. *Language as Discourse Perspective for Language Teaching*. Beijing: Beijing Press.

[9] Nation, P. & Coxhead, A. 2002. RANGE. http: / / www. vuw. ac. nz/ lals/ staff/ Paul_Nation.

[10] Nation, P. 1990. *Teaching and Learning Vocabulary*. New York: Newbury House Publishers.

[11] Suo Xinjia. 2012. Independent study of College English from the perspective of data-driven learning. *Journal of Congqing Technical College*, 4, 195-198.

[12] Wang Junsong. 2010. A model of data-driven learning and fostering of the ability to take independent study of vocabulary--based on a teaching experiment from COCA corpus. *China's Foreign Language Teaching*, 1, 24-30.

[13] Wei Jingzhu. 1996. The coherence of text and reading comprehension. *Foreign Language Teaching and Learning*, 1, 45-51.

[14] Zhang Renxia. 2010. The application of *Range* in the comment of college English writing. *Journal of Yinchun College*, 9, 184-187.

[15] Zhou Shen. 2011. *A New Guide for TEM-8 Exam*. Shanghai: Shanghai Foreign Language Education Press.