A Corpus-based Study on the Lexical Complexity of Writings by Business English Majors in Higher Vocational Colleges

Ronggen Zhang

Shanghai Publishing and Printing College , Shanghai, 200093
zrgen@163.com

ABSTRACT. Based on data from two corpora of 82 pieces of writings by Business English majors from a higher vocational college in Shanghai, the paper attempts to make a study of the lexical complexity of those writings. All the data are processed oneline by the Web-based Lexical Complexity Analyzer, and then processed with the corpus tool AntConc and IBM SPSS Statistics 20. The findings include: First, the students's proficiency is getting improved in their writing with time passing and through more writing pratice. Second, the writings by vocational college students are lack of nouns and verbs in varied forms. Third, their writings are sometimes full of adverbs in colloquial form. Finally, some corresponding pedagogical suggestions are put forward on how to improve teaching the writing course.

KEYWORDS: Corpus, Lexical Complexity, Business English Majors, Higher Vocational College

1. Introduction

Lexemes are the building blocks of language. Needles to say, researches on lexemes have always been hot topics to linguists, language teachers, and even computer scientists. In the 1980s , lexical complexity became appealing to the experts concerned[1]. According to Dr. Lu, lexical complexity entails lexical density, lexical sophistication, and lexical Variation[2] [3]. Lexical density is the proportion of the text made up of lexical word tokens, including nouns, lexical verbs, adjectives, and adverbs[4]. Lexical sophistication refers to“the proportion of relatively unusual or advanced words in the learner’s text”[5]. Lexical Variation, also called lexical diversity, or lexical range, refers to the range of a learner’s vocabulary as displayed in his or her language use[2]. And the measure of lexical variation is the number of different words(NDW) in a language sample.

Johnson studies research synthesis and quantitative meta-analysis, contributing to recent L2 writing research on task complexity and its impact on the syntactic complexity, accuracy, lexical complexity, and the study suggests that features of task complexity may promote attention to the formulation and monitoring systems of the writing process[6]. Frear,and Bitchener's study reports the findings of a within-subject experimental study that examined the relationship between increases in cognitive task complexity and the writing of intermediate L2 writers of English [7]. Lahmann concludes that lexical complexity at the stage of L2 ultimate attainment is the result of a complex interplay of variables general to language learning and performance rather than L2 specific[8]. Tabari's findings reveal that choosing suitable task-based implementational conditions can help L2 writers improve their lexical complexity[9].

Zhang investigates the features of Chinese and international English learners' lexical complexity, based on the copora of SWECCL, ICLE, and BNC[10]. Jin finds genre has a major effect on lexical complexity, which is reflected in the fact that the lexical complexity of argumentative essays is significantly greater than that of narrative essays[11]. Bao thinks, there is no interaction between school type and writing proficiency on lexical complexity[12]. Zhang finds the students prefer to use overlapping pronouns and content words in their writings[13].

Since the subjects of the researches on lexical complexity mentioned above are mainly university undergraduates or postgraduate students, few researches involving vocational college students are found. Hence, this paper attempts to fill in this gap. This study is based on data from two corpora of 82 pieces of writings by Business English majors from a higher vocational college in Shanghai. Those 82 pieces of writings are done in the 1st term and the 2nd term of their fresh year respectively.The topics of the writing are "My Campus Life" in the 1st term, and,"My View on Money", and "Rising Divorce Rates in China"in the 2nd term. All the writings are done online and scored through the scoring system provided by http://pigai.org/guest.php, just for reference.

This study attempts to answer the following questions: What are the lexical features of writings by Business English majors in higher vocational college in China? How are those lexical features correlated with the proficiency of theirs writings? What are the pedagogical implications in teaching English writing course?

2. Data Processing and Analysis

All the data are processed oneline by the Web-based Lexical Complexity Analyzer [2] [3].The terms concerned are as follows:

Lxical density(LD), Lexical sophistication -I (LS1), Lexical sophistication-II (LS2),Number of different words (NDW), NDWERZ (expected random 50) (NDW-ER50), Type/Token ratio (TTR), Corrected TTR(CTTR), Verb variation-1 (VV1), Lexical word variation (LV),Verb variation-II (VV2),Noun variation (NV),Adjective variation (ADJV),Adverb variation (ADVV),Modifier variation (MODV).

Here are the definitions of some of the terms, for measuring lexical complexity.

Lexical sophistication -I (LS1), refers to the ratio of the number of tokens of sophisticated lexical words to the number of tokens of lexical words. Lexical sophistication-II (LS2), refers to the ratio of the number of types of sophisticated words(Ts) to the number of types of words(T). Type/Token ratio (TTR), is the ratio of the number of words types to the number of words in a text. Corrected TTR(CTTR), is the ratio of the number of words types to the square root of the twofold number of words in a text. VV1 refers to the ratio of the number of types of verbs to the number of tokens of verbs. VV2 refers to the ratio of the number of types of verbs to the number of tokens of lexical words. NV refers to the ratio of the number of types of nouns to the number of tokens of nouns. And so on, for the definition of ADJV, ADVV), and MODV.

Table 1 Correlation between Lexical Complexity Measures for the 1st Term

	Score	LD	LS1	LS2	NDW	NDWERZ	TTR	CTTR	VV1	LV	VV2	NV	ADJV	ADVV	MODV
Score	1	-0.048	-.416**	-.432**	0.193	-0.295	-.319*	0.034	-0.18	-0.275	-0.171	-.384*	0.143	0.252	0.28
LD	-0.048	1	0.139	0.108	-0.258	.371*	0.24	-0.101	-0.265	-0.238	-0.111	-0.27	-0.213	0.044	-0.143
LS1	-.416**	0.139	1	.822**	-.308*	.407**	.547**	0.048	.357*	.476**	0.277	.337*	-0.054	-0.185	-0.14
LS2	-.432**	0.108	.822**	1	-0.128	0.268	.316*	0.098	0.229	.363*	0.244	.376*	-0.062	-0.126	-0.116
NDW	0.193	-0.258	-.308*	-0.128	1	-0.041	-.481**	.787**	-0.164	-0.222	-0.001	-0.197	0.134	.317*	0.302
NDWERZ	-0.295	.371*	.407**	0.268	-0.041	1	.728**	.457**	0.278	.383*	.360*	0.214	-0.177	-0.024	-0.151
TTR	-.319*	0.24	.547**	.316*	-.481**	.728**	1	0.15	.467**	.742**	.378*	.595**	-0.112	-0.153	-0.168
CTTR	0.034	-0.101	0.048	0.098	.787**	.457**	0.15	1	0.133	0.251	0.242	0.172	0.088	0.237	0.228
VV1	-0.18	-0.265	.357*	0.229	-0.164	0.278	.467**	0.133	1	.630**	0.302	.336*	0.12	0.013	0.127
LV	-0.275	-0.238	.476**	.363*	-0.222	.383*	.742**	0.251	.630**	1	.482**	.868**	0.228	-0.039	0.18
VV2	-0.171	-0.111	0.277	0.244	-0.001	.360*	.378*	0.242	0.302	.482**	1	0.291	-0.167	-.353*	-.346*
NV	-.384*	-0.27	.337*	.376*	-0.197	0.214	.595**	0.172	.336*	.868**	0.291	1	0.207	-0.042	0.152
ADJV	0.143	-0.213	-0.054	-0.062	0.134	-0.177	-0.112	0.088	0.12	0.228	-0.167	0.207	1	-0.018	.763**
ADVV	0.252	0.044	-0.185	-0.126	.317*	-0.024	-0.153	0.237	0.013	-0.039	-.353*	-0.042	-0.018	1	.630**
MODV	0.28	-0.143	-0.14	-0.116	0.302	-0.151	-0.168	0.228	0.127	0.18	-.346*	0.152	.763**	.630**	1

* P < 0.05; ** p < 0.01

Table 1 shows, in the 1st term, first, there exist negative correlations between Score and LS1(-.432**), LS2 (-.432**), TTR (-.319*), and NV (-.384*); that is, the students better at writing are more likely to use less sophisticated words, especially the simple noun words. Second, there are the positive correlations between LS1 and TTR(.547**), VV1 (.357*), LV (.476**), and NV (.337*), which suggests the writings with higher lexical density are usually full of verbs and nouns in varied forms. Third, the positve corrections beween MODV and ADJV(.763**),and ADVV (.630**), demonstrates that adjectives and adverbs are frequently used as modifiers by the students. The adjectives with more than 5 concordance hits are:

beautiful,different, few,first, future, good, great, happy, high, important, interesting, many, meaningful, military, my, new, other, past, same, small,spare,wonderful, and the like. With these words , a vivid picture of a new campus life for a fresh college student is spread before the readers: farewelling to the past middle school life, through the great happy/interesting/meaningful military training, the freshman begins his/her wonderful campus life. And the adverbs with more than 5 concordance hits are: also, finally, forward, hard, here, just, not, n’t, only, so, still, very, and so forth. It can be seen most of the adverbs are common, or even colloquial words.

Table 2 Correlation between Lexical Complexity Measures for the 2nd Term

Term 2	Score	LD	LS1	LS2	NDW	NDWERZ	TTR	CTTR	VV1	LV	VV2	NV	ADJV	ADVV	MODV
Score	1	0.106	0.209	0.065	0.149	-0.137	-0.111	0.096	0.012	-0.054	-0.23	-0.029	0.102	-0.176	-0.053
LD	0.106	1	0.09	0.127	0.236	0.123	-0.126	0.241	-0.104	-.340*	-0.209	-.467**	0.023	0.073	0.106
LS1	0.209	0.09	1	.882**	.346*	-0.021	0.006	.419**	.314*	.317*	0.042	.309*	.374*	-.349*	0.012
LS2	0.065	0.127	.882**	1	.310*	-0.091	-0.078	.341*	0.22	0.187	-0.039	0.161	0.292	-.373*	-0.069
NDW	0.149	0.236	.346*	.310*	1	.437**	-.455**	.900**	-0.185	-0.195	-0.069	-0.21	0.05	0.049	0.051
NDWERZ	-0.137	0.123	-0.021	-0.091	.437**	1	.328*	.679**	-0.163	.324*	.345*	0.263	-0.044	.443**	0.297
TTR	-0.111	-0.126	0.006	-0.078	-.455**	.328*	1	-0.069	.472**	.854**	.500**	.722**	0.269	.349*	.473**
CTTR	0.096	0.241	.419**	.341*	.900**	.679**	-0.069	1	-0.055	0.149	0.151	0.077	0.176	0.188	0.253
VV1	0.012	-0.104	.314*	0.22	-0.185	-0.163	.472**	-0.055	1	.602**	0.219	.385*	0.219	0.084	0.235
LV	-0.054	-.340*	.317*	0.187	-0.195	.324*	.854**	0.149	.602**	1	.528**	.884**	.373*	0.235	.451**
VV2	-0.23	-0.209	0.042	-0.039	-0.069	.345*	.500**	0.151	0.219	.528**	1	.363*	-0.106	0.225	0.104
NV	-0.029	-.467**	.309*	0.161	-0.21	0.263	.722**	0.077	.385*	.884**	.363*	1	0.226	0.075	0.215
ADJV	0.102	0.023	.374*	0.292	0.05	-0.044	0.269	0.176	0.219	.373*	-0.106	0.226	1	-0.049	.691**
ADVV	-0.176	0.073	-.349*	-.373*	0.049	.443**	.349*	0.188	0.084	0.235	0.225	0.075	-0.049	1	.677**
MODV	-0.053	0.106	0.012	-0.069	0.051	0.297	.473**	0.253	0.235	.451**	0.104	0.215	.691**	.677**	1

*** P < 0.05; ** p < 0.01**

In Table 2, in the 2nd term, first, there seems no significant correlation between Score and other lexical density meaures. Second, LD is negatively correlated with LV(-.340*),and NV(-.467**). This is a surprising phenomenon that means the more noun words the student use, the lower lexical density his writing contains. Third, LS1 is positively correlated with VV1(.309*), and NV(.374*), but negatively with ADVV(-.349*), i.e. those writings with higher lexical sophistication are full of verbs and nouns in varied forms, but lack of adverbs in varied forms. In addition, the adverb n’t , the colloquial form of adverb not, has 46 concordance hits, with a frequency rate of 0.18%.

Table 3 Descriptive Statiscs of the Lexical Complexity Measures for the 1st & the 2nd Term

	Min1	Min2	Max1	Max2	Mean	Mean	Std 1	Std 2
Score	50	43.5	84	90.5	68.98	74.04	9.10	9.41
LD	0.45	0.45	0.63	0.59	0.54	0.52	0.04	0.04
LS1	0.12	0.04	0.4	0.46	0.27	0.20	0.07	0.08
LS2	0.12	0.05	0.34	0.44	0.22	0.18	0.05	0.07
NDW	55	41	119	184	83.41	95.34	13.19	27.67
NDWERZ	34.4	33.2	41.9	41.7	38.60	37.97	1.43	1.96
TTR	0.51	0.43	0.73	0.71	0.62	0.56	0.06	0.06
CTTR	4.29	3.51	5.74	6.51	5.05	5.09	0.36	0.65
VV1	0.61	0.52	1	1	0.84	0.79	0.09	0.11

LV	0.62	0.57	0.94	0.92	0.81	0.72	0.07	0.08
VV2	0.12	0.1	0.29	0.3	0.20	0.17	0.04	0.04
NV	0.6	0.39	0.96	0.91	0.78	0.65	0.09	0.10
ADJV	0.09	0.08	0.26	0.22	0.16	0.14	0.04	0.03
ADVV	0.02	0.04	0.15	0.17	0.08	0.10	0.03	0.03
MODV	0.14	0.14	0.35	0.32	0.24	0.24	0.05	0.05

Apparently, there is progress in the writing during the 2nd term compared with that of the 1st term. First, the mean score of each piece of writing is raised from 68.98 to 74.04, which is a significant rise. Second, the number of different words in the writings of the 2nd term is greater than that of the counterpart of the 1st term. This is, with time passing and through more writing practice, the students writing proficiency is getting higher and higher. In view of the slightly greater standard deviation from 9.10 to 9.41, this shows it is common in China the students in vocational colleges have different levels of English proficiency on their entrance to college. And those at the lowest level are always a tough nut for their teachers.

Table 4 Correlation Between Lexical Complexity Measures for the 1st & the 2nd Term

Term	.267*	-0.163	-.383**	-.360**	.268*	-0.184	-.431**	0.035
	Score	LD	LS1	LS2	NDW	NDWERZ	TTR	CTTR
Term	0.035	-.246*	-.508**	-.385**	-.563**	-0.203	.240*	0.019
	CTTR	VV1	LV	VV2	NV	ADJV	ADVV	MODV

*** P < 0.05; ** p < 0.01**

Table 4 summarizes the changes of the lexical complexity measures in the 1st and the 2nd term. Most importantly, as just mentioned above, score is positively correlated with Term, i.e. the students' proficiency is getting improved in their writing with time passing. Secondly, Term is also positively correlated with NDW(.268*), and ADVV(.240*), which means, in the 2nd term their writings are becoming longer in length and with more adverbs, althogh those adverbs are in less varied forms. Thirdly, Term is negatively correlated with NV(-.563**), LV(-.508**), TTR (-.431**), LS1(-.383**), LS2 (-.360**), and VV1(-.246*). This suggests, the lexical density of the students' writings in the 2nd term is lower than that of the counterpart in the 1st term, consisting in the lowerness of LS, VV1, LV , and NV. That is, most of the students prefer to nouns and verbs in less varied forms in their writings in the 2nd term. This confirms Jin's finding that genre has a major effect on lexical complexity[11], for the narrative genre is chosen in the 1st term, while the argumentative in the 2nd term.

3. Pedagogical Implications

To summarize the results of the analyses from above, here come the following points:

First, the students' proficiency is getting improved in their writing with time passing and through more writing pratice. Second, the writings by vocational college students are lack of nouns and verbs in varied forms. Third, their writings are sometimes full of adverbs in colloquial form.

Corresponding to findings above, some pedagogical suggestions are put forward as follows:

First, the vocational college students should be kept practising writing more often, for the writing course is a practical course needing frequent exercise.

Second, the students are encouraged to read something more on grammar and lexicology, to obtain more knowledges on varied forms of content words, especially noun words and verb words, and to memorize more synonyms of nouns and verbs.

Third, the students are required to learn something more on the style of words, especially of adverbs.

Finally, in light of the uneven levels of English language proficiency of the vocational college students, graded teaching methods may be applied in the writing course, and special attentions and patience should be given to those students at lower levels in order that they are able to do their best in writing.

References

[1] P. J.L.Arnaud (1984).The lexical richness of L2 written production and the validity of vocabulary tests[A]. In T.Culhane, C.Klein-Braley&D.K.Stevenson(eds.). Practice and Problems in Language Testing[C]. Colchester: University of Essex, p.14- 28.

[2] Ai, Haiyang and Lu, Xiaofei (2010). A web-based system for automatic measurement of lexical complexity. Paper presented at the 27th Annual Symposium of the Computer-Assisted Language Consortium (CALICO-10). Amherst, MA. June 8-12.

[3] Lu, Xiaofei (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. The Modern Language Journal, 96(2):190-208.

[4] Douglas Biber , Stig Johansson , Geoffrey Leech , Susan Conrad , Edward Finegan (1999). Johansson S, Leech G, Conrad S, Finegan, E. Student Grammar of Spoken and Written English. London:Longman.

[5] J. Read (2000). Assessing Vocabulary. Cambridge University Press.

[6] Mark D. Johnson (2017). Cognitive task complexity and L2 written syntactic complexity, accuracy, lexical complexity, and fluency: A research synthesis and meta-analysis. Journal of Second Language Writing, Vol.37, pp.13-38.

[7] Mark Wain Frear, John Bitchener (2015). and Bitchener, J. The effects of cognitive task complexity on writing complexity.Journal of Second Language Writing, Volume 30, 2015, pp. 45-57.

[8] Cornelia Lahmann , Rasmus Steinkrauss , Monika S. Schmid (2016). Factors Affecting Grammatical and Lexical Complexity of Long‐Term L2 Speakers’ Oral Proficiency. Language Learning, Vol.66 (2), pp.354-385.

[9] Abdi Tabari (2016). The effects of planning time on complexity, accuracy, fluency, and lexical variety in L2 descriptive writing. Asian-Pacific Journal of Second and Foreign Language Education 1:10.

[10] P. Zhang (2007). A Corpus-based Contrastive Analysis on Lexical Complexity of Chinese and International EFL Learners. Foreign Languages in China, (03), pp.54-59.

[11] X. W. Jin (2010). A Study on the Factors Affecting the Lexical Complexity of English Majors'Compositions. Journal of Nanjing Institute of Technology(Social Science Edition), (02), pp.30-34.

[12] G. Bao (2011). The Effects of School Type and Writing Proficiency on EFL Learners' Lexical Complexity Across Course Program Levels. Journal of PLA University of Foreign Languages, 04:55-60.

[13] Ronggen, Zhang (2015). A Coh-metrix Study of Writings by Majors of Mechanic Engineering in the Vocational College. Theory and Practice in Language Studies, vol.5, no.5, p. 1929-1934.