International Journal of English Language and Linguistics Research
Vol.9, No 6, pp.47-67, 2021
Print ISSN: ISSN 2053-6305
Online ISSN: ISSN2053- 6313(online)

LINGUISTIC COMPLEXITY OF FOREIGN AND CHINESE MASTER THESES ABSTRACTS

Zhang Qi
Lanzhou University, Gansu Province, 730000

ABSTRACTS: *Language complexity has been proposed as a valid and basic descriptor of L2 performance, and as an index of language development and progress. With the assistance of L2SCA, the present corpus-based contrastive study aims at analyzing the linguistic features of theses abstracts written by foreign Masters and Chinese Masters. It is found that with respect to lexical complexity, foreign Masters outperform Chinese Masters in lexical diversity but fall behind in lexical sophistication. Chinese-English learners seem to involve more academic and advanced but less diversified words in their academic writing; in terms of syntactic complexity, natives tend to product relatively longer mean length of T-unit and employ much more subordination (including clauses and dependent clauses) and verb phrases (including finite or non-finite verb phrases) in their academic writing, while nonnatives are inclined to output shorter mean length of T-unit and involve more coordination in their theses abstracts, which is consistent with what has been verified in previous studies. The findings of the study reveal some common features of academic writing in terms of vocabulary and sentence patterns, which further provides references for the teaching practice of academic thesis writing.*

KEYWORDS: lexical complexity, syntactic complexity, thesis abstract, corpus-based contrastive study

INTRODUCTION

The notion of complexity, accuracy and fluency are believed to capture the constructs of L2 performance and L2 proficiency, which is multi-componential in nature (Skehan 1998; Ellis 2003, 2008; Ellis and Barkhuizen 2005). Complexity has been proposed as a valid and basic descriptor of L2 performance, an indicator of proficiency as well as an index of language development and progress. According to Norris and Ortega (2009), complexity is a highly sophisticated construct, consisting of several sub-constructs, dimensions, levels, and components, each of which can, at least in principle, be independently evaluated. With respect to L2 complexity, it is widely accepted to stand for a progressively more elaborate language and a greater variety of syntactic patterning (Foster & Skehan, 1996). Bulté and Housen (2012) distinguish between three components of L2 complexity in a narrow sense of the term: propositional complexity,

discourse-interactional complexity, and linguistic complexity. Having received by far the most attention in L2 writing research, linguistic complexity logically comprises lexical and syntactic complexity.

Lexical complexity is a multidimensional construct form which several variables have been distinguished: lexical diversity, lexical richness, and lexical sophistication (Skehan,2009a, b). Such a classification is also supported by Johnson (2017), who believes that there are two indications of lexical complexity: a) text internal measures including lexical diversity (type-token ratio) and lexical density (content word ratio); b) text external measures, namely, lexical sophistication. Vocabulary knowledge, embedded in lexical complexity, is crucial in determining how well second language (L2) learners can express themselves. It has been considered as a significant indicator of the quality of L2 academic writing (Nation, 2013) since there is always an expectation to adhere to a set of precisely defined words that frequently occur in academia.

Syntactic complexity is manifested in second language writing in terms of how varied and sophisticated the production units or grammatical structures are (Foster& Skehan 1996, Ortega 2003, Wolfe-Quintero et al. 1998). To put it another way, it is about the range of syntactic structures produced and the degree of sophistication of such structures. Generally speaking, syntactic complexity is closely correlated with and quantified by the length of production unit, the amount of subordination and coordination, and the degree of sophistication of such particular structures as verb phrases and nominals. There have been a lot of studies on syntactic complexity in the field of language testing and language acquisition and a large number of measures have been proposed for characterizing syntactic complexity in second language writing, which has led to a flurry of research aimed at examining the extent to which these measures, along with measures of accuracy, fluency, and grammatical complexity, can be used as reliable and valid indices of the learners' developmental level or overall proficiency in L2. With the continuous expansion of the depth and breadth of research, it is more plausible to think of syntactic complexity as a continuum stretching from very low complexity (e.g., short texts produced by children or beginner English L2 learners) to very high complexity (e.g., very abstract and embedded philosophical and formal texts). Researchers has realized that syntactic complexity is such a multidimensional construct that should be explored in a comprehensive way. There has been an implicit consensus in L2 writing research that nativeness is a key factor influencing the quality of writing so that many studies compare L2 academic writing with those of native experts or native students. Therefore, among the studies focusing on the linguistic features in the academic writing of L2 learners, experts' writing has been widely used as the reference corpus. In view of what has been mentioned above,

the present study explores the linguistic complexity of theses abstracts written by foreign and Chinese Masters with the help of corpus comparative analysis. Such a comparison may help domestic academic beginners better realize the shortcomings of their own writing and thus improve the quality of their academic writing.

Measures of Linguistic Complexity

As mentioned above, L2 linguistic complexity covers lexical and syntactic complexity. Vocabulary knowledge has been considered as a significant indicator of the quality of L2 academic writing and is traditionally operationalized as lexical richness. Following Read (2000), lexical richness is conceptualized as a multidimensional feature of a learner's language use that consists of the following four interrelated components: lexical density, lexical sophistication, lexical variation, and number of errors in vocabulary use.

Lexical density reflects the proportion of content words in a text. The higher the lexical density, the more information contained in a limited space, the more condensed the language is. Lexical sophistication, also known as lexical rareness, measures the proportion of relatively unusual or advanced words in the learner's text (Read, 2000:203). Lexical variation, also labeled lexical diversity (Malvern et al., 2004) or lexical range, refers to the range of a learner's vocabulary as displayed in his or her language use, which is measured by the ratio of different word types divided by the total number of tokens in a text or standardized length of samples, i.e., Type-Token Ratio.

Researchers interested in language testing or language acquisition have put forward or employed different measures to evaluate the role lexical richness played in the quality of learners' writing or task performance. These three dimensions have been respectively proved to be positively related to writing proficiency or quality such as lexical diversity by Gebril and Plakans (2016), lexical density by Gregori-Signes and Clavel-Arroitia (2015) and lexical sophistication by Zheng (2016) and Higginbotham and Reid (2019). Lexical density is reflected by the ratio of the number of lexical (as opposed to grammatical) words, namely, nouns, adjectives, verbs, and adverbs, to the total number of words in a text. When it comes to lexical sophistication, it is much more complicated to measure in that how sophistication is defined varies among researchers and thus different lexical sophistication metrics coexist (see Table 1). Regarding lexical variation, Malvern et al. (2004) and Wolfe-Quintero et al. (1998) provided a thorough discussion of different measures of lexical variation that exist. An intuitively straightforward measure of lexical variation is the number of different words (NDW) used in a language sample, which has proved to be a potentially useful measure of child language development. However, NDW is dependent on the length of the language sample, given

the various length of the thesis abstract, it is not an optimal choice. Though Type–token ratio (TTR), that is, the ratio of the number of word types (T) to the number of words(N) in a text, is widely used in both first language (L1) and L2 acquisition studies to measure lexical diversity (Malvern, Richards, Chipere, &Durán, 2004), it has been criticized as an unsatisfactory measure of lexical diversity, given that the ratio tends to decrease as the size of the sample increases, that is, a long language sample is more likely to have a lower TTR value than a short language sample, which causes TTR to be an unreliable measure of lexical diversity. Mean segmental TTR (MSTTR) constitutes one way to improve TTR, which is computed by dividing a sample into successive segments of a given length and then calculating the average TTR of all segments. Although MSTTR reduces the sample size problem, Malvern et al. (2004) pointed out that it is not without its disadvantages; for example, samples don not always divide into standard sized segments, resulting in some waste of data. Other transformations of TTR include Corrected TTR (CTTR; Carroll, 1964), Root TTR (RTTR; Guiraud, 1960), Bilogarithmic TTR (LogTTR; Herdan, 1964), and the Uber Index (Dugast, 1979), as listed in Table 2. Although some measures have obvious problems and have been criticized more than others, there is no consensus among researchers concerning a single best measure. Results for most of the measures that have been examined in more than one study tend to be mixed, and due to the variability in research design and measure definition, it is not easy to compare the results.

Table 1. *Measures of Lexical Sophistication*

Measure	Code	Calculation
Lexical Sophistication-I	LS1	number of sophisticated lexical words/number of words
Lexical Sophistication-II	LS2	number of sophisticated word type/ total number of word types
Verb Sophistication-I	VS1	number of sophisticated verb types / total number of verbs
Corrected VS1	CVS1	number of sophisticated verb types/√2 total number of words
Verb Sophistication-II	VS2	number of sophisticated verb types ^2/ total number of verbs

Table 2. *Measures of Lexical Variation*

Measure	Code	Calculation
number of different words	NDW	number of different words/ total number of words (T)
type-token ratio	TTR	number of types/number of tokens(T/N)
mean segmental TTR (50)	MSTTR	mean TTR of all 50-word segments
corrected TTR	CTTR	T/√2N
Root TTR	RTTR	T/√N
Bilogarithmic TTR	logTTR	logT/logN
Uber Index	Uber	log^2N/log(N/T)

Syntactic complexity is about the variety and complexity of sentence patterns and the information embedded in a sentence. It has been considered an important construct in second language teaching and research in that SC measures can be used to describe L2 learners' written performance and predict their writing proficiency (Ortega, 2003). The number of (independent) clauses, dependent or subordinate clauses, and complex T-units is frequently used to quantify syntactic subordination. A clause (e.g., independent clauses, nominal, adjective and adverb clauses) is defined as a structure comprising a subject and a finite verb (Lu, 2010, 2015) and a dependent clause is specified as a finite nominal, adjective, or adverb clause (Cooper, 1976; Lu, 2010, 2014). A complex T-unit includes at least one dependent clause of any type (e.g., relative, noun, and adverbial) but does not differentiate between the types and number of dependent clauses, and therefore, is regarded as a global syntactic structure.

Researchers have made attempts to find effective SC measures to further make contribution to distinguish a good writing in view of linguistic features. Coh-Metrix, initially designed for analysis of cohesion and coherence, has also added some indicators to measure syntactic complexity and sentence pattern density. However, there still exist some limitations as for the application of Coh-Metrix to syntactic analysis due to its limited number of measures and word limit on texts to be analyzed. L2 Syntactic Complexity Analyzer, or L2SCA (Lu, 2010) was a computational system for automatic analysis of syntactic complexity using 14 different measures, which were selected from the large set of measures reviewed in Wolfe-Quintero et al. (1998) and Ortega (2003). These measures fall into 5 types (see Table 3): the first type consists of three measures that gauge length of production at the clausal, sentential, or T-unit level; the second type consists of a sentence complexity ratio; the third and fourth type comprise ratios that reflect the amount of subordination and coordination respectively; the final type consists of three ratios that consider the relationship between particular syntactic structures and larger production units. L2SCA has achieved relatively higher reliability in that among 14 indexes, 11 have been confirmed to be significantly related to second language development and writing, and the other 3 are indicators recommended for L2 writing researchers by Wolfe-Quintero et al. In addition, L2SCA can be used to process any number of writing samples using any or all of the 14 complexity measures, which improves the efficiency of data analysis, thus leading to relatively rich and reliable research results.

One of the most-studied subordination indices is the T-unit complexity ratio (C/T) which measures the number of clauses per T-unit and includes all types of clauses. It is reported in six studies reviewed by Wolfe-Quintero et al. (1998) as a strong indicator of language proficiency. This index in Nasseri (2017) is shown to distinguish between the EFL and English L1 groups' academic writing. The complex T-unit ratio (CT/T)

index of subordination is also investigated in some works (e.g., in Casanave,1994; Lu, 2010, 2011; Lu & Ai, 2015; Kyle, 2016). Casanave (1994) is among the few studies that found a positive relationship between this measure and language development; Neither Lu (2010) nor Lu (2011), on the other hand, found any significant relationship between this measure's values in argumentative essays and increased proficiency levels.

Table 3. *The Fourteen Syntactic Complexity Measures*

Measure	Code	Calculation
Type 1: Length of Production Unit		
Mean Length of Clause	MLC	number of words / number of clauses
Mean Length of Sentence	MLS	number of words / number of sentences
Mean Length of T-unit	MLT	number of words / number of T-units
Type 2: Sentence Complexity		
Sentence Complexity Ratio	C/S	number of clauses / number of sentences
Type 3: Subordination		
T-unit Complexity Ratio	C/T	number of clauses / number of T-units
Complex T-unit Ratio	CT/T	number of complex T-units / number of T-units
Dependent Clause Ratio	DC/C	number of dependent clauses / number of clauses
Dependent Clauses per T-unit	DC/T	number of dependent clauses / number of T-units
Type 4: Coordination		
Coordinate Phrases per Clause	CP/C	number of coordinate phrases / number of clauses
Coordinate Phrases per T-unit	CP/T	number of coordinate phrases / number of T-units
Sentence Coordination Ratio	T/S	number of T-units / number of sentences
Type 5: Particular Structures		
Complex Nominals per Clause	CN/C	number of complex nominals / number of clauses
Complex Nominals per T-unit	CN/T	number of complex nominals / number of T-units
Verb Phrases per T-unit	VP/T	number of verb phrases / number of T-units

The next clausal subordination index is DC/C which measures the number of dependent clauses (DC) per clause. This index is also investigated in Lu (2010, 2011), Lu and Ai (2015) and is one of the important indicators of language proficiency. Ai and Lu (2013) reported the effectiveness of this measure in capturing proficiency differences between the texts of English L1 vs. EFL groups, with the English L1 group producing larger

amounts of clausal subordination. A similar measure is the DC/T index which calculates dependent clauses per T-unit. In her review of subordination indices, Ortega (2003) also documents the studies that show the values of DC/T linearly increase with writing experience and proficiency.

The use of coordination structures is also believed to be a characteristic of syntactic complexity in early L2 development (e.g., Ortega, 2003; Wolfe-Quintero et al.,1998) in that the increase in coordination is marked as a developmental stage in L2 writing complexification (Bardovi-Harlig, 1992; Wolfe-Quintero et al., 1998). Coordination structures include coordinate phrases (CP), coordinate clauses, and sentence-level coordination. Among the global coordination indices, the two measures of coordinate phrases per clause (CP/C) and per T-unit (CP/T) are used more frequently. The CP/C index is shown in Lu (2010) to discriminate between three school proficiency levels. The findings of Lu (2011) also indicate that phrasal coordination increases with the increase in proficiency levels. Lu (2010) also found CP/C to discriminate between proficiency levels; Ai and Lu (2013) equally reported that this measure could distinguish between English L1 and EFL groups' academic writing.

Phrasal complexity and sophistication indices are among the most-commonly-used syntactic measures in studies on linguistic performance, proficiency (differences) as well as register variation studies. Prominent among phrasal sophistication and complexity measures are the CN/C and CN/T indices which calculate complex nominals per Clause and T-units. These two indices are selected and analyzed in Lu (2010, 2011) and Lu and Ai (2015). The CN/C values are reported to show significant differences between English L1 vs. EFL students in Lu and Ai (2015), and between the combined EFL learners and the English L1 group in Ai and Lu's (2013) corpus. Its conceptually related measure, CN/T, was also employed in several studies (Ai & Lu, 2013; Lu, 2010; Lu & Ai, 2015) where significant between-group and between-proficiency differences were reported. Lu & Ai (2015), for instance, found a similar pattern in CN/T values and the findings of CN/C regarding the differences between MA dissertations and published research articles.

Wolfe-Quintero et al. (1998), however, seem to favor the CN/C index when it comes to capturing group differences in second language development. The effectiveness of this selection is partly examined in Lu's (2011) work where CN/T could not capture differences between adjacent proficiency levels (marked as school years) while CN/C captured between-proficiency of all levels except one. The VP/T index (verb phrases per T-unit) is another important phrasal complexity measure that calculates the ratio of verb phrases to T-units and includes verb phrases with both finite and non-finite verbs. This index is recommended by Wolfe-Quintero et.al. (1998) and was reported to

distinguish between English L1 and EFL groups' academic writing. However, in Lu's (2011) study no relationship was found between the values of this measure and proficiency.

With the deepening of the study on syntactic complexity, researchers gradually realize that syntactic complexity is a multi-dimensional concept, and each dimension does not share a balanced and linear development, but varies in different stages of language development. Besides, other factors such as learners' background, writing task or teaching writing environment have different effects on the syntactic complexity of second language writing.

The present study aims at unlocking the particular lexical and syntactic features in theses abstracts written by foreign and Chinese Masters, the findings of which may provide teachers with a more accurate picture of lexical and syntactic progress and academic beginners with references to pointedly sharpen their academic writing skills.

Corpora Establishment and Analysis

Focusing on the lexical and syntactic features of foreign and Chinese Master theses abstracts, this paper aims to expose the linguistic features of the two by establishing Foreign Linguistics Thesis Abstract Corpus (FLC) and Chinese Linguistics Thesis Abstract Corpus (CLC), with FLC serving as a reference corpus. Each corpus consists of 50 Master theses abstracts in the field of linguistics published from 2015 to 2020, which were sampled from PQDT (ProQuest Dissertations and Theses) and CNKI (China National Knowledge Infrastructure) respectively. To ensure that all papers in the FLC corpus are written by native speakers, the name and affiliation of the first author were reexamined although this is one of the selection criteria for the corpus. In order to avoid the influence of subject attribute on linguistic features in academic writing, the two corpora involved in this study have the same distribution of sub-disciplines of linguistics and roughly equal number of papers are extracted under the same discipline (see Table 4). It is undeniable that the two corpora are never absolutely equivalent to each other. However, considering that lexical use and syntactic structure instead of writing style are of importance in this study, the two corpora are comparable.

Table 4. *The Distribution of Sub-disciplines of Linguistics across Groups*

Corpora	First/Second Language Acquisition	Discourse Analysis	Corpus-based Studies	Pragmatics	Semantics
FLC	11	12	9	14	4
CLC	12	9	8	15	6

Based on previous studies on the effectiveness of measures of different dimensions of lexical and syntactic complexity, only some of the indices that have been shown to be effective in distinguishing between different writing proficiency groups were included in this study. Therefore, in terms of lexical complexity, LD, LS2 and Uber index were selected to measure lexical density, lexical sophistication and lexical diversity respectively. Regarding syntactic complexity, MLT was chosen to stand for the length of production unit, C/T and DC/C for the amount of subordination, CP/C for the amount of coordination, and CN/C and VP/T for the degree of sophistication of particular structures.

The L2 Syntactic Complexity Analyzer (L2SCA; Lu, 2010) was used to analyze lexical and syntactic complexity measures in the present study. SPSS 22.0 was used for Independent-samples T-test to figure out whether there are significant differences in terms of lexical or syntactic metrics. Prior to conducting t-test, Q-Q plots (quantile-quantile plot) of each set of data were plotted to see whether the data are at normal distribution (as shown in Figure 1 and 2). The following graphs show these measures with a dot (e.g., CN.C instead of CN/C), which confirm homoscedasticity for all of these measures.

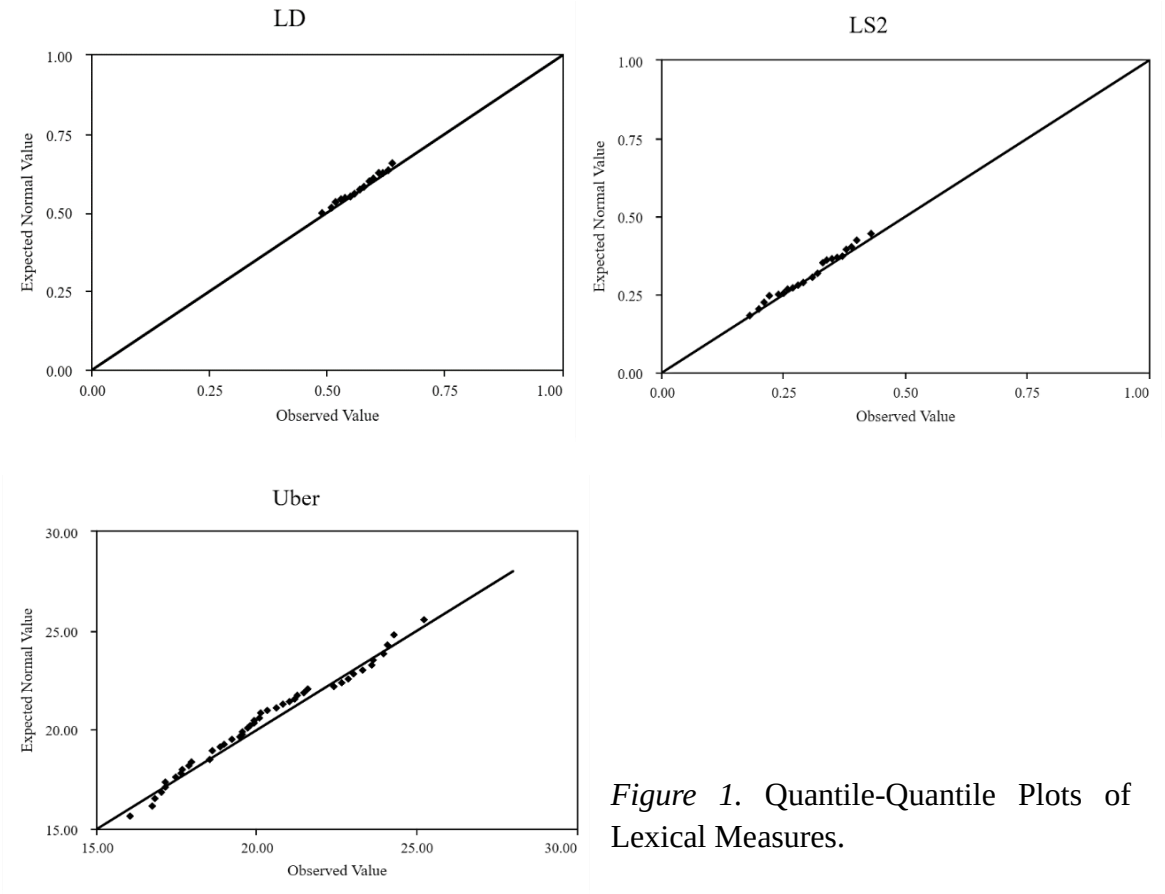

Figure 1. Quantile-Quantile Plots of Lexical Measures.

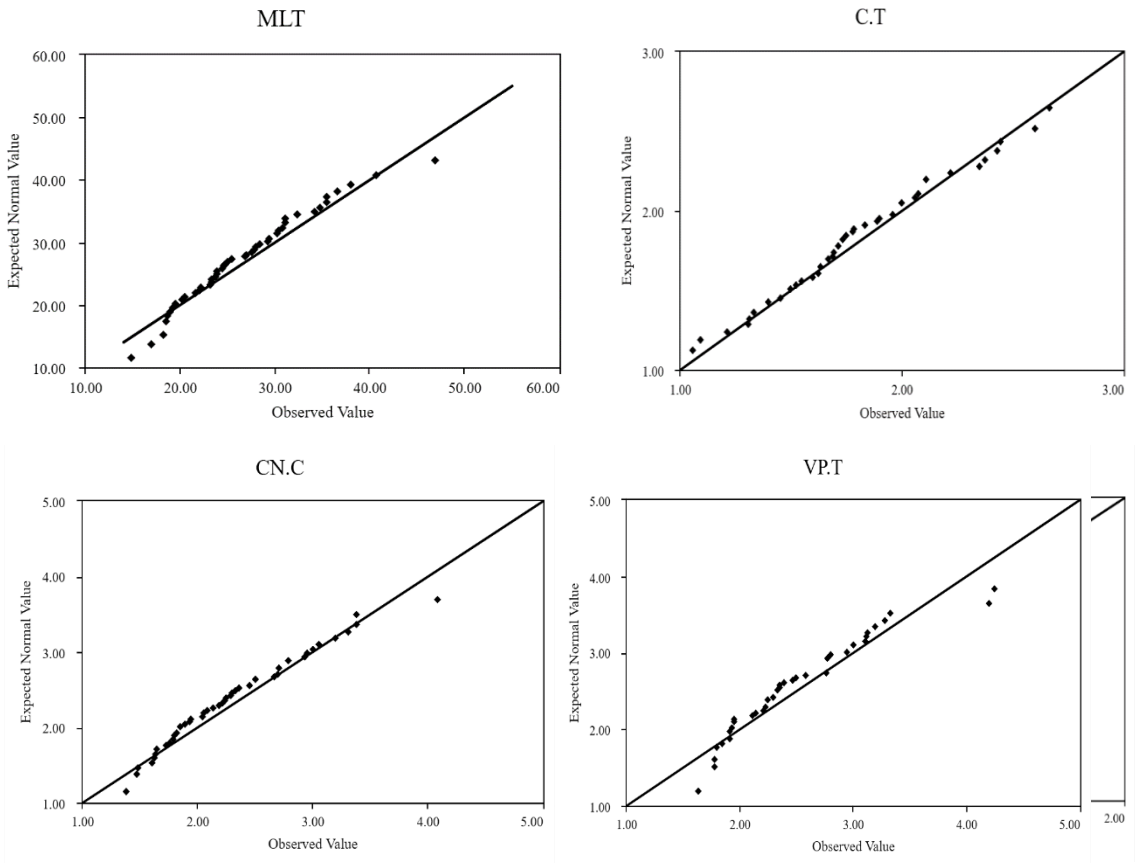

Figure 2. Quantile-Quantile Plots of Syntactic Measures.

Comparison of lexical complexity between foreign and Chinese Masters are shown in Figure 3-5. It can be seen from the above graphs that the LD curves of the abstracts of the two groups almost coincide with each other; the LS2 curve shows a complementary trend; the Uber curve clearly shows that the lexical richness of abstracts written by natives is higher than that by nonnatives.

Independent-samples T-test was further carried out to see whether there is statistical significance between two groups with regard to lexical density, sophistication and diversity. The results are shown in Table 5.

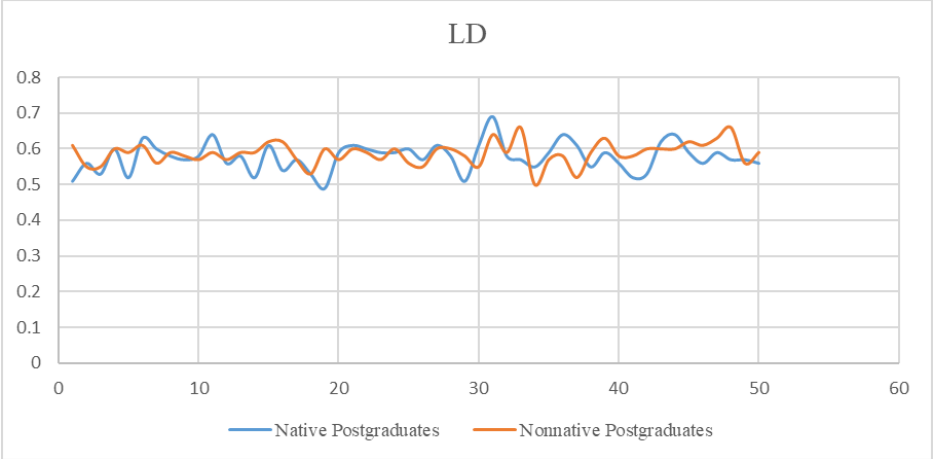

Figure 3. Comparison of LD between Natives and Nonnatives

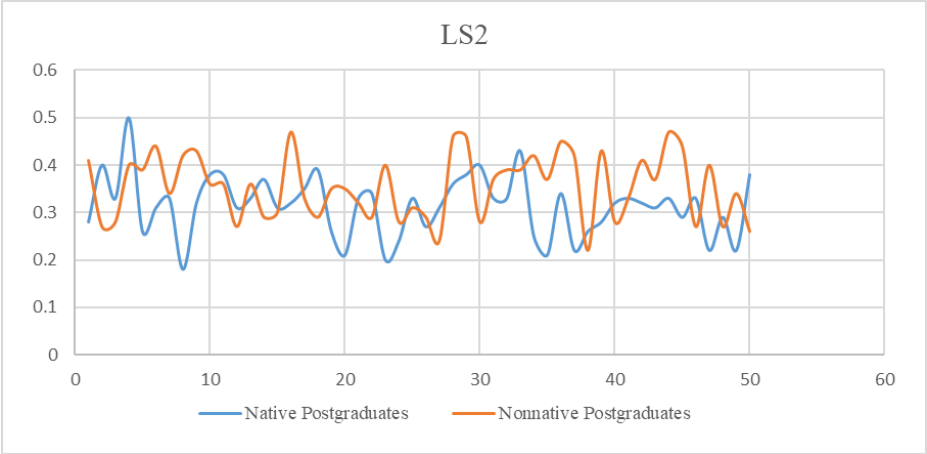

Figure 4. Comparison of LS2 between Natives and Nonnatives

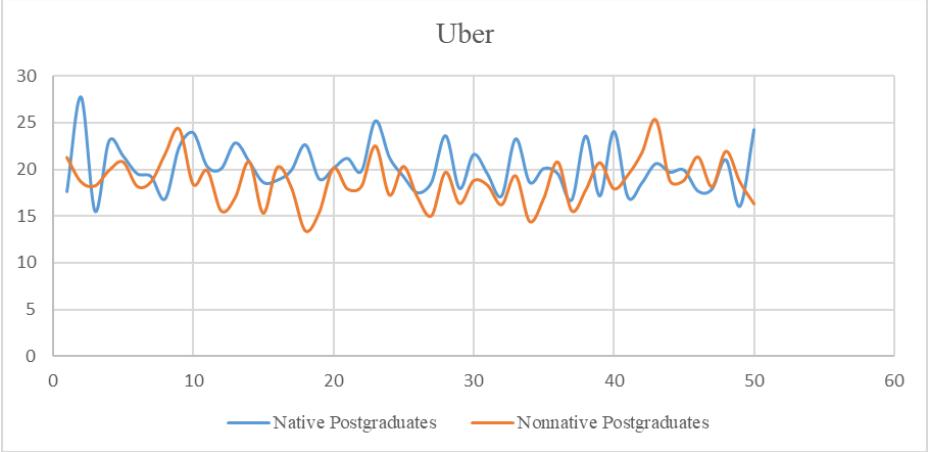

Figure 5. Comparison of Uber Index between Natives and Nonnatives

Table 5. *Comparison of Lexical Complexity between Natives and Nonnatives*

	Natives (n=50)		Nonnatives (n=50)		t	p
	M	SD	M	SD		
LD	0.58	0.04	0.59	0.03	-1.433	0.155
LS2	0.31	0.06	0.35	0.07	-3.129	0.002
Uber	20.19	2.61	18.76	2.46	2.805	0.006

As shown in Table, the two groups show similar performance in terms of lexical density with 0.58 for natives and 0.59 for Chinese-English learners, which suggest that both natives and nonnatives are aware that research papers are densely-packed information text. However, there is statistical significance in lexical sophistication (t=-3.129, p=0.002) and lexical diversity (t=2.805, p=0.006), with natives using less complicated but more diversified words compared with nonnatives. That is to say, Chinese English learners, compared with natives, tend to use more complex but less various words to write an abstract that contains roughly the same amount of information.

Comparison of syntactic complexity between native postgraduates and Chinese postgraduates are shown in Figure 6-11. As shown in these line graphs, the MLT curve of native speakers' abstracts is higher than that of Chinese-English learners. C/T, DC/C and VP/T curve are higher than those of non-native speakers. The CP/C curves of the two groups are almost coincident, and the CN/C curves are roughly complementary. That is, native speakers tend to use relatively longer T units that contain more phrasal verbs than non-native speakers and employ more subordinate structures but less coordinate constructions and complex nominals in their academic writing. T-test was conducted for further exploration of the syntactic features of the two groups of abstracts, and the results are shown in Table 6.

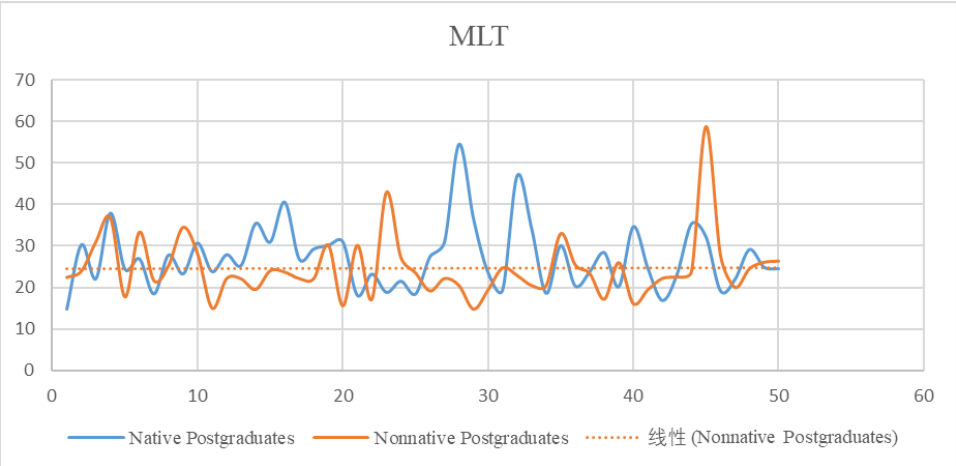

Figure.6. Comparison of MLT between Natives and Nonnatives

https://www.eajournals.org/ https://doi.org/10.37745/ijellr.13

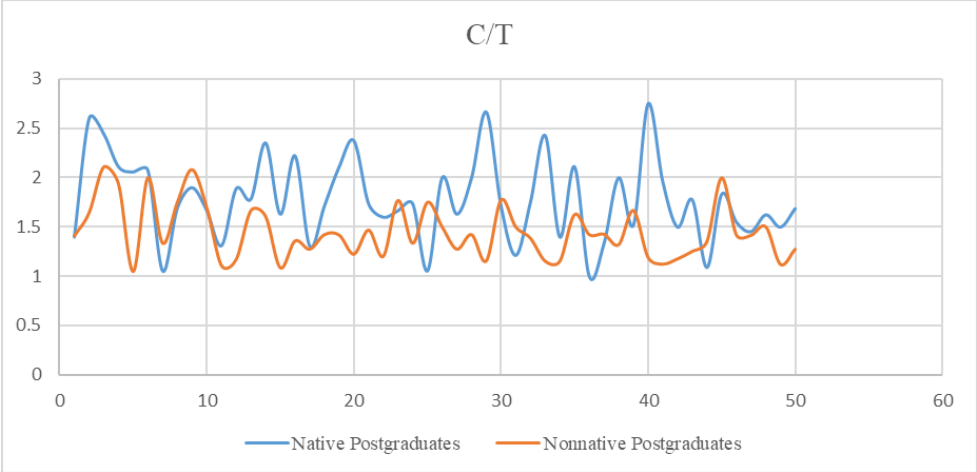

Figure 7. Comparison of C/T between Natives and Nonnatives

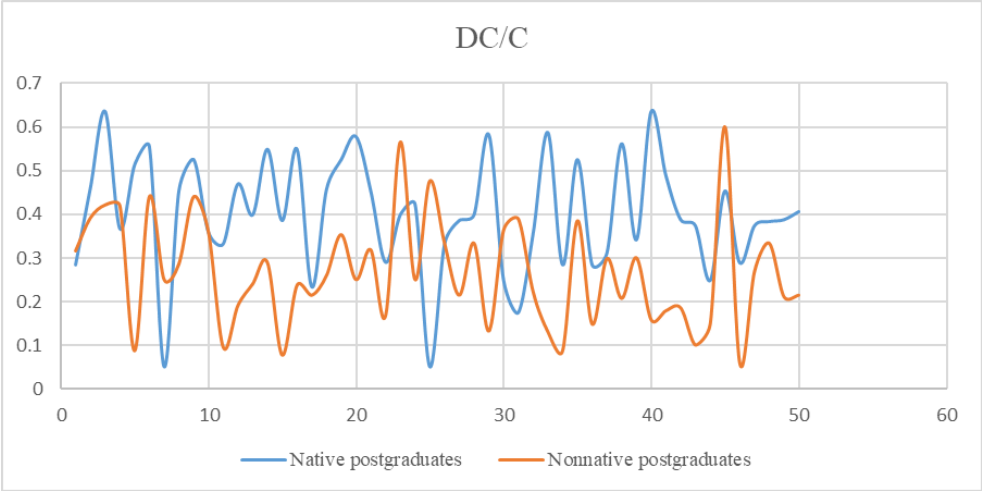

Figure 8. Comparison of DC/C between Natives and Nonnatives

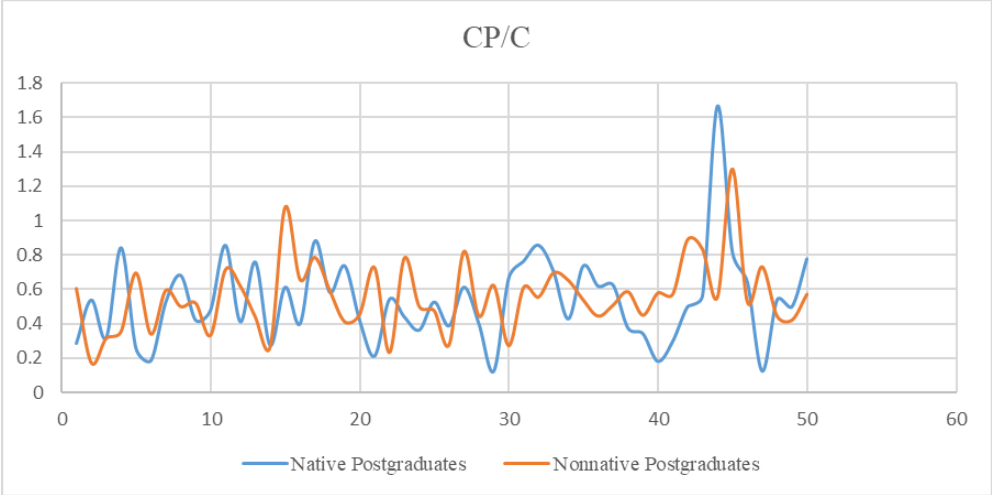

Figure 9. Comparison of CP/C between Natives and Nonnatives

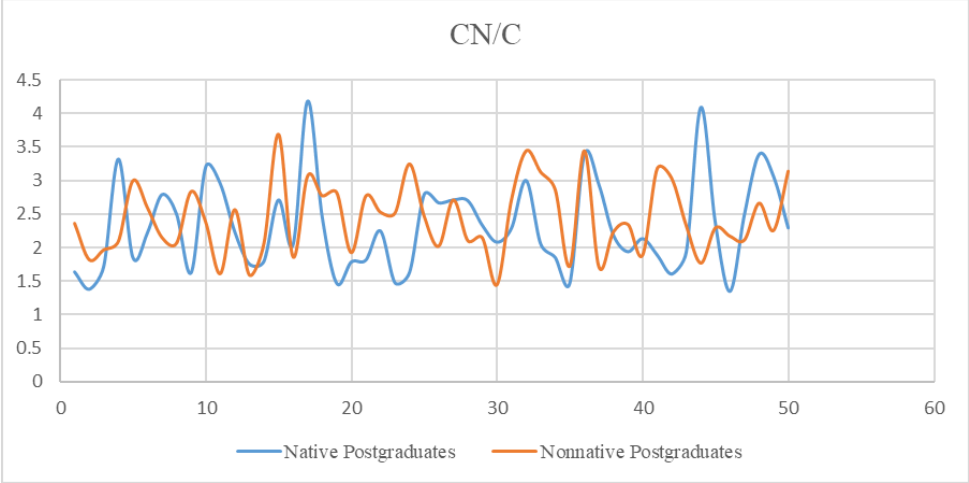

Figure 10. Comparison of CN/C between Natives and Nonnatives

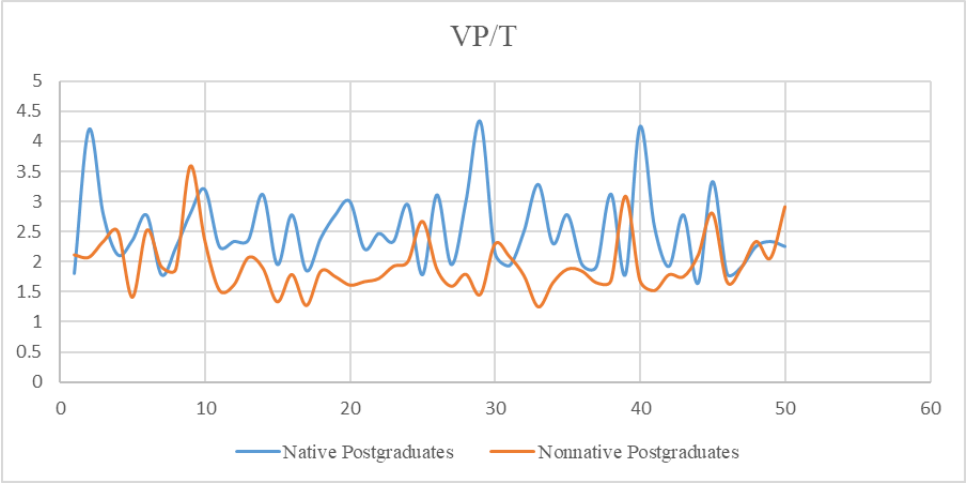

Figure 11. Comparison of VP/T between Natives and Nonnatives

Table 6. *Comparison of Syntactic Complexity between Natives and Nonnatives*

	Natives		Nonnatives		t	p
	M	SD	M	SD		
MLT	27.3	7.73	24.7	7.51	1.70	0.092
C/T	1.78	0.42	1.45	0.28	4.64	0.000
DC/C	0.40	0.13	0.27	0.13	5.27	0.000
CP/C	0.54	0.26	0.56	0.21	-0.53	0.597
CN/C	2.32	0.67	2.44	0.54	-0.99	0.325
VP/T	2.51	0.64	1.95	0.47	4.97	0.000

It can be seen from Table 6 that there is no significant difference in the mean length of T-unit between the two groups with 27.3 for natives and 24.7 for nonnatives. There are big gaps in T-unit complexity ratio (t=4.64, p=0.000) and dependent clause ratio (t=5.27,

p=0.000) as well as verbs phrase per T-unit (t=4.97, p =0.000). Regarding the number of coordinate phrases and complex nominals in clauses, the results of the two groups indicate no big differences with 0.54 and 0.56 for coordinate phrases and 2.32 and 2.44 for complex nominals. That is to say, natives use significantly more subordinate structures (including clauses and dependent clauses) and more verb phrases in the same length of T-units but less coordinate phrases and complex nominals in clauses.

RESULTS AND DISCUSSION

In the present study, two corpora (FLC and CLC) were established and L2SCA was employed to analyze the lexical and syntactic complexity of these abstracts. Lexical complexity mainly involves such three dimensions as lexical density, sophistication and diversity. Lexical density represents the proportion of content words in a text. The higher the lexical density, the more information contained in a limited space, and the more condensed the language is. Lexical sophistication reflects the distribution of words with different frequency in the text. The more high-frequency words, the less sophisticated the language is. Lexical diversity refers to the variety of words in the text, which is the embodiment of the author's linguistic ability. In the present study, LD, LS2 and Uber index were selected respectively to investigate the three dimensions of lexical complexity of Chinese and foreign master thesis abstracts.

The results show that the difference in lexical density of the two series of papers is very small, which means that both natives and Chinese-English learners are aware of the information function of academic writing and try to achieve that function by using as few function words as possible. In addition, nonnative postgraduates seem to attach more importance to the amount of information they can convey and consciously use more content words, resulting in a slightly higher lexical density.

However, the difference in lexical sophistication and diversity are statistically significant. To put it more specific, containing the same amount of information, theses abstracts written by natives show higher lexical diversity but lower lexical sophistication. That is, natives tend to involve relatively simple but varied words in theses abstracts to convey meaning, while Chinese-English learners are more engaged in using more sophisticated words. The reason for the high lexical sophistication in Chinese Master these abstracts may be attributed to their intention to demonstrate academic professionalism and rigor by using more complex and academic words, which often at the expense of lexical diversity. It should be noted that the use of academic terms or the proportion of that in the text actually do not determine the quality of academic writing. The academic thought revealed between the lines is the core of academic writing, of which natives are clearly aware.

Syntactic complexity is closely related to the length of production unit, the amount of subordination and coordination, and the proportion of particular structures. In the present study, mean length of T-unit (MLT) serves as the indicator for the length of production unit, T-unit complexity ratio (C/T) and dependent clauses ratio (DC/C) for subordination, coordinate phrases per clause (CP/C) for coordination and complex nominals per clause (CN/C) as well as verb phrases per T-unit (VP/T) for phrasal sophistication.

The data indicates that there is no significant difference between two groups with regard to the mean length of T-unit although natives use slightly longer MLT. However, there is statistical difference in terms of the amount of subordination, whereas the opposite is true for the proportion of coordination. That is, natives use much more clauses or dependent clauses instead of more coordinate phrases in their writing. With respect to complexity at the phrasal level, natives involve significantly much more verb phrases and slightly less complex nominals in their writing. That is to say, more clauses and dependent clauses as well as finite or non-finite verb phrases are involves in foreign Master theses, and relatively more coordinate phrases (adjective, adverb, non and verb phrases) and complex nominals participate in Chinese Master theses.

In conclusion, the current corpus-based contrastive study found that at the lexical level, foreign Masters outperform Chinese Masters in lexical diversity but fall behind in lexical sophistication, meaning they are inclined to use more diversified but relatively simple words to express their thoughts. At the syntactic level, foreign Masters excel in the amount of subordination and the proportion of verb phrases, whereas Chinese Masters use more coordinate phrases and complex nominals in academic writing. That is to say, more clauses and dependent clauses as well as finite or non-finite verb phrases are involves in foreign Master theses, and relatively more coordinate phrases (adjective, adverb, non and verb phrases) and complex nominals make up Chinese Master theses.Through the comparison of linguistic features of theses abstracts written by foreign and Chinese Masters, the study indirectly investigates the role nativeness plays in the academic writing.

The results of the present study showed that foreign Masters excelled in lexical diversity but fell behind in lexical sophistication, which means that nativeness may not the core factor determining the expansion of vocabulary size. There may be other factors such as expertise participating in academic writing. Therefore, the role of nativeness at least in theses writing may be limited.

Some researchers and previous studies have come to similar conclusions. Hyland (2016)

indicated that there is little evidence of claiming L2 writers' linguistic uses at a disadvantaged position. Marti et al. (2019) revealed that nativeness is not the major block that leads to the difference in stance construction between native and nonnative novice writers. In their comparative analysis of four corpora of RAs in the field of applied linguistics written by expert native English writers, Turkish non-native expert English writers, novice native English writers, and Turkish novice non-native English writers with regard to the patterns of use and construction of stance, they found that expertise level is an important factor in RA writing as there is little variation in the reporting practices by native and non-native expert writers whereas remarkable variation is found between non-native novice writers and the other groups. In Cotos, Link, and Huffman (2017), they reported that L1 English novice writers did not show significant differences with L2 learners in the learning of genre features. After pedagogical intervention of Research Writing Tutor (an automatic RA writing feedback system), they found that novice native and nonnative writers produced similar texts with the conventional features of RA genre. Though stance construction and genre features are not equal to vocabulary, the results, as revealed by the results of Marti et al. (2019), Cotos et al. (2017), and the current study, seem to point to a similar conclusion that nativeness has a limited role to play in the quality of academic writing.

CONCLUSION

The current corpus-based contrastive study aims to analyze the linguistic features of thesis abstracts written by Foreign Masters and Chinese Masters, from which some significant differences concerning lexical and syntactic complexity are observed. Results showed that, at the lexical level, foreign Masters are inclined to use less complicated but more diversified words, while Chinese Masters tend to use more sophisticated but less various words to write an abstract that contains roughly the same amount of information, which could lead to such a speculation that nativeness does not play a decisive role regarding lexical richness. There may be other factors such as expertise involving in academic writing. When it comes to the syntactic features, the most distinguishing features of foreign Masters' thesis abstract is the use of clauses or dependent clauses and definite or non-definite verb phrases. Such a feature is in accordance with what has been confirmed by previous studies that the proportion of subordination in writing increases with the improvement of language proficiency.

The findings of the study provide references not only for academic beginners to sharpen their academic writing skills but also for teachers to reflect on their teaching practice and the suitability of their teaching materials. Such salient linguistic features revealed by the study first shed light on the differences between foreign and Chinese Masters' academic writing, which is conducive for academic beginners to better understand their

writing shortcomings, and then gradually and pointedly improve their writing ability under the guidance of teachers. Greater awareness of lexis and syntax on the part of the teachers is also essential since they make suitable plans to train students' writing skills and decide on the order of writing exercises, the degree pf difficulty in terms of writing tasks and on the writing models for students to refer to.

Due to limited condition and time, there are still some shortcomings in the present study: 1) The size of the corpora should be expanded and the selection of Master theses needs to be more representative. Only 100 Master theses related to linguistics are selected for the present study, and the number of papers in each subfield is different, which may influence the validity of the conclusion, resulting in the failure to comprehensively and systematically reflect the linguistic features of foreign and Chinese Master theses writing; 2) The internal and external validity of the study remains to be further verified. Involving limited measures, the present study is only a tentative research to provide the statistical results of lexical and syntactic complexity of Master thesis abstract writing. To put it more specific, the number of errors in vocabulary use is not taken into consideration in the present study, which may influence the speculation of the role of nativeness to a certain degree. Therefore, detailed and expended study of lexical and syntactic complexity from different aspects should be carried out to further verify the conclusion of the present study. 3) There is room for improvement with regard to the design of the study. L2SCA has been evaluated as a reliable computational system to analyze syntactic complexity. Although it can also measure lexical complexity, the results drawn from which are relatively less detailed than that from some specific vocabulary analyzer such as Lextutor and LFP (Lexical Frequency Profile).

In the future study of linguistic complexity of academic writing, the following aspects require additional attention: 1) learner corpus should also serve as reference corpus to provide a multidimensional view of lexical and syntactic features in articles written by L2 learners. Through comparing L2 students with native writers at different levels, i.e., native beginner students and native experts, the role of nativeness and expertise in terms of lexical richness in English research articles could be further exposed. 2) The influence of disciplinary expertise on academic writing should receive more attention. As inspired by Marti et al.'s (2019) result that the major block leading to the differences of nonnative novice writers derive largely from insufficient disciplinary expertise and awareness rather than nativeness, the future research may delve into the analysis of the influence of academic expertise on the linguistic features in research article writings. 3) There are few studies on linguistic complexity in Chinese. Lexical and syntactic complexity in English academic writing has long been the focus in the field of language testing or language acquisition, whereas the analysis of the relationship of syntactic complexity to L2 Chinese acquisition, development and production quality is quite few.

References

Ai, H. & Lu, X., 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.

Bardovi-Harlig, K., 1992. A second look at the T-unit analysis: Reconsidering the sentence. *TESOL Quarterly* 26: 390-5.

Bulté, B., & Housen, A., 2012. Defining and operationalizing L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), Dimensions of L2 performance and proficiency. Amsterdam: John Benjamins. 21–46.

Casanave, C., 1994. Language development in students' journals. *Journal of Second Language Writing*, 3 (3), 179–201.

Cooper, T. C., 1976. Measuring written syntactic patterns of second language learners of German. *The Journal of Educational Research*, 69 (5), 176–183.

Ellis, R., 2003. Task-based Language Learning and Teaching. Oxford University Press.

Ellis, R. & G. Barkhuizen., 2005. Analyzing Learner Language. Oxford University Press.

Foster, P. & Skehan, P., 1996. The influence of planning and task type on second language per-formance. *Studies in Second Language Acquisition*, 18 (3), 299–323.

Hyland, K., 2016. Academic publishing and the myth of linguistic injustice. *Journal of Second Language Writing*, 31, 58-69.

Johnson M. 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.

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

Lu, X., 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4):474-496.

Lu, X., 2011. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers's language development. *TESOL Quarterly, 45*(1):36-62.

Lu, X. & Ai, H., 2015. Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing, 29*, 16-27

Malvern, D., Richards, B., Chipere, N., & Duran, P., 2004. Lexical diversity and language development: Quantification and assessment. London: Palgrave.

Nation, I. S. P., 2013. Learning vocabulary in another language. Cambridge: Cambridge University Press.

Norris, J. M., & Ortega, L., 2009. Towards an organic approach to investigating CAF

in instructed SLA: The case of complexity. *Applied Linguistics,* 30,555–578.

Ortega, L., 2003. Syntactic complexity measures and their relationship to L2proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*,24, 492-518.

Read, J., 2000. Assessing vocabulary. Cambridge: Cambridge University Press.

Skehan, P., 1998. A Cognitive Approach to Language Learning. Oxford: Oxford University Press.

Skehan, P., 2009a. Lexical performance by native and non-native speakers on language learning tasks. In B. Richards, H.M. Daller, D. Malvern. P. Mears, J.Milton, & J. Treffers-Daller (Eds.), Vocabulary studies in first and second language acquisition: The interface between theory and application (pp.107-124). London: Palgrave Macmillan.

Skehan, P., 2009b. Modelling second language performance: Integrating complexity, accuracy, fluency and lexis. *Applied Linguistics*, 30, 510-532.

Wolfe-Quintero, K., Inagaki, S. & Kim, H.-Y., 1998. Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity. Honolulu, HI: University of Hawaii Press.