ISSN 2522-6398 Vol. 4, Issue 1: 20-25, DOI: 10.25236/FER.2021.040105

On the Differentation of Texts by L1 and L2 Learners

Ronggen Zhang

Shanghai Publishing and Printing College , Shanghai, 200093

ABSTRACT. *Based on data of 90 pieces of writings from two corpora WECCL and LOCNESS, the paper makes a study of the differentiation of texts by L1 and L2 learners in respect of lexical complexity. The tools used to process the data include AntConc, SPSS 19, the Web-based Lexical complexity analyzer - Single Mode etc. The findings of the research are as follows: first, there are significant underuses of modifiers and adverbs in L2 learners' writings; second, there are more sophisticated words such as verbs, nouns, adjectives, adverbs, and modifiers in the L1 learners' writings than in the L2 learners' writings. Finally, some pedagogical suggestions are raised to enhance the L2 writing instruction. For instance, first, the L2 instructors should arouse the learners' attention to the differences between the learners' mother language and L2, to eradicate with efforts the negative transfer of the mother language of the L2 learner; second, the instructors should implant in the students more knowledge on English stylistics, by making clear difference between oral style and written style; third, the instructors should compile more updated English text books with authentic English material from online English corpora.*

KEYWORDS: *Corpus, Data Mining, WECCL, LOCNESS, Lexical complexity*

1. Introduction

This study is based on corpus and all the data are processed via computer. The data include 90 argumentative writings by Chinese undergraduate English majors and native undergraduates based on the two corpora, namely the Louvain Corpus of Native English (LOCNESS) [1] and Written English Corpus of Chinese Learners (WECCL) [2]. The 45 texts from LOCNESS covers the topic on some social issues such money as the root of all evil, crime, male or femaleis social contribution, feminism, death penalty, legalization of marijuana, teachers role, euthanasia, rules and regulations, etc. The other 45 texts from WECCL include essays on some social issues such as consequences of a failure to expensive education, plastic pollution, and college students living outside campus. The study attempts to find the differentation of texts by L1 and L2 learners in respective of lecical complexity. Lexical density is the proportion of the text made up of lexical word tokens, including nouns, lexical verbs, adjectives, and adverbs [3].

Wilkens makes a study of the effectiveness of the cues like the length of the word or its frequency by using them in a classification task for separating words as simple or complex, and finds that word length is not important, while corpus frequency is enough to correctly classify a large proportion of the test cases [4]. Frear and Bitchener's study examines the relationship between increases in cognitive task complexity and the writing of intermediate L2 writers of English [5]. Johnson finds that features of task complexity may promote attention to the formulation and monitoring systems of the writing process [6]. Lexical simplification is the task of replacing a complex term with its simpler alternatives and finds that the Conditional Random Field based approach for complex word identification gives better accuracy and the substitution generation method out performs previous approaches [7].

Mahajan and Zaveri proposed a machine learning system uses lexical features and dependency based features for sentence level paraphrase identification, and finds results with dependency features are highly sensitive to minor syntactic change [8]. The accuracy of 76% could be improved by adding a step in the phrasal compounds compounder module which specified user-defined contexts being sensitive to the part of speech of the non-head parts and by using TreeTagger, in line with our approach [9]. Pietro et al investigates the interplay between lexical complexity and syntactic complexity with respect to nominal lexicon and how it is affected by textual genre and level of linguistic complexity within genre[10].

Yang et al finds that there are significant differences between Chinese students and native speakers in the choice of position adverbials. In most cases, Chinese students prefer to put it at the beginning of

ISSN 2522-6398 Vol. 4, Issue 1: 20-25, DOI: 10.25236/FER.2021.040105

the sentence, while native speakers will use it flexibly at the beginning, before and after the verb [11]. Chinese undergraduates overuse some lexical bundles and most frequently used bundles by Chinese undergraduates seldom appear in the native undergraduates' writings [12]. There are significant differences between Chinese English learners and native speakers in the use frequency of time series conjunctions: the use frequency of time series conjunctions of Chinese English learners is higher than that of native speakers [13].

The above literature review shows there are few researches on the comparative study of the texts by L1 and L2 learners in respective of lecical complexity. Hence this paper tries to fill in this gap by comparing the texts from two corpora LOCNESS and WECCL.

Research Questions

What are the significant differences in lexical complexity of the texts by L1 and L2 learners?

What are the pedagogical implications of the reseach findings?

Objective of this Research

The purpose of the study is, by date mining of the corpora and analyzing the data, to find the differentiation of the texts by L1 and L2 learners and give some suggestions for L2 learners. .

Significance of this Research

The significance of this research lies in that it provides both the instructors and the students with a systematic computer-assisted way of analyzing texts in terms of lecical complexity, to enhance the teaching and learning of English writing courses.

2. Methodology

1) Sampling

The corpora concerned are based on the 90 pieces of argumentative writings by Chinese undergraduate English majors and native undergraduates based on the two corpora, namely the Louvain Corpus of Native English (LOCNESS) [1] and Written English Corpus of Chinese Learners (WECCL) (Wen, Liang, & Yan, 2008).

2) Data mining: all the data are processed by using the software such as AntConc, SPSS 19, etc.

3. Data Processing and Analysis of Lexical Complexity

All the data are processed oneline by the Web-based Lexical complexity analyzer - Single Mode [14].The special terms concerned are as follows:

LD stands for Lexical density, LS1 short for Lexical sophistication-I, LS2 for Lexical sophistication-II, NDW for Number of different words, NDWERZ for Number of different words expected random, TTR for Type/Token ratio, CTTR for Corrected TTR, VV1for Verb variation-1, LV for Lexical word variation, VV2 for Verb variation-II, NV for Noun variation, ADJV for Adjective variation, ADVV for Adverb variation, and MODV for Modifier variation.

Some of the above terms for measuring lexical complexity are further defined and explained as follows:

Lexical sophistication-I (LS1), stands for the ratio of the number of tokens of sophisticated lexical words to the number of tokens of lexical words. And Lexical sophistication-II (LS2), represents 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. And 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. And VV2 stands for the ratio of the number of types of verbs to the number of tokens of lexical words. Finally, NV refers to the ratio of the number of types of nouns to the number of tokens of nouns. And so on, comes the definition of ADJV, ADVV), and MODV.

Table 1 Descriptive Statistics of the Lexical Complexity Measures for WECCL and LOCNESS

ISSN 2522-6398 Vol. 4, Issue 1: 20-25, DOI: 10.25236/FER.2021.040105

	Min-wec	Min-loc	Max-wec	Max-loc	Mean-wec	Mean-loc	Std-wec	Std-loc
LD	0.44	0.42	0.6	0.57	0.51	0.5	0.03	0.03
LS1	0.06	0.1	0.31	0.45	0.2	0.25	0.07	0.09
LS2	0.08	0.12	0.3	0.41	0.19	0.25	0.05	0.06
NDW	83	125	239	479	138.93	263.29	36.67	82.45
NDWERZ	34.3	34.4	42	42.5	38.96	39.15	1.66	1.74
TTR	0.39	0.22	0.61	0.54	0.5	0.38	0.06	0.08
CTTR	4.7	4.9	7.41	10.01	5.82	6.9	0.62	0.97
VV1	0.54	0.36	1	0.88	0.74	0.63	0.1	0.14
LV	0.49	0.35	0.91	0.74	0.68	0.57	0.08	0.11
VV2	0.11	0.09	0.25	0.2	0.18	0.15	0.03	0.03
NV	0.41	0.29	0.85	0.73	0.64	0.54	0.11	0.12
ADJV	0.09	0.06	0.22	0.18	0.14	0.11	0.03	0.03
ADVV	0.06	0.04	0.16	0.12	0.11	0.07	0.03	0.02
MODV	0.19	0.1	0.32	0.28	0.25	0.18	0.04	0.04

The descriptive statistics data in table 1 shows, that the values of the most of the 17 lexical complexity measures for WECCL and LOCNESS are quite different from each other except 4 of them; that is, the values of LD, LS1, NDWERZ, and VV2 of the two corpora are quite similar to each other. And the 5 most significantly different values are of NDW, TTR, MODV, ADVV, and LV, which means that the L1 learners can use more different words, especially different modifiers and adverbs, in their writings than their L2 counterparts. The concordance hit results by AntConc3.2.2 are as follows: Interestingly, the words all and such are the 2 most frequently used modifiers in the both corpora. Nevertheless, the 5 most frequently used adverbs by the L2 learners are most (frequency 11), so (7), faster (4), even (4), further(2), while the 5 most frequently used adverbs by the L1 learners are more(146), most(31), less (14), even(8), and there(7).

Table 2 Correlation between Lexical Complexity Measures for WECCL

	LD	LS1	LS2	NDW	NDWERZ	TTR	CTTR	VV1	LV	VV2	NV	ADJV	ADVV	MODV
LD	1	-0.107	-0.171	-0.293	0.078	.299*	-0.168	0	-0.067	-0.077	-0.055	0.006	0.005	-0.007
LS1	-0.107	1	.914**	0.249	.440**	0.196	.367*	0.166	.397**	.368*	.398**	0.111	0.063	0.138
LS2	-0.171	.914**	1	.458**	.411**	0.073	.559**	0.105	.354*	.346*	.333*	0.092	0.053	0.123
NDW	-0.293	0.249	.458**	1	0.275	-.541**	.889**	-.304*	-0.14	-0.028	-0.061	-0.235	-0.155	-0.266
NDWERZ	0.078	.440**	.411**	0.275	1	.301*	.492**	0.068	.395**	0.251	.498**	0.254	0.105	0.25
TTR	.299*	0.196	0.073	-.541**	.301*	1	-0.113	.647**	.812**	.560**	.658**	.559**	0.24	.574**
CTTR	-0.168	.367*	.559**	.889**	.492**	-0.113	1	-0.015	0.255	0.238	0.266	0.03	-0.041	0.004
VV1	0	0.166	0.105	-.304*	0.068	.647**	-0.015	1	.625**	.370*	0.236	.627**	0.147	.591**
LV	-0.067	.397**	.354*	-0.14	.395**	.812**	0.255	.625**	1	.588**	.848**	.521**	.344*	.631**
VV2	-0.077	.368*	.346*	-0.028	0.251	.560**	0.238	.370*	.588**	1	.496**	-0.005	0.092	0.06
NV	-0.055	.398**	.333*	-0.061	.498**	.658**	0.266	0.236	.848**	.496**	1	.316*	0.292	.422**
ADJV	0.006	0.111	0.092	-0.235	0.254	.559**	0.03	.627**	.521**	-0.005	.316*	1	-0.097	.701**
ADVV	0.005	0.063	0.053	-0.155	0.105	0.24	-0.041	0.147	.344*	0.092	0.292	-0.097	1	.627**
MODV	-0.007	0.138	0.123	-0.266	0.25	.574**	0.004	.591**	.631**	0.06	.422**	.701**	.627**	1

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

ISSN 2522-6398 Vol. 4, Issue 1: 20-25, DOI: 10.25236/FER.2021.040105

According to table 2, the correlations between the lexical complexity measures for WECC are as follows: First, LD is only positively correlated with TTR; i.e. the higher Type/Token ratio is, the higher the Lxical density is. Second, LS1 is positively correlated with LS2, NDWERZ, CTTR, LV, VV2, and NV; i.e. the sophisticated lexical words mainly consist of different types of words, especially of verbs and nouns. Third, LS2 is positively correlated with LS1, NDW, NDWERZ, CTTR, LV , and VV2; i.e. the greater number of types of sophisticated words means the greater number of tokens of sophisticated lexical words, the more different words, especially more verbs and nouns. Fourth, TTR is positively correlated with LD, NDW, NDWERZ, VV1, LV, VV2, NV, ADJV, and MODV; i.e. the greater number of words types means the higher lexical density, which persists in larger numbers of different words, especially of verbs, nouns, adjectives and modifiers. Fifth, VV2 is positively correlated with LS1, LS2, TTR, VV1, LV, NV; i.e. the lexical variation mainly persists in the variations of verbs and nouns. Sixth, NV is positively correlated with LS1, LS2, NDWERZ, CTTR, LV, VV2, ADJV, MODV; i.e. the greater number of nouns comes with the greater number of verbs, adjectives and modifiers which contribute to the higher lexical word variation. And MODV is positively correlated with TTR, VV1, LV, NV, ADJV, and ADVV; i.e. the greater number of modifiers contributes to the greater number of words types, along with more verbs, nouns, adjectives and adverbs. To conclude the data in table 2, the L2 learners' writings of high lexical density tend to be filled with different verbs, nouns, adjectives, adverbs, and modifiers.

Table 3 Correlation between Lexical Complexity Measures for LOCNESS

	LD	LS1	LS2	NDW	NDWERZ	TTR	CTTR	VV1	LV	VV2	NV	ADJV	ADVV	MODV
LD	1	.412**	.401**	0.085	.468**	.335*	.374*	0.2	0.101	-0.253	0.101	0.13	-0.105	0.058
LS1	.412**	1	.893**	0.22	.377*	.392**	.581**	.452**	.413**	-0.052	.352*	.467**	0.079	.389**
LS2	.401**	.893**	1	.409**	.530**	.326*	.729**	.310*	.379*	-0.154	.384**	.447**	0.083	.374*
NDW	0.085	0.22	.409**	1	.381**	-.514**	.710**	-.486**	-.422**	-.622**	-.299*	-0.262	-.347*	-.344*
NDWERZ	.468**	.377*	.530**	.381**	1	.347*	.747**	0.102	.309*	-0.131	.413**	0.24	0.277	.313*
TTR	.335*	.392**	.326*	-.514**	.347*	1	0.223	.838**	.942**	.668**	.875**	.675**	.468**	.712**
CTTR	.374*	.581**	.729**	.710**	.747**	0.223	1	0.112	0.288	-0.163	.382**	0.255	0.021	0.206
VV1	0.2	.452**	.310*	-.486**	0.102	.838**	0.112	1	.848**	.619**	.665**	.740**	0.245	.663**
LV	0.101	.413**	.379*	-.422**	.309*	.942**	0.288	.848**	1	.739**	.938**	.742**	.494**	.768**
VV2	-0.253	-0.052	-0.154	-.622**	-0.131	.668**	-0.163	.619**	.739**	1	.679**	.429**	.327*	.422**
NV	0.101	.352*	.384**	-.299*	.413**	.875**	.382**	.665**	.938**	.679**	1	.618**	.496**	.668**
ADJV	0.13	.467**	.447**	-0.262	0.24	.675**	0.255	.740**	.742**	.429**	.618**	1	.386**	.891**
ADVV	-0.105	0.079	0.083	-.347*	0.277	.468**	0.021	0.245	.494**	.327*	.496**	.386**	1	.744**
MODV	0.058	.389**	.374*	-.344*	.313*	.712**	0.206	.663**	.768**	.422**	.668**	.891**	.744**	1

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

According to table 3, the correlations between the lexical complexity measures for LOCNESS are as follows: First, LD is positively correlated with LS1, LS2, NDWERZ, TTR, CTTR, ; i.e. the higher the Lexical density lies in the greater number of tokens or types of sophisticated lexical words. Second, LS1 is positively correlated with LD, LS2, NDWERZ, TTR, CTTR, VV1, LV, NV, ADJV, MODV; i.e. the sophisticated lexical words mainly consist of different types of words, especially of verbs, nouns, adjectives and modifiers. Third, LS2 is positively correlated with LD, LS1, NDW, NDWERZ, CTTR, LV , NV, ADJV,and MODV; i.e. the greater number of types of sophisticated words means the greater number of tokens of sophisticated lexical words, the more different words, especially more verbs, nouns, adjectives and modifiers. Fourth, TTR is positively correlated with LD, LS1, LS2, NDWERZ, VV1, LV, VV2, NV, ADJV, ADVV and MODV; but negatively correlated with NDW, i.e. the greater number of words types means the higher lexical density, which persists in larger numbers of sophisticated lexical words, especially of verbs, nouns, adjectives, adverbs, and modifiers. Fifth, VV2 is positively correlated with TTR, VV1, LV, NV, ADJV, ADVV and MODV; but negatively correlated with NDW, i.e. the lexical variation persists in the variations of verbs, nouns, adjectives, adverbs, and modifiers. Sixth, interestingly, either NV or MODV is positively correlated with all the lexical density measures except LD; but negatively correlated with NDW, i.e. the greater number of nouns or modifiers comes with the greater number of verbs, adjectives or adverbs which contribute to the higher

ISSN 2522-6398 Vol. 4, Issue 1: 20-25, DOI: 10.25236/FER.2021.040105

lexical word variation. To conclude the data in table 3, the L1 learners' writings of high lexical density tend to be filled with sophisticated verbs, nouns, adjectives, adverbs, and modifiers.

Table 4 T-test, Correlation Coefficients of Paired Samples

		N	CC	Sig.
pair 1	CORP & LD	90	-.070	.514
pair 2	CORP & LS1	90	.289	.006
pair 3	CORP & LS2	90	.453	.000
pair 4	CORP & NDW	90	.702	.000
pair 5	CORP & NDWERZ	90	.059	.582
Pair 6	CORP & TTR	90	-.651	.000
pair 7	CORP & CTTR	90	.557	.000
pair 8	CORP & VV1	90	-.406	.000
pair 9	CORP & LV	90	-.514	.000
pair 10	CORP & VV2	90	-.484	.000
pair 11	CORP & NV	90	-.433	.000
pair 12	CORP & ADJV	90	-.483	.000
pair 13	CORP & ADVV	90	-.564	.000
pair 14	CORP & MODV	90	-.636	.000

Table 4 shows the correlations between either of the two corpora and each of its lexical density measures. Obviously, only 3 measures are not correlated with the type of corpora, i.e. LD, LS1, and NDWERZ, which has been confirmed in table 1, where the values of LD, LS1, and NDWERZ are correspondingly similar in both corpora. Each pair of them in WECCL and LOCNESS is respectively as follows: LD (0.51- 0.5);LS1 (0.2- 0.25), and NDWERZ (38.96- 39.15). Among the rest 11 lexical density measures, 3 of them (LS2, NDW, and CTTR) are positively correlated with LOCNESS, while the other 8 are negatively correlated with WECCL (for the convenience of statistics, 1 stands for WECCL and 2 for LOCNESS). That is, the L1 learners' writings of high lexical density tend to be filled with sophisticated, while the L2 learners' writings of high lexical density tend to be filled with different words, especially verbs, nouns, adjectives, adverbs, and modifiers. To summarize, if of the same lexical density, there are more sophisticated words such as verbs, nouns, adjectives, adverbs, and modifiers in the L1 learners' writings than in the L2 learners' writings.

4. Conlusion

From the above data analyses, at least two conclusions may be made: First, the L1 learners can use more different words, especially different modifiers and adverbs, in their writings than their L2 counterparts; i.e. there are significant underuses of modifiers and adverbs. Second, there are more sophisticated words such as verbs, nouns, adjectives, adverbs, and modifiers in the L1 learners' writings than in the L2 learners' writings.

As to the above findings, some pedagogical suggestions are to be raised to enhance the L2 writing instruction as follows:

First, the L2 instructors should arouse the learners' attention to the differences between the learners' mother language and L2, to eradicate with efforts the negative transfer of the mother language of the L2 learner. For instance, since there are fewer transitional adverbs in Chinese than in English, this may lead to the Chinese students' underuses of some adverbs in their English writings.

Second, in the light of the fact that there are fewer sophisticated words in the L2 learners' writings, the instructors should implant in the students more knowledge on English stylistics, to make clear difference between oral style and written style. Since many Chinese students tend to use more oral words in their English writings, this contributes to their underuse of sophisticated words.

Third, the instructors should compile for the students more updated English text books by utilizing all kinds of resources online and offline, especially online English corpora, to provide them with more authentic English material.

ISSN 2522-6398 Vol. 4, Issue 1: 20-25, DOI: 10.25236/FER.2021.040105

5. Limitations of the Research

The data in this study consist of 45 argumentative timed writings by Chinese undergraduate English majors in grade three, and 45 argumentative timed writings from LOCNESS by undergraduates from Indiana University at Indianapolis and University of South Carolina. And some of writings by US students are abridged within 10,000 words, which is the maximum of words to be processed by the Web-based Lexical complexity analyzer - Single Mode. Hence the limitations of the study persist in at least the following 2 points.

First, the two corpora WECCL and LOCNESS selected here are not so updated.

Second, the data from WECCL are only from English major in grade 3, who may not possibly represent all the L2 learners.

Therefore, future researches concerned are to be further carried out based on more updated corpora.

References

[1] Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In Granger, S. (ed.) Learner English on Computer. (pp. 3–18). New York: Addison Wesley Longman.

[2] Wen Q.F., Liang M. C., & Yan, X. Q. (2008). Beijing: Foreign Language Teaching and Research Press.

[3] Biber D., Johansson S., Leech G., Conrad S., & Finegan, E. (1999). Student Grammar of Spoken and Written English. London: Longman.

[4] Wilkens, R., & Dalla, V., Alessandro, B., Marcely Z., Padró, M., &Villavicencio, A. (2014). Size does not matter. Frequency does. A study of features for measuring lexical complexity. Lecture Notes in Computer Science, 8864, 129-140.

[5] Frear M. and Bitchener, J. (2015). The effects of cognitive task complexity on writing complexity. Journal of Second Language Writing, 30: 45-57.

[6] Johnson, D. (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, 37:13-38.

[7] Silpa, K & Irshad, M. (2018). Lexical Simplification of Complex Scientific Terms. 1-5. 10.1109/ICETIETR.2018.8529069.

[8] Mahajan, R. & Zaveri, M. (2016). Machine Learning based paraphrase identification system using lexical syntactic features. 2016 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2016

[9] Trips, C. (2016). Syntactic analysis of phrasal compounds in corpora: A challenge for NLP tools, Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, 1092-1097.

[10] Dell'Oglio, P., Brunato, D. & Dell'Orletta, F. (2018). Lexicon and Syntax: Complexity across Genres and Language Varieties, CEUR Workshop Proceedings, v 2253, 2018, Proceedings of the 5th Italian Conference on Computational Linguistics, CLiC-it 2018.

[11] Yang, J., Huang, P. & Wu, L. (2013).An analysis of Chinese students' improper understanding of adverbial position, A comparative study based on the corpora of CEEW and LOCNESS, Journal of Huaibei Normal University (Philosophy and Social Sciences) (in Chinese), 34 (5), 190-193.

[12] Yin, H.Y. (2017).A Contrastive Study on Four-Word Lexical Bundles in Argumentative Writing Between Chinese Undergraduates and Native Undergraduates —A Corpus-Based Approach, Xi 'an Polytechnic University (in Chinese), MA Thesis, 51-52.

[13] Zhang Rui; Lv Changhong. (2019). On the application of time series conjunctions by English learners- A study based on the corpora of TECCL and LOCNESS, Journal of Changji University(in Chinese), 5,36-39.

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