ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 4, No. 4, pp. 743-749, April 2014
© 2014 ACADEMY PUBLISHER Manufactured in Finland.
doi:10.4304/tpls.4.4.743-749

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). Pre-task planning condition was found to have a small significant effect on writing fluency, whereas pre-task planning condition was found to have no impact on lexical complexity and grammatical complexity(Johnson , Mercado and Acevedo 2012). Kormos discussed the role of three important individual difference factors, aptitude, working memory capacity, and motivation, in the different stages of writing and the processes of learning through writing (Kormos, 2012).

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

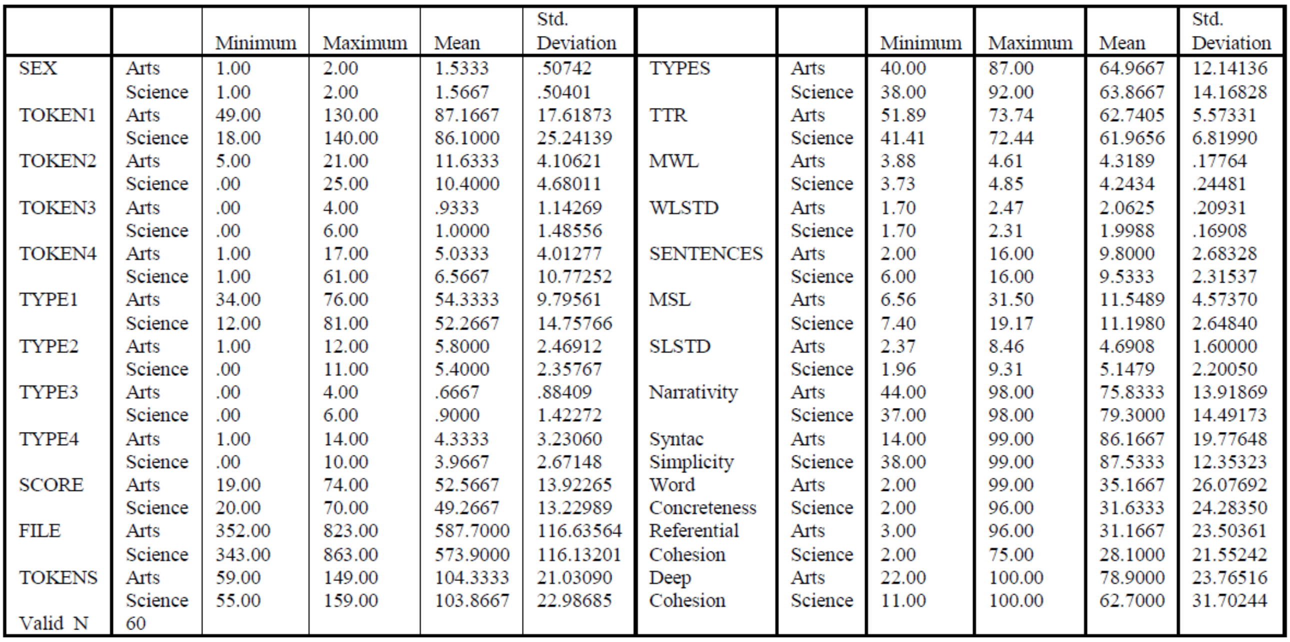
TABLE I.
DESCRIPTIVE STATISTICS OF MAJORS OF ARTS AND SCIENCE

	Minimum	Maximum	Mean	Std.Deviation		Minimum	Maximum	Mean	Std.Deviation
TOKEN1	18.00	140.00	86.6333	21.58779	TTR	41.41	73.74	62.3531	6.18722
TOKEN2	.00	25.00	11.0167	4.40913	MWL	3.73	4.85	4.2811	.21546
TOKEN3	.00	6.00	.9667	1.31441	WLSTD	1.70	2.47	2.0307	.19136
TOKEN4	1.00	61.00	5.8000	8.09645	SENTENCES	2.00	16.00	9.6667	2.48839
TYPE1	12.00	81.00	53.3000	12.46187	MSL	6.56	31.50	11.3735	3.70958
TYPE2	.00	12.00	5.6000	2.40198	SLSTD	1.96	9.31	4.9193	1.92132
TYPE3	.00	6.00	.7833	1.18023	Narrativity	37.00	98.00	77.5667	14.19521
TYPE4	.00	14.00	4.1500	2.94483	SyntacSimplicity	14.00	99.00	86.8500	16.36224
SCORE	19.00	74.00	50.9167	13.56752	WordConcreteness	2.00	99.00	33.4000	25.04518
FILE	343.00	863.00	580.8000	115.60318	ReferentialCohesion	2.00	96.00	29.6333	22.41063
TOKENS	55.00	159.00	104.1000	21.84436	DeepCohesion	11.00	100.00	70.8000	28.95397
TYPES	38.00	92.00	64.4167	13.09327	Valid N	60			

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.
COMPARATIVE DESCRIPTIVE STATISTICS OF MAJORS OF ARTS AND SCIENCE



		Minimum	Maximum	Mean	Std. Deviation			Minimum	Maximum	Mean	Std. Deviation
SEX	Arts	1.00	2.00	1.5333	.50742	TYPES	Arts	40.00	87.00	64.9667	12.14136
	Science	1.00	2.00	1.5667	.50401		Science	38.00	92.00	63.8667	14.16828
TOKEN1	Arts	49.00	130.00	87.1667	17.61873	TTR	Arts	51.89	73.74	62.7405	5.57331
	Science	18.00	140.00	86.1000	25.24139		Science	41.41	72.44	61.9656	6.81990
TOKEN2	Arts	5.00	21.00	11.6333	4.10621	MWL	Arts	3.88	4.61	4.3189	.17764
	Science	.00	25.00	10.4000	4.68011		Science	3.73	4.85	4.2434	.24481
TOKEN3	Arts	.00	4.00	.9333	1.14269	WLSTD	Arts	1.70	2.47	2.0625	.20931
	Science	.00	6.00	1.0000	1.48556		Science	1.70	2.31	1.9988	.16908
TOKEN4	Arts	1.00	17.00	5.0333	4.01277	SENTENCES	Arts	2.00	16.00	9.8000	2.68328
	Science	1.00	61.00	6.5667	10.77252		Science	6.00	16.00	9.5333	2.31537
TYPE1	Arts	34.00	76.00	54.3333	9.79561	MSL	Arts	6.56	31.50	11.5489	4.57370
	Science	12.00	81.00	52.2667	14.75766		Science	7.40	19.17	11.1980	2.64840
TYPE2	Arts	1.00	12.00	5.8000	2.46912	SLSTD	Arts	2.37	8.46	4.6908	1.60000
	Science	.00	11.00	5.4000	2.35767		Science	1.96	9.31	5.1479	2.20050
TYPE3	Arts	.00	4.00	.6667	.88409	Narrativity	Arts	44.00	98.00	75.8333	13.91869
	Science	.00	6.00	.9000	1.42272		Science	37.00	98.00	79.3000	14.49173
TYPE4	Arts	1.00	14.00	4.3333	3.23060	Syntac	Arts	14.00	99.00	86.1667	19.77648
	Science	.00	10.00	3.9667	2.67148	Simplicity	Science	38.00	99.00	87.5333	12.35323
SCORE	Arts	19.00	74.00	52.5667	13.92265	Word	Arts	2.00	99.00	35.1667	26.07692
	Science	20.00	70.00	49.2667	13.22989	Concreteness	Science	2.00	96.00	31.6333	24.28350
FILE	Arts	352.00	823.00	587.7000	116.63564	Referential	Arts	3.00	96.00	31.1667	23.50361
	Science	343.00	863.00	573.9000	116.13201	Cohesion	Science	2.00	75.00	28.1000	21.55242
TOKENS	Arts	59.00	149.00	104.3333	21.03090	Deep	Arts	22.00	100.00	78.9000	23.76516
	Science	55.00	159.00	103.8667	22.98685	Cohesion	Science	11.00	100.00	62.7000	31.70244
Valid N	60										

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

	SCORE	TYPE2	Referential Cohesion	TOKEN2	SEN TENCES	TOKEN3	FILE	TYPE4	Narra tivity	SEX	TTR	Word Concreteness
Levene statistics	0	0.017	0.017	0.026	0.074	0.085	0.2	0.206	0.224	0.236	0.363	0.382
df1	1	1	1	1	1	1	1	1	1	1	1	1
df2	58	58	58	58	58	58	58	58	58	58	58	58
Significance	0.997	0.896	0.896	0.872	0.786	0.772	0.657	0.651	0.638	0.629	0.549	0.539
	TOKENS	TYPES	WLSTD	Syntac Simplicity	TYPE3	TOKEN4	MSL	MWL	TYPE1	TOKEN1	SLSTD	Deep Cohesion
Levene statistics	0.672	0.825	0.874	1.013	1.083	1.1	1.157	1.627	2.739	2.791	4.86	6.67
df1	1	1	1	1	1	1	1	1	1	1	1	1
df2	58	58	58	58	58	58	58	58	58	58	58	58
Significance	0.416	0.367	0.354	0.318	0.302	0.299	0.287	0.207	0.103	0.1	0.031	0.012

Through comparisons between the majors 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, while 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 I

TABLE IV.
PEARSON CORRELATION FOR MAJORS OF ARTS AND SCIENCE

	STU	SEX	TOKEN1	TOKEN2	TOKEN3	TOKEN4	TYPE1	TYPE2	TYPE3	TYPE4	SCORE	FILE	TOKENS	TYPES	TTR
STU	1	.034	-.025	-.141	.026	.095	-.084	-.084	.100	-.063	-.123	-.060	-.011	-.042	-.063
SEX	.034	1	-.213	.049	-.177	.082	-.171	.045	-.139	-.057	-.172	-.166	-.181	-.118	.198
SCORE	-.123	-.172	.566**	.408**	.356**	-.001	.569**	.507**	.346**	-.108	1	.724**	.689**	.695**	-.094
TOKEN1	-.025	-.213	1	.454**	-.083	-.419**	.922**	.381**	-.130	.013	.566**	.790**	.833**	.692**	-.354**
TOKEN2	-.141	.049	.454**	1	-.155	-.302*	.396**	.694**	-.162	.048	.408**	.496**	.464**	.313*	-.305*
TOKEN3	.026	-.177	-.083	-.155	1	.482**	-.064	-.042	.935**	.054	.356**	.301*	.239	.318*	.116
TOKEN4	.095	.082	-.419**	-.302*	.482**	1	-.410**	-.200	.559**	.254*	-.001	.150	.115	.198	.132
TYPE1	-.084	-.171	.922**	.396**	-.064	-.410**	1	.445**	-.075	.097	.569**	.723**	.750**	.790**	.001
TYPE2	-.084	.045	.381**	.694**	-.042	-.200	.445**	1	-.019	.253	.507**	.447**	.388**	.484**	.157
TYPE3	.100	-.139	-.130	-.162	.935**	.559**	-.075	-.019	1	.058	.346**	.278*	.230	.364**	.224
TYPE4	-.063	-.057	.013	.048	.054	.254*	.097	.253	.058	1	-.108	.164	.115	.244	.227
FILE	-.060	-.166	.790**	.496**	.301*	.150	.723**	.447**	.278*	.164	.724**	1	.983**	.879**	-.298*
TOKENS	-.011	-.181	.833**	.464**	.239	.115	.750**	.388**	.230	.115	.689**	.983**	1	.874**	-.342**
TYPES	-.042	-.118	.692**	.313*	.318*	.198	.790**	.484**	.364**	.244	.695**	.879**	.874**	1	.148
TTR	-.063	.198	-.354**	-.305*	.116	.132	.001	.157	.224	.227	-.094	-.298*	-.342**	.148	1
MWL	-.177	.139	-.390**	.004	.285*	.133	-.287*	.133	.212	.221	-.047	-.152	-.321*	-.176	.334**
WLSTD	-.168	-.106	-.180	-.052	.455**	.248	-.105	.255*	.360**	.150	.353**	.073	-.017	.121	.250
SENTENCES	-.054	-.054	.374**	.223	.193	.125	.274*	.170	.183	.007	.236	.506**	.510**	.374**	-.319*
MSL	-.048	-.017	.129	.033	-.055	-.001	.180	.099	-.043	.119	.082	.108	.121	.184	.148
SLSTD	.120	-.043	.449**	.238	-.070	-.033	.437**	.098	-.084	.159	.254	.403**	.454**	.411**	-.150
Narrativity	.123	.010	.070	-.015	-.333**	-.260*	-.047	-.274*	-.401**	-.242	-.090	-.171	-.097	-.270*	-.366**
Syntac Simplicity	.042	.080	.057	.084	.084	-.151	-.006	-.073	.040	-.245	.124	.014	.017	-.092	-.237
Word Concreteness	-.071	-.025	-.177	.071	-.026	.049	-.079	.154	.085	.107	-.060	-.103	-.141	-.014	.271*
Referential Cohesion	-.069	.175	-.263*	.002	-.256*	-.243	-.413**	-.408**	-.326*	-.348**	-.403**	-.422**	-.409**	-.651**	-.394**
Deep Cohesion	-.282*	-.145	.061	.279*	-.100	-.069	.097	.270*	-.053	.022	.207	.077	.069	.084	-.008

(TO BE CONTINUED)

	MWL	WLSTD	SENTENCES	MSL	SLSTD	Narrativity	Syntac Simplicity	Word Concreteness	Referential Cohesion	Deep Cohesion
STU	-.177	-.168	-.054	-.048	.120	.123	.042	-.071	-.069	-.282*
SEX	.139	-.106	-.054	-.017	-.043	.010	.080	-.025	.175	-.145
SCORE	-.047	.353**	.236	.082	.254	-.090	.124	-.060	-.403**	.207
TOKEN1	-.390**	-.180	.374**	.129	.449**	.070	.057	-.177	-.263*	.061
TOKEN2	.004	-.052	.223	.033	.238	-.015	.084	.071	.002	.279*
TOKEN3	.285*	.455**	.193	-.055	-.070	-.333**	.084	-.026	-.256*	-.100
TOKEN4	.133	.248	.125	-.001	-.033	-.260*	-.151	.049	-.243	-.069
TYPE1	-.287*	-.105	.274*	.180	.437**	-.047	-.006	-.079	-.413**	.097
TYPE2	.133	.255*	.170	.099	.098	-.274*	-.073	.154	-.408**	.270*
TYPE3	.212	.360**	.183	-.043	-.084	-.401**	.040	.085	-.326*	-.053
TYPE4	.221	.150	.007	.119	.159	-.242	-.245	.107	-.348**	.022
FILE	-.152	.073	.506**	.108	.403**	-.171	.014	-.103	-.422**	.077
TOKENS	-.321*	-.017	.510**	.121	.454**	-.097	.017	-.141	-.409**	.069
TYPES	-.176	.121	.374**	.184	.411**	-.270*	-.092	-.014	-.651**	.084
TTR	.334**	.250	-.319*	.148	-.150	-.366**	-.237	.271*	-.394**	-.008
MWL	1	.455**	-.264*	.030	-.281*	-.402**	-.126	.255*	.086	.050
WLSTD	.455**	1	-.131	.113	-.151	-.255*	-.085	-.010	-.294*	.217
SENTENCES	-.264*	-.131	1	-.651**	-.208	.034	.449**	-.270*	-.304*	-.139
MSL	.030	.113	-.651**	1	.428**	-.270*	-.717**	.308*	.150	.236
SLSTD	-.281*	-.151	-.208	.428**	1	.202	-.157	-.175	-.065	.071
Narrativity	-.402**	-.255*	.034	-.270*	.202	1	.418**	-.391**	.318*	-.059
Syntac Simplicity	-.126	-.085	.449**	-.717**	-.157	.418**	1	-.357**	-.098	-.096
Word Concreteness	.255*	-.010	-.270*	.308*	-.175	-.391**	-.357**	1	.070	.157
Referential Cohesion	.086	-.294*	-.304*	.150	-.065	.318*	-.098	.070	1	.054
Deep Cohesion	.050	.217	-.139	.236	.071	-.059	-.096	.157	.054	1

*. Correlation is significant at the 0.05 level (2-tailed); **. Correlation is significant at the 0.01 level (2-tailed)

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 TOKENS (.689**), TYPES (.695**), TOKEN1 (.566**), TOKEN2 (.408**), TOKEN3 (.356**), 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 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). Lastly, female students score higher than male ones, as generally expected, thanks to the females' more interest and harder work in English as usual.

Third, Word Concreteness positively correlates with TTR (.271*), and MWL (.255*), yet negatively correlates with Syntactic Simplicity (-.357**), SENTENCES (-.270*), and SCORE (-.060). This indicates that those prefer to use more content words tend to use wider varieties of longer multi-syllabic words. Meanwhile, they also like to write longer and more complex sentences, but out of their expectations their writings usually result in lower scores.

Fourth, Referential Cohesion negatively correlates with many variances such as TYPES (-.651**), FILE (-.422**), TYPE1 (-.413**), TOKENS (-.409**), TYPE2 (-.408**), SCORE (-.403**), TTR (-.394**), SENTENCES (-.304*), TOKEN3 (-.256*), TOKEN4 (-.243), TYPE3(-.326*), and TYPE4(-.348**). This suggests that those pay more attention to Referential Cohesion tend to use fewer varieties of lexemes and less difficult words; despite they do better in local coherence, they tend to get lower scores in writing as mentioned in above paragraphs (Liang, 2006).

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 (.069), and TYPES (.084). 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.

E. Pearson Correlation II

TABLE V.
PEARSON CORRELATION FOR MAJORS OF ARTS VS MAJORS OF SCIENCE

Pearson R	TOKEN1	TOKEN2	TOKEN3	TOKEN4	TYPE1	TYPE2	TYPE3	TYPE4	FILE	TOKENS	TYPES
SCORE-ARTS	.703**	.518**	.215	.033	.687**	.550**	.212	.067	.746**	.728**	.705**
SCORE-SCI	.488**	.289	.490**	.004	.505**	.449*	.481**	-.351	.699**	.662**	.694**
Pearson R	TTR	MWL	WLSTD	SENTENCES	MSL	SLSTD	Narrativity	Syntac Simplicity	Word Concreteness	Referential Cohesion	Deep Cohesion
SCORE-ARTS	-.230	.001	.304	.139	-.027	.270	.093	.309	-.198	-.575**	.063
SCORE-SCI	.006	-.126	.389*	.345	.272	.283	-.247	-.165	.077	-.234	.278

*. CORRELATION IS SIGNIFICANT AT THE 0.05 LEVEL (2-TAILED); **. CORRELATION IS SIGNIFICANT AT THE 0.01 LEVEL (2-TAILED)

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 a11 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 using 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

[1] Baba, Kyoko. (2009).Aspects of lexical proficiency in writing summaries in a foreign language. *Journal of Second Language Writing*, 18, 191-208.

[2] Corder, S. P. (1981). Error Analysis and Interlanguage. Oxford University Press, UK

[3] Crossley, S.A., McNamara, D. S. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing*, 18, 119-135.

[4] Graesser, A.C., McNamara, D.S., Kulikowich, J. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40, 223-234. http://cohmetrix.memphis.edu/cohmetrixpr/cohmetrix3.html#References (accessed 21/12/2013)

[5] Huang, A., H. (2005). Stylistic Examination of English at the Lexical Level. MA Thesis of WUT (Wuchang University of Technology, China).

[6] Halliday, M. A. K., Hasan R. (1976). Cohesion in English. London: Longman.

[7] Johnson, M. D., Mercado, L., Acevedo, A. (2012).The effect of planning sub-processes on L2 writing fluency, grammatical complexity, and lexical complexity. *Journal of Second Language Writing*, 21, 264-282.

[8] Kormos, Judit. (2012). The role of individual differences in L2 writing. *Journal of Second Language Writing*, 21, 390-403.

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

[10] Leki, I., Carson, J. (1994). Students' perceptions of EAPwriting instruction and writing needs across the discipline. *TESOL Quarterly*, 28, 81–101.

[11] Liang, Maocheng. (2006). A study of coher ence in EFL learner s written production. *Modern Foreign Languages (Quarterly)*, 29, 284-285.

[12] Li, Y. J. (2012). Negative echoic effects of the writing topic of college English test band 4 - some afterthought of a CET4 writing rater. *College English Academic Edition*, 99, 273-277.

[13] Nitschke, S., Kidd, E., Serratrice, L. (2010). First language transfer and long-term structural priming in comprehension. *Language and Cognitive Processes*, 25, 94–114.

[14] Nation, Paul. (2011). Range program with British National Corpus list [CP]. http://www.victoria.ac.nz/lals/about/staff/paul-nation (accessed 23/11/2013).

[15] Nitschke, S., Kidd, E., Serratrice, L. (2010). First language transfer and long-term structural priming in comprehension. *Language and Cognitive Processes*, 25, 94–114.

[16] Scott, M. (2010). WordSmith Tools Version 5, Oxford University Press, Oxford, UK.

[17] Templin, M. (1957). Certain language skills in children: Their development and interrelationships. Minneapolis, MN: The University of Minnesota Press. http://cohmetrix.memphis.edu/cohmetrixpr/cohmetrix3.html#References (accessed 21/12/2013).

[18] Widdowson, H. (1992). Practical Stylistics. London: Oxford University Press. http://uastudent.com/stylistics-theoretical-issues-of-stylistics/ (accessed 21/12/2013).

[19] Xu, H., E. (2011). The Implications of Language Iconicity for the English Expository Writing Teaching -Taking LinYutang's Human Life like a Poem as an Example. *Education and Teaching Research*, 25, 90-93.

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 JCIT (Journal of Convergence Information Technology).