Automated Evaluation of Text and Discourse with RANGE

Ying Shen[1], Kai Li[2], Ting Guo[2]

[1] Foreign Languages College Guangxi University, Guangxi Province 530004, China; Institute of Intelligent Systems The University of Memphis, Tennessee State 38111, USA

[2] Foreign Languages College Guangxi University, Guangxi Province, China

shenying388@gmail.com, 309609422@qq.com

Abstract - This paper analyzes the features of vocabulary use in relation to reading comprehension in Test for English Majors-Band 4 and Band 8 (TEM-4 and TEM-8 hereafter) from 2000 to 2014 by using corpus analysis software *RANGE* and data analysis software SPSS 19. The vocabulary use features are as follows: (1) pronouns, conjunctions and propositions are used more frequently than other word categories, such as modal auxiliaries; (2) there is a significant difference between conjunctions, modal auxiliaries and notional verbs between TEM-4 and TEM-8; (3) the topic scope of reading comprehension in TEM-8 is much wider than in TEM-4.

Index Terms - Software *RANGE*; SPSS 19; TEM-8; TEM-4; reading comprehension; vocabulary

1. Introduction

Test for English Majors---Band 4 and Band 8 (TEM-4 and TEM-8) have been used for over 20 years. The number of students attending the tests is growing quickly and is forty times higher than before (Zhou, 2010). TEM-4 and TEM-8 have become the most important evaluation tools of university-level teaching quality and English proficiency of English majors in China. Reading comprehension has an important role in TEM-4 and TEM-8, responsible for 20% of the total score. The requirements of *Teaching Syllabus of English Majors of High Education Universities* (2000) specifies that (1) students can read essays and comments in English newspapers as well as historical biography and literature works with difficulty published by English-speaking country; (2) students can catch the main ideas, text structure, language features and rhetoric uses in reading materials; (3) students reading rate should be 140-180 words per minute. With globalization and the development of China's English teaching population, there will be more and more requirements for English majors. Vocabulary is one of the basic constituents of text. An analysis of vocabulary in reading comprehension in TEM-4 and TEM-8 may provide some implication for reading and vocabulary teaching.

RANGE corpus software can be used to analyze a text or discourse. The working principle is to compare object text with recognized authoritative vocabularies with results shown in tables. The results show the words of the text in recognized authoritative vocabularies, word frequency and its percentage of the whole text. By analyzing this data, we can understand lexicon use and other features of the text.

Many scholars have made use of *RANGE* corpus analysis software for research. Bao & Wang (2005) use *RANGE* corpus analysis software to evaluate productive vocabulary of second foreign language. Liu (2003) discusses the functions of vocabulary in English writing. Cheng (2009) explores the stylistic use of *RANGE* corpus analysis software by taking example of Inaugural Addresses. This research laid a foundation for related studies.

This paper first analyzes lexicon use features of reading comprehension text in TEM-4 and TEM-8 with the aid of *RANGE* corpus analysis software. Next, SPSS data analysis software will be used to draw conclusions for English majors in their preparation of TEM-4 and TEM-8 reading comprehension test.

2. Research Objectives and Methods

The research data is composed of the vocabulary usage in each original text of TEM-4 and TEM-8 from 2000 to 2014. Data and materials are collected from Internet, library and related books. SPSS 19 data analysis software is used to make a comparison of each item, including pronouns, conjunctions, prepositions, modal forms, notional verbs and nouns, according to the results from *RANGE*. A paired-sample *t*-test and a Pearson correlation of each pair are conducted.

The reading section of the TEM-4 and TEM-8 tests is an important constituent, taking 20% of the whole score. These texts cover a variety of topics, including economy, politics, technology, culture and arts. Having a broad view of lexicon use features in texts like this may help students prepare for these two tests and further guidance in English teaching.

Corpus analysis software *RANGE* based on word frequency is used to compare the vocabulary and diction of different texts. The results are reported using a number of key terms: *tokens* refers to the frequency and rate of some words; *types* defines the word classes and the rate of each in text; *families* refer to word families, including infection forms, derivative forms and so on. The vocabulary analysis is based on three basic vocabularies, called Basewrd1.txt, Basewrd2.txt, and Basewrd3.txt, as reported by Nation (1990:19).

The Basewrdl.txt vocabulary includes 998 of the most commonly used word families, for a total of 4119 words. The Basewrd2.txt vocabulary includes 988 commonly used word families, for a total number of 3,708 words. According to

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).

Nation (1990:19), these two basic vocabularies cover 87% of English words in discourse. In addition, the Basewrd 3.txt vocabulary includes academic words not found in Basewrd 1.txt and Basewrd 2.txt. This vocabulary represents words frequently used in college and high school textbooks, including 570 words families, for a total of 3107 words, comprising 8.50% percent of all words in academic books. Range corpus analysis software will compare the object texts with these three basic vocabularies and output the total number and percentage of words found in these basic vocabularies. The frequency and percentages are also ranked from high to low.

3. Results of the Reading Texts in TEM-4

A. Types of Vocabularies in the Reading Texts in TEM-4

There are 4,716 *types*, among which 1887 coming from Basewrd1.txt, comprising 40.01%; 764 *types* from Basewrd2.txt, totaling 16.20%, and 532 *types* from Basewrd3.txt, making up 11.28%. The remaining 32.51% represent 1533 *types* not included in these three basic vocabularies. The rate of *tokens* from these three basic vocabularies is 80.87%, 6.06%, and 3.95%, respectively, with 9.12% of *tokens* not included in any of these lists. This data demonstrates that most vocabulary, including staple words, hypo-ordinary words and academic words, are also widely used.

B. Conjunctions in the Reading Texts in TEM-4

The top ten conjunctions are AND, AS, BUT, OR, IF, SO, WELL, FIRST, NOW and WHILE. The cohesion and coherence theory put forward by Halliday and Hasan is by far the most comprehensive treatment of discourse analysis and has become the standard text in this area. According to Halliday and Hasan, there are five methods of cohesion. They are cohesive ties, reference, substitution, ellipsis, conjunction and lexical cohesion. The current discussion of conjunctions is based on this theory. Conjunctions in discourse not only function as connections and transitions, but also bring in new information and new topics. A study of conjunctions can help readers understand the logical structure of the discourse and provides a better idea of the discourse.

C. Pronouns in the Reading Texts in TEM-4

The pronoun is an important form of ellipsis, which is also a method of cohesion. Pronouns can also be a way of substitution to avoid repetitions, and to connect the whole text. In discourse grammar, personal pronouns and demonstrative pronouns can be considered part of the same system. For example, the pronoun IT is not only associated with HE or SHE, but can also be a substitute for THIS or THAT (Song Hong, 2010).

As this paper tries to study the features of vocabulary use on discourse level, the pronouns analyzed here include personal pronouns, demonstrative pronouns, and other related pronouns. The top pronouns in the relevant texts are THAT, IT, I, YOU, HE, THEY, and WE. The accumulative percentage is 6.9% of all words, with personal pronouns ranked at the top. Halliday & Hasan (1985) demonstrate that under the framework of discourse analysis, anaphoric pronouns are not be limited to content words, but also play a significant role in cohesion. Studying pronouns from discourse level can help the interpretation of the text, but may also present some ambiguity. For example, English students should read the title and context, then try to determine a clear idea of the meaning of the pronouns, as well as the discourse. Usages of pronouns can vary, so students must consider the use of pronouns in the discourse and the consistency of those pronouns.

D. Prepositions in the Reading Texts in TEM-4

The top ten prepositions in the TEM-4 reading texts are TO, OF, IN, FOR, WITH, ON, and FROM. Prepositions can denote relationships between nouns, pronouns and other parts of speech. Pronouns alone cannot be taken as sentence constituents; they must be used together with nouns, pronouns or other word classes that can be used as nouns. Repetition and ellipsis are very common in pronoun usage, so students should treat pronouns seriously to judge the meaning of pronouns according to the context and word features (Lin Xiangzhou, 1985).

E. Notional Verbs in the Reading Texts in TEM-4

A verb is a part of speech that can be used to express action or state, and is the foundation of a complete discourse. Notional verbs, however, are the key to the meaning of discourse. The top twenty notional verbs in the reading texts of the TEM-4 test are LIKE, WORK, MAKE, GET, SAID, GO, MADE, USED, and SEE. These verbs are most frequently used and can express a writer's attitude or bring in new topics, such as LIKE, or SAY.

F. Modal Forms in the Reading Texts in TEM-4

Modal auxiliaries play an important role in discourse. According to McCarthy & Carter (2005), modal forms from the broad view refer to the attitude of the speaker or writer towards the information or ideas expressed. Various lexical means can express modality. This paper studies modality from the text level, and other words that can perform modality are also considered. Main modal forms are CAN, WOULD, WILL, MAY, SHOULD, and COULD . From this we know that low or middle modal verbs, such as WILL, SHOULD, and WOULD are more frequently used while the high modal verbs, like MUST, which can express strong personal emotions and ideas, are less used. From this, we can conclude that writer attitude is euphemistic and objective.

G. Nouns in the Reading Texts in TEM-4

The top twenty nouns in the reading texts in the TEM-4 test are PEOPLE, TIME, WAY, LIFE, SCHOOL, WORLD, DAY, YEAR, MAN, MUSIC, YEARS, COUNTRY, FAMILY, MEN, THINGS, AMERICAN, HOME, NAME, EXAMPLE, HOUSE. According to Halliday and Hasan (1976), lexical cohesion in English discourse can be classified into reiteration and collocation. Reiteration is when general words or umbrella forms appear in discourse in the forms of synonym, hyponymy, or other forms to make cohesion of the discourse. Many general words, such as PEOPLE, THINGS and MEN are obviously seen. The usage of these words can avoid lack of lexical and strength cohesion in discourse.

On the other hand, this reflects the depth and breadth of the topic source domain. Van Dijk (1977:152) points out that topics can settle down the range of concepts that may be used in discourse or part of discourse, so as to limit lexical insertion. Fowler (1986:39) reports that the topic is the context of reference in discourse and thus lexical cohesion is the result of topic choice. In a complicated discourse, the collection and interrelationships of the lexicon is necessary to maintain the unity and coherence. These general words demonstrate that topics of the reading texts in TEM-4 are more about school education, social life, people, family and culture. Reading comprehension discourse may cover heated topics home and abroad, so students should expand their readings and focus more on what is happening around the country and the world, to better understand a discourse with a more abundant reading background. This represents the so-called "Top-Down" approach in discourse analysis and psycholinguistics.

4. Results of the Reading Texts in TEM-8

A. Types of Vocabularies in the Reading Texts in TEM-8

There are 9,784 *types*, including 2,635 from the Basewrd1.txt vocabulary, comprising 26.93%; 1443 *types* were found in the Basewrd2.txt vocabulary, comprising 14.75% of the total. 1018 *types* come from the Basewrd3.txt vocabulary, accounting for 10.40% of the total. The results returned 4688 *types* not found in any of these three basic vocabularies, for 47.91% of the total. The rate of *tokens* found in each of the three basic vocabularies is 78.62%, 5.60%, and 4.26%, with 11.52% of the tokens not found in any of the vocabularies. From this, we know that most vocabulary is formed from staple words, hypo-ordinary words and academic words.

B. Conjunctions in the Reading Texts in TEM-8

The top appearing conjunctions in the TEM-8 reading texts are AND, AS, BUT, OR, and IF, SO, NOW, and FIRST.

C. Pronouns in the Reading Texts in TEM-8

The top seen pronouns are THAT, IT, HE, THEIR, THEY, HIS, I and THIS. The total percentage of these pronouns is 6.187% of all words, which may be evidence that the reading section of the TEM-8 is more comprehensive and more difficult than in the TEM-4.

D. Prepositions in the Reading Texts in TEM-8

The preposition is one of the word classes that share various usages and is of great number in closed word classes. The top prepositions are OF, TO, IN, FOR, WITH, ON, AT, and FROM. Prepositions in discourse occur in the form of preposition clauses, and "preposition clauses can function as adverbials, modifiers as well as object complement" (Yang, 1985). As a result, prepositional clauses may cause ambiguity in reading discourse, so students should pay special attention to their usage in text.

E. Notional Verbs in the Reading Texts in TEM-8

The top notional verbs are LIKE, WORK, SAID, SEE, MADE, MAKE, GET,NEED, SAY, and GO. This list includes both the present tense and the past tense of some words, including pairs like MAKE-MADE and SAY-SAID.

F. Modal Forms in the Reading Texts in TEM-8

According to Halliday (2000), CAN is one of low value modal auxiliaries, and its basic modal meaning includes providing suggestions, permissions and so on; WILL is a medium value modal auxiliary with strong modal meaning; MUST is a high value modal auxiliary of very strong modal meaning. From the data we find that low and medium value modal forms are more common, possibly reflecting that the writer's attitude is implicit.

G. Nouns in the Reading Texts in TEM-8

The top nouns are PEOPLE, TIME, WORLD, LIFE, and WAY. To some degree, these nouns are topics of discourse. They come from various source domains, including society, economy, culture, general knowledge, education and so on.

The top noun is PEOPLE, and its frequency is 116, accounting for 2.8% of the total word count. This shows that the focus of these discourses is related to human beings and their development. Students preparing for the TEM-8 should not only focus on basic knowledge, but also on reading materials from English-speaking newspapers. Abundant knowledge of what happened home and abroad can help the interpretation of discourses.

5. Comparison of Results of TEM-4 by *RANGE* Analysis

A. Comparison of the Whole Lexicon Use Between the TEM-4 and TEM-8

The findings show that the total number of *types* found in the TEM-8 is twice of that in the TEM-4, and this shows that the reading text in the TEM-8 is composed of more words than the TEM-4 reading discourse and the requirements for reading speed are much higher than the TEM-4 reading. Besides, the percentage of words not in the vocabulary lists for the TEM-8 is 47.91%, which is much higher than the 32.51% result from the TEM-4 vocabulary analysis. This can, to some degree, explain that the readings of the TEM-8 are more difficult than the TEM-4 readings. The finding shows that the independent *t*-test and clearly demonstrates the differences between the two tests after inputting the data from SPSS 19. There is no significant difference between the TEM-4 and TEM-8 readings in terms of *tokens*, *types* and *families*, or the distribution of words on three vocabulary levels. A correlation analysis was also conducted, and the results show that the rate of *tokens*, *types* and *families* in the TEM-4 and TEM-8 at three basic vocabularies is statistically related with r between 0.722 and 0.999.

B. Comparison of Conjunctions in the TEM-4 and TEM-8 Reading Discourses

Conjunctions are frequently used in the TEM-4 and TEM-8 reading discourses, and the top five conjunctions are AND, AS, BUT, OR, IF and SO. These words can be used to express transition, conditional relation and cohesion. The rate of conjunctions is statistically significant, at the 95% confidence interval. The result of the correlation analysis is that the rate of conjunction between TEM-4 and TEM-8 reading discourses is closely related (r= 0.999 (p<0.01, bilateral)).

C. Comparison of Pronouns in the TEM-4 and TEM-8 Reading Discourses

Pronouns are key constituents of discourse cohesion and coherence, and play an important role in interpreting and constructing the meaning and framework of a discourse. In the reading discourses of the TEM-4 and TEM-8, the demonstrative pronoun THAT ranks first in rate column, with a frequency higher than 100. Therefore, students should be sensitive to personal pronouns, and pay close attention to anaphoric and cataphoric pronominals. On the basis of the results of the Paired-Sample t-test of pronouns in the TEM-4 and TEM-8 readings, it is easy to know that there is no significant difference between pronouns in these two tests, and the Pearson correlation coefficient is 0.564.

D. Comparison of Prepositions in the TEM-4 and TEM-8 Reading Discourses

From a paired-sample t-test of prepositions in the TEM-4 and TEM-8 readings it is easy to know that there is a significant difference between the prepositions on these two tests, and the Pearson correlation coefficient is 0.986 (p<0.01) bilateral.

E. Comparison of Notional Verbs in the TEM-4 and TEM-8 Reading Discourses

There are twelve notional verbs in the top twenty of the TEM-4 and TEM-8 readings. According to SPSS 19, it is evident that the incidence of notional verbs is statistically different between these two tests, and the Pearson correlation coefficient is 0.858 (p<0.01) bilateral.

F. Comparison of Modal Forms in the TEM-4 and TEM-8 Reading Discourses

Modality is a subsystem of interpersonal function in systemic functional grammar and undergoes interpersonal meaning. Fowler (1979) points out that modality can reflect the opinions and attitude of the writer or speaker. Quirk considers modality to be like the announcer's judgment towards propositions (Quirk et al., 1985:219). The top five modal forms in the TEM-4 and TEM-8 readings are CAN, WOULD, WILL, MAY, and SHOULD, making up 1.2% of the total words on both tests. This table shows that these modal forms are more frequently used in the TEM-8 than in the TEM-4, meaning that the writer's attitude is more objective and that it may be harder for students to grasp the intention of the writer. After the twelve most common modal forms of the top thirteen in the TEM-4 and TEM-8 tests are inputted into SPSS, it is there is apparent that significant difference in modal forms and the Pearson correlation coefficient is 0.948 (p<0.01) bilateral.

G. Comparison of Nouns in the TEM-4 and TEM-8 Reading Discourses

There are obvious distinctions between the results of these analyses on nouns. The breadth and scope of nouns used in the TEM-8 is much wider than the list found in the TEM-4. The included topics cover education, children, language, economy, family and so on. There is a significant difference of noun usage between these tests, and the Pearson correlation coefficient is 0.920 (p<0.01) bilateral. Therefore, it is suggested that English students should read various materials to enlarge their reading scope and to be more skillful in catching the main ideas and framework.

6. Conclusion

This paper analyzes the reading discourses of the TEM-4 and TEM-8 tests, makes a comparison of lexical use features, taking English majors and the guidance of the TEM-4 and TEM-8 tests as starting points, making full use of the *RANGE* corpus analysis software and SPSS data analysis software. The research results demonstrate that the TEM-4 and TEM-8 tests show no significant difference between the use of pronouns, prepositions and nouns, but do show a significant difference in the use of conjunctions, notional verbs and modal forms. As a result, when preparing for and teaching the TEM-4 and TEM-8 tests, English Majors and English teachers should pay special attention to the function of conjunctions in discourse, various uses of prepositions, different modal meanings of different modal forms, and try to read as much as possible about the outside world. Although, there is no significant difference of pronoun usage between the two tests, the importance of pronouns cannot be ignored.

References

[1] Bao Gui and Wang Xia. "The application of RANGE in the evaluation of productive vocabulary of second language," *Electronic Teaching and Learning of Foreign Language,* vol.8, pp. 54-58, 2005.

[2] Cai Hui. "On the research of interactive factors in context," *PLA Journal of Foreign Languages,* vol.1, pp. 45-48, 2000.

[3] Cheng Shi. "The application of *Range* in the research of text," *Journal of Chinese,* vol. 9, pp.42-46, 2009.

[4] Coxhead, A. "A new academic word list," *TESOL Quarterly,* vol.34, no.2, pp. 213-238, 2000.

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

[6] Halliday, M.A.K and R. Hasan. *Language, Context, and Text: Aspects of Language in a Social-semiotic perspective*. Victoria: Deakin University Press, 1985.

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

[8] Hu Zhuanglin. "The application of discourse analysis to teaching and learning," *Foreign Language Teaching,* vol.1, pp. 1-10, 2001.

[9] Lin Xinagzhou. "The repetition and substitution of English pronouns," *Foreign Language,* vol.4, pp. 40-43, 1985.

[10] Liu Donghong. "The application of vocabulary in English writings" *Modern Foreign Language,* vol.26, no.2, pp. 180-187, 2003.

[11] Michal McCathy and Ronald Carter. *Language as Discourse Perspective for Language Teaching*. Beijing: Beijing Press, 2005.

[12] Nation, P. And A. Coxhead. RANGE. http://www.vuw.ac.nz/lals/staff/Paul_Nation. 2003.

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

[14] Quirk, R. *A Comprehensive Grammar of the English Language*. London: Longman Group Limited, 1985.

[15] Song Hong. *On the Research of Anaphora of Personal Pronouns*. Beijing: National Defense Industry Press, 2010.

[16] Van Dijk, T.A. *Text and Context: Explorations in the Semantics and Pragmatics of Discourse*. London: Longman, 1997.

[17] Wei Jinmei. "Coherence and reading comprehension of text," *Foreign Language Teaching,* vol.1, pp. 45-51, 1996.

[18] Yang Huixin. "Various interpretations of English pronouns in syntax," *Foreign Language,* vol.7, pp. 40-44, 1985.

[19] Zhou Shen. *The New Guidance of English TEM-8 Test*. Shanghai: Shanghai Foreign Language Teaching and Learning Press, 2011.