A Corpus-based Comparative Study of Lexical Proficiency of Writings by Majors of Arts v.s. Those of Science

Ronggen Zhang
Shanghai Publishing and Printing College, Shanghai, China

Abstract—With aids of corpus linguistics technology, WordSmith Tools, Range, and other software such as Coh-Metrix, this paper attempts to give a scientific analysis on lexical proficiency of writings in placement test for vocational college majors of arts and those of science. First, there exist huge lexical proficiency discrepancies between the students, and they should be implanted more stylistic knowledge in writing and be input more original English materials to enlarge their English vocabulary size. Second, those using more referential cohesive devices and more difficult words tend to score lower in their writings, and the students should be instructed to pay attention to the surface coherence, but also to focus more attention to the global coherence. Finally, the students of arts do better than those of science in deep cohesion, and can score higher, despite that their words are usually the basic common words, suggesting that global coherence is of the greatest importance in all types of writings.

Index Terms—English exposition, corpus, lexical proficiency, writing proficiency

I. INTRODUCTION

In the light of recent reform of China’s College English Test (CET for short) policy, the Chinese college students are required to lay more emphasis on discourse reading, textual translation and writing. The salience of textual abilities for the students becomes more and more obvious. Nevertheless, the writing proficiency of the students in CET seems to be lowering in recent years, with an average score of 40 out of 100 for each student’s writing (Li, 2012). Therefore, it is urgent to enhance China’s college students’ writing abilities.

Researches on relationship between lexical knowledge and L2 writing proficiency are abundant. Leki and Carson found that more proficient L2 users use a wider variety of words and more sophisticated (e.g., low-frequency) the size of vocabulary available to the writer plays an important role in L2 writing (Leki and Carson 1994). It was also found more proficient L2 users use a wider variety of words in their writing than do less proficient L2 users (Laufer and Nation, 1995). Crossley and McNamara found the lexical differences between L1 and L2 writers will highlight the restricted lexical proficiency common in L2 learners (Crossley and McNamara, 2009). Baba suggested that different aspects of L2 lexical proficiency have a differential impact on EFL learners’ summary writing (Baba, 2009).

In response to the above researches, this paper attempts to give a scientific analysis on lexical proficiency of writings in the placement test for college majors of arts and those of science, by using both quantitative and the qualitative methods. By adoption of the technology from the corpus linguistics, WordSmith Tools, Range, and other software such as Coh-Metrix are used in this study. Nesi and Gardner aimed to improve understanding of the writing demands placed on today's university students, by using corpus tools, such as WordSmith Tools (Scott, 2010) to analyze variation in keywords, lemmas, and clusters across genre families, disciplines, and student levels. RANGE is used to compare the vocabulary of up to 32 different texts at the same time. RANGE is designed by Paul Nation to provide a table which shows how much coverage of a text each of the three base lists (BASEWRD1, BASEWRD2, and BASEWRD3) provides. With the three base lists, RANGE can provide information on word frequency, such as TOKENS, TYPES, FAMILIES and the like(Nation, 2011). Coh-Metrix is a computational tool that produces indices of the linguistic and discourse representations of a text. These values can be used in many different ways to investigate the cohesion of the explicit text and the coherence of the mental representation of the text. From the homepage of Department of Psychology of University of Memphis (http://cohmetrix.memphis.edu/cohmetrixpr/index.html), calculated by Coh-Metrix, the Text Easability Assessor provides percentile scores on five characteristics of text, including Narrativity, Syntactic Simplicity, Word Concreteness, Referential Cohesion, and Deep Cohesion (Graesser , McNamara and Kulikowich , 2011).
First, narrativity tells a story, with characters, events, places, and things that are familiar to the reader. Then, syntactic simplicity reflects the degree to which the sentences in the text contain fewer words and use simpler, familiar syntactic structures, which are less challenging to process. Third, high word concreteness means that texts that contains content words that are concrete, meaningful, and evoke mental images are easier to process and understand, while texts that contain more abstract words are more challenging to understand. Furthermore, a text with high referential cohesion contains words and ideas that overlap across sentences and the entire text, forming explicit threads that connect the text for the reader; whereas a low cohesion text is typically more difficult to process because there are fewer connections that tie the ideas together for the reader. Finally, deep cohesion reflects the degree to which the text contains causal and intentional connectives when there are causal and logical relationships within the text. If the text is high in deep cohesion, then those relationships and global cohesion are more explicit.

With regard to the above case, this study mainly focuses on the following points:

a) What lexical features the English writings by China’s vocational college students may display.

b) What differences may exist between the writings by Majors of Arts and those of Science.

c) Pedagogical implications of the findings.

II. METHODOLOGY

A. Sampling:

The corpora concerned are based on the 60 pieces of students’ writings, randomly sampled among the 1600 pieces of writings from the placement test for freshmen majors of arts and those of science in Shanghai Publishing & Printing College in September of 2013. The topic of the writing is an expository composition on ‘No Smoking in Public Places’, with each student given a picture in which many people are smoking at a restaurant at the time. And the writing is required to be finished within half an hour.

B. Data Processing:

Including scoring, concordancing, and editing, by using the software such as AntConc 3.2, WordSmith Tools 5.0, Range 32, SPSS 19, Coh-Metrix, etc. Each piece of writing is scored through the scoring system provided by http://pigai.org/guest.php, just for reference.

C. Concepts Concerned in Data Processing:

FILE - file size of each text; Tokens-the running words; Types - distinct words
TOKEN1, TOKEN2, and TOKEN3 belong to BASEWRD1, BASEWRD2, and BASEWRD3 within the three base lists respectively, while TOKEN4 is out of the base list (Nation, 2011).

TYPE1, TYPE2, TYPE3, and TYPE4 corresponding to TOKEN1, TOKEN2, TOKEN3 and TOKEN4 respectively
TTR - type/token ratio; STU1 - majors of arts; STU2 - majors of science
MWL - mean word length (in characters); WLSTD - word length standard deviation
Sentences - the total number of sentences in the text; MSL - mean sentence length (in words)
SLSTD- sentence length standard deviation

According to Templin, Type-token ratio (TTR) is the number of unique words (called types) divided by the number of tokens of these words. Each unique word in a text is considered a word type. Each instance of a particular word is a token (Templin, 1957).

III. DATA ANALYSES AND FINDINGS

A. Descriptive Statistics I

<table>
<thead>
<tr>
<th>MINIMUM</th>
<th>MAXIMUM</th>
<th>MEAN</th>
<th>STD.DEV</th>
<th>TTR</th>
<th>MINIMUM</th>
<th>MAXIMUM</th>
<th>MEAN</th>
<th>STD.DEV</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOKEN1</td>
<td>18.00</td>
<td>140.00</td>
<td>86.6333</td>
<td>21.58779</td>
<td>41.41</td>
<td>73.74</td>
<td>62.3531</td>
<td>6.18722</td>
</tr>
<tr>
<td>TOKEN2</td>
<td>0.00</td>
<td>25.00</td>
<td>11.0167</td>
<td>4.40913</td>
<td>1.73</td>
<td>4.85</td>
<td>4.2811</td>
<td>0.21546</td>
</tr>
<tr>
<td>TOKEN3</td>
<td>0.00</td>
<td>6.00</td>
<td>9667</td>
<td>1.31441</td>
<td>1.70</td>
<td>2.47</td>
<td>2.0307</td>
<td>0.19136</td>
</tr>
<tr>
<td>TOKEN4</td>
<td>1.00</td>
<td>61.00</td>
<td>58000</td>
<td>8.09645</td>
<td>2.00</td>
<td>16.00</td>
<td>9.6667</td>
<td>2.48839</td>
</tr>
<tr>
<td>TYPE1</td>
<td>12.00</td>
<td>81.00</td>
<td>533000</td>
<td>12.46187</td>
<td>6.56</td>
<td>31.50</td>
<td>11.3735</td>
<td>3.70958</td>
</tr>
<tr>
<td>TYPE2</td>
<td>0.00</td>
<td>12.00</td>
<td>56000</td>
<td>2.40198</td>
<td>1.96</td>
<td>9.31</td>
<td>4.9193</td>
<td>1.92132</td>
</tr>
<tr>
<td>TYPE3</td>
<td>0.00</td>
<td>6.00</td>
<td>7833</td>
<td>1.18023</td>
<td>3.70</td>
<td>98.00</td>
<td>77.5667</td>
<td>14.19521</td>
</tr>
<tr>
<td>TYPE4</td>
<td>0.00</td>
<td>14.00</td>
<td>41500</td>
<td>2.94483</td>
<td>14.00</td>
<td>99.00</td>
<td>86.8500</td>
<td>16.36224</td>
</tr>
<tr>
<td>SCORE</td>
<td>19.00</td>
<td>74.00</td>
<td>509167</td>
<td>13.56752</td>
<td>2.00</td>
<td>99.00</td>
<td>33.4000</td>
<td>25.04518</td>
</tr>
<tr>
<td>FILE</td>
<td>343.00</td>
<td>863.00</td>
<td>580800</td>
<td>115.60318</td>
<td>2.00</td>
<td>96.00</td>
<td>29.6333</td>
<td>22.41063</td>
</tr>
<tr>
<td>TKON5</td>
<td>55.00</td>
<td>159.00</td>
<td>1041000</td>
<td>2184436</td>
<td>11.00</td>
<td>100.00</td>
<td>70.8000</td>
<td>28.95397</td>
</tr>
<tr>
<td>TYPES</td>
<td>38.00</td>
<td>92.00</td>
<td>644167</td>
<td>1309327</td>
<td>60.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

© 2014 ACADEMY PUBLISHER
Table I shows, the overall writing proficiency of the 60 vocational college students is far from satisfactory, with a mean score of less than 51 out of 100 and a large score standard deviation of more than 13. What’s more, there exist huge lexical proficiency discrepancies between the students, seen from the high standard deviations of TOKENS, TYPES TTR and FILE. The vast discrepancies are also demonstrated by the great disparity between the percentile scores on five characteristics of each text, including Narrativity, Syntactic Simplicity, Word Concreteness, Referential Cohesion, and Deep Cohesion, i.e. with the standard deviation of 14.19521, 16.36224, 25.04518, 22.41063, and 28.95397 respectively. Considering the means of TOKENS and TYPES, both TOKENS1 and TYPES1 are the highest of the same kind correspondingly; and the mean of TTR is over 62. This illustrates the fact that due to the small vocabulary size of the student, most of the words in the writing are simple common words and the same word reappears very often in the same text. This result may be further confirmed by the high means of Narrativity (77.5667) and Syntactic Simplicity (86.8500). High narrativity indicates that the text is more story-like and may have more familiar words, and high syntactic simplicity means that the text has simple sentence structures and it is easier to process. Nevertheless, and the low mean of Word Concreteness (33.4000) is quite out of the researcher’s expectation, for the low word concreteness means there are many abstract words that are hard to visualize, causing the text to be more difficult to understand. In addition, the low referential cohesion (29.6333) suggests that the reader may have to infer the relationships between sentences and ideas, also causing the text to be more difficult to understand. Anyway, lower referential cohesion may result from the negative transfer of the students’ mother tongue Chinese, which further confirms their low writing proficiency. In Chinese, high frequency of repetitions of words is rather preferred, and so is low referential cohesion.

B. Descriptive Statistics II

Table II displays the comparative descriptive statistics for the 60 vocational college majors of arts and those of science. For one thing, there exist some differences between the two majors in TOKEN1, TYPE1, TYPES, TTR, and TOKEN4. The means of TOKEN1, TYPE1, TYPES, TTR for majors of arts are a bit higher than those of science respectively, while the situation is opposite in terms of the mean of TOKEN4. This illustrates that students of arts are more likely to use more simple familiar words and wider varieties of words in their writings, whereas those of science are more liable to use more abstract unfamiliar words. This may result from their different ways of thinking, where students of arts tend to think in a concrete way while those of science prefer to think in an abstract manner. As to the standard deviations of TOKEN1, TYPE1, TYPES, TTR, and TOKEN4, the students of arts are all lower than those of science correspondingly, indicating that the lexical proficiency discrepancy between the students of arts is smaller than that between those of science. For another thing, the means of Narrativity and Syntactic Simplicity for the majors of arts are a little lower than those of science, while it is on the contrary in view of Word Concreteness, and Referential Cohesion. This suggests that majors of science prefer to write with simple short words and with simple short sentences, and those of arts prefer to use more reference words and more complex sentences in their writings. In the end, the mean of Deep Cohesion for the students of arts (78.9) is much higher than that of science (62.7), suggests that the former are better at using connecting words and other cohesive means to make their writing more coherent than the latter.
C. Homogeneity Test of Variance

| TABLE III. HOMOGENEITY TEST OF VARIANCE STATISTICS OF MAJORS OF ARTS AND SCIENCE |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Levene statistics | SCORE | TOKEN1 | TOKEN2 | TOKEN3 | FILE | TYPE2 | TYPE3 | MSL | Deep Cohesion |
| df1 | 9.03 | 6.02 | 3.01 | 2.00 | 1.00 | 0.01 | 0.00 | 0.00 | 0.00 |
| df2 | 9.03 | 6.02 | 3.01 | 2.00 | 1.00 | 0.01 | 0.00 | 0.00 | 0.00 |
| Significance | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |

Through comparisons between the scores of arts and those of science by One-Way ANOVA analysis, it can be seen that the scores between the two majors are not significantly different (its significance is 0.997, far greater than 0.05), which may be due to many factors, such as the students’ individual intelligence, motivation, and some other environmental factors. However, there are two variances (SLSTD and Deep Cohesion) significantly different, suggesting that the students of two majors demonstrate strong differences in sentence length standard deviation and Deep Cohesion, which has been mentioned in the above paragraph. That is, majors of science prefer to write simple sentences, whereas those of arts rather than use complex sentences on the one hand; the former pay less attention to the coherence of the writing, whereas the latter do better at making their writing more coherent on the other hand.

D. Pearson Correlation 1

| TABLE IV. PEARSON CORRELATION FOR MAJORS OF ARTS AND SCIENCE |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| STU | SEX | TOKEN1 | TOKEN2 | TOKEN3 | FILE | TYPE2 | TYPE3 | MSL | Deep Cohesion |
| 0.34 | 0.05 | 0.02 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.05 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

© 2014 ACADEMY PUBLISHER
E. Pearson Correlation II

Table IV shows, there exist quite a few significant positive or negative correlations between the variances concerned. First, the two majors of students display a negative correlation in terms of Deep Cohesion (−.282*), indicating the majors of arts do better than those of science in coherence as confirmed above.

Second, score positively correlates with TOKE NKS (.689**), TYPES (.695**), TOKEN1 (.566**), TOKEN2 (.408**), TOKEN3 (.556**), TYPE1 (.569**), TYPE2 (.346**), WLSTD (.353**), FILE (.724**); but it negatively correlates with Referential Cohesion (−.403**), SEX (−.172), and TOKEN4 (−.001). This means those students with better mastering of base words and in more various wordings can score higher. However, it is surprising that those using more referential cohesive devices and more difficult words (out of the basewords list) tend to score lower in their writings. This finding partly corresponds to that of Liang, who found that due to the overuse of some surface features, cohesive referential cohesive devices and more difficult words (out of the basewords list) tend to score lower in their writings.

Lastly, Deep Cohesion negatively correlates with STU (−.282*), SEX (−.145), and TTR (−.008); while it positively correlates with TOKEN2 (279*), TYPE2 (270*), SCORE (207), TOKENS (0.69), and TYPES (0.84). This indicates that, above all, students of arts do better than those of science in deep cohesion, i.e. the former, especially the females prefer to use more logic connectives in writing and can score higher, despite that their words are usually the basic common words. This finding is of especially pedagogical implication, to be further detailed in the latter part of the paper.

© 2014 ACADEMY PUBLISHER
Table V is a comparison of the two majors of students in terms of the Pearson correlations between score and other variances. It is easy to see that the two majors of students share many similarities in writing, such as their significant positive correlations of TOKEN1, TYPE1, TYPE2, FILE, TOKENS, and TYPES in terms of score. Nevertheless, the two majors display some discrepancy in their significant correlations of TOKEN2, TOKEN3, TYPE3, and Referential Cohesion, either positively or negatively, in terms of score. That is, on the one hand, both majors of students’ writing proficiency matches with using common words from the baseword1 list or the baseword2 list; on the other hand, students of science with higher writing proficiency prefer to use more complex words from the baseword3 list, while those of arts doing better in local coherence get lower scores in writing as mentioned above.

IV. PEDAGOGICAL IMPLICATIONS

A. Stylistics

Stylistics is a branch of linguistics, the study and interpretation of texts in regards to their linguistic and tonal style (Widdowson, 1992). Style is classically viewed as using proper words in proper places by Jonathan Swift. Style is generated at two levels: lexical level and syntactic level (Huang, 2005). Each style is exercised in two forms (written and oral). There is also strong iconicity in expository text construction (Xu, 2011). As mentioned above, our vocational college students’ writings display low word concreteness, that is, there lack concrete content words to evoke iconic of the reader, causing the writings more difficult to understand. Therefore, the students should be implanted more stylistic knowledge in various types of writings such as narration, argumentation, description, exposition, and the like.

B. Language Transfer

Language transfer refers to speakers or writers applying knowledge from their native language to a second language. Negative transfer occurs when speakers and writers transfer items and structures that are not the same in both languages (Nitschke, 2010). Corder claims that it is a generally agreed observation that many but not necessarily all the idiosyncratic sentences of a second language learner bear some sort of regular relation to the sentences of his mother tongue (Corder, 1981). Negative transfer occurs most often at the early stage of language acquisition, especially among L2 learners of lower proficiency. As mentioned above, our vocational college students demonstrate lower referential cohesion, which may result from negative transfer of Chinese, i.e. preferring high frequency of repetitions of words in Chinese rather than frequent use of referential cohesive devices in English. Hence, it is necessary to minimize the impacts of negative transfer of mother tongue in teaching Chinese students writing courses by inputting them more original English materials and by enlarging their English vocabulary size.

C. Cohesion

Cohesion is the grammatical and lexical linking within a text or sentence that holds a text together and gives it meaning. Cohesion takes place on two levels: lexical level and grammatical level. M. A. K. Halliday and R. Hasan identify five general categories of cohesive devices that create coherence in texts: reference, ellipsis, substitution, lexical cohesion and conjunction (Halliday and Hasan, 1976). As mentioned above, our Chinese students display low referential cohesion suggesting that the reader may have to infer the relationships between sentences and ideas, and also causing the text to be more difficult to understand. Despite that lower referential cohesion may result from the negative transfer of the students’ mother tongue, it may also be due to their lack of knowledge on English cohesion. As mentioned above, the score of the student’s writing negatively correlates with Referential Cohesion, confirming that the overuse of some surface features, cohesive ties such as pronouns and connectives do not contribute much to the coherence in the texts (Liang, 2006). Therefore the students should be instructed to pay attention to the surface coherence, but also to focus more attention to the global coherence, i.e. to let the writing come to the point.

V. CONCLUSION

This paper has attempted to investigate what lexical features the English writings by China’s vocational college students may display and what differences may exist between the writings by Majors of Arts and those of Science. First, it is found that, as expected, the overall writing proficiency of the 60 vocational college students is far from satisfactory, that there exist huge lexical proficiency discrepancies between the students, and that due to the small
vocabulary size of the student, most of the words in the writing are simple common words and the same word reappears very often in the same text. And the paper assumes that lower referential cohesion in the findings may result partly from the negative transfer of the students’ mother tongue, and partly from the students’ lack of knowledge concerned. Therefore it is suggested that the students should be implanted more stylistic knowledge in various types of writing such as narration, argumentation, description, exposition, and the like, and that they should be instructed to input more original English materials and enlarge their English vocabulary size.

Second, it is also found that those using more referential cohesive devices and more difficult words tend to score lower in their writings, this partly corresponding to that of Liang, who found that due to the overuse of some surface features, cohesive ties such as pronouns and connectives do not contribute much to the coherence in EFL writers texts, and that high-proficiency EFL writers’ texts tend to be globally more coherent, while low-proficiency EFL writers texts are likely to be locally more coherent (Liang, 2006). Hence, the students should be instructed to pay attention to the surface coherence, but also to focus more attention to the global coherence, i.e. to let the writing come to the point.

Finally, it is found that students of arts do better than those of science in deep cohesion, i.e. the former, especially the females prefer to use more logic connectives in writing and can score higher, despite that their words are usually the basic common words. This suggests that clarity and accuracy are the core principles in expository writings, but what’s more important, global coherence is of the most importance in all types of writings, and this further confirms the above findings.

REFERENCES


Ronggen Zhang, Associate professor, holds an MA degree in ESP from University of Shanghai for Science and Technology in China. His research interest lies in corpus linguistics and investigation of language teaching strategies. He is now teaching in Shanghai Publishing and Printing College. He has recently published in JCT (Journal of Convergence Information Technology).