International Journal of English Language and Literature Studies
ISSN(e): 2306-0646
ISSN(p): 2306-9910
DOI: 10.55493/5019.v12i3.4852
Vol. 12, No. 3, 213-225.

URL: www.aessweb.com

Corpus-based comparison of lexical complexity in L1 and L2 postgraduate academic writing

Maha Al-Harthi
Department of Applied Linguistics, College of Languages, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.
Email: mnalharthi@pnu.edu.sa

Article History
Received: 2 May 2023
Revised: 20 June 2023
Accepted: 18 July 2023
Published: 15 August 2023

Keywords
First language
Lexical complexity
Lexical density
Lexical sophistication
Lexical variation
Postgraduate academic writing
Second language.

ABSTRACT

This study adopted a corpus-based, contrastive approach to lexical complexity in the academic writing of first language (L1) and second language (L2) postgraduates. Lexical complexity scores were extracted using the Lexical Complexity Analyzer from the Corpus of Arab Proficient Users of English (CAPUE), consisting of Saudi academics' dissertations in applied linguistics. To investigate the potential differences between this corpus and native speakers' corpus, the lexical complexity of writing material from the CAPUE and Corpus of English Native Speakers (CENS) were compared. The computational system employed 25 lexical complexity measures to investigate differences in the two groups' lexical density, sophistication, and variation. The results revealed similar lexical density in the writing of both groups; however, the texts by L1 researchers were more lexically complex for most measures of sophistication and variation. The results have implications for teaching English for academic purposes and highlights areas with inappropriate lexical choices. These findings call for the design of pedagogical interventions to enhance the lexical complexity development of L2 postgraduates.

Contribution/ Originality: It is a corpus-based study of the lexical complexity of Saudi postgraduates' writings of a special kind of academic texts that are rarely investigated. This was conducted in comparison with L1 postgraduate researchers in terms of lexical density, sophistication, and variation. The focus on Saudi academics' dissertations contributes to the originality of this study.

1. INTRODUCTION

The past three decades have witnessed a prominent investigation of second language (L2) proficiency to identify factors that enhance L2 learners' proficiency and how it can best be measured (Housen, Kuiken, & Vedder, 2012). Three different components—complexity, accuracy, and fluency (CAF)—can be used as valid and reliable indicators of learners' general proficiency (Housen et al., 2012; Norris & Ortega, 2009). L2 proficiency is primarily measured quantitatively in the CAF framework using frequencies, indices, and ratios. Although the three dimensions are closely interrelated, this study focuses on complexity. Bulté and Housen (2014) reported that in first language (L1) and L2 research, complexity has proved to be a "valid and basic descriptor of L2 performance, as an indicator of proficiency and as an index of language development and progress." Therefore, complexity measures can be used as indicators for more general constructs, such as L2 writing proficiency, development, or maturity (Ortega, 2003; Wolfe-Quintero, Inagaki, & Kim, 1998).

However, there is a lack of consensus among L2 researchers on the definition of complexity (Bulté & Housen, 2012; Bulté & Housen, 2014; Housen et al., 2012). Housen et al. (2012) defined complexity as "the ability to use a wide and varied range of sophisticated structures and vocabulary in the L2." Lexical and syntactic are the most popular complexity measures (Ortega, 2012). This study investigates the hypothesis that lexical complexity may contribute to the differences in academic English discourse written by high-proficiency English as a Foreign Language (EFL) learners and natives (i.e., L1 and L2 postgraduates).

Lexical competence measures can be classified into breadth of knowledge measures (e.g., lexical variation and word frequency), depth of knowledge measures (e.g., synonymy, polysemy, and word associations), and accessibility of lexical items (e.g., word concreteness and familiarity) (for details, see Crossley and McNamara (2009)). Lexical competence in learner production involves a proficient selection of diverse and sophisticated words in a particular genre.

The practical significance of this study is that the academic lingua franca of the international academic research community and an increasing number of L2 postgraduate students write their dissertations in English, yet many of them face linguistic challenges at the lexico-grammatical level. Their ability to compose high-quality dissertations indicates their academic achievement in research-based postgraduate programs and international publications. The development of L2 competence involves knowledge of the lexical and grammatical features of the language and a clear understanding of the linguistic variation across genres. It includes a skillful application of these linguistic resources and conformity to the conventions of academic research.

This study explores the lexical distinctions of high-proficiency Saudi academics compared to L1 researchers, and raises awareness regarding the linguistic and academic requirements and constraints of specific genres. While such academic discourse has been investigated in several studies (e.g., ElMalik and Nesi (2008)), this study examines the lexical complexity of postgraduate academic discourse within the corpus-based framework. I have employed two corpora of academic writing: the Corpus of Arab Proficient Users of English (CAPUE) and Corpus of English Native Speakers (AlNafjan, 2015). The primary objective is to examine the different patterns of lexical complexity in the academic writing produced by L1 and L2 writers through the constructs of diversity, sophistication, and density. Second, I identify potential differences between the two and evaluate their impact on the use of lexical complexity as an indicator of linguistic proficiency.

Using a native-speaker baseline to examine and evaluate the performance of L2 learners appears to be a rather neglected dimension (e.g., Foster and Tavakoli (2009)). I obtain data to address this research gap. Therefore, I systematically use a native-speaker baseline to investigate the lexical complexity of the academic writing produced by Saudi postgraduate students. Comparing the performance of non-native speakers (NNS) and native speakers (NS) allows us to determine whether and to what extent NNS performance deviates from or approximates that of NS. This comparison provides valuable information that L2 educators and curriculum designers can utilize to devise appropriate pedagogical interventions targeting specific problem areas.

2. LITERATURE REVIEW: LEXICAL COMPLEXITY IN L2 WRITING

Lexical complexity can be defined as the range and degree of density, sophistication, and variation in lexical constructions during language production. As it reflects the learner's ability to communicate effectively in L2, it has been recognized as a highly salient construct in L2 writing, teaching, and research (e.g., (Bulté & Housen, 2012; Daller, Milton, & Treffers-Daller, 2007; Lu, 2012; Malvern, Richards, Chipere, & Durán, 2004; Read, 2000; Wolfe-Quintero et al., 1998)).

Nevertheless, there is a terminological problem associated with the term "lexical complexity." Various analogous terms, such as lexical richness (e.g., (Daller et al., 2007; Malvern et al., 2004)), lexical density (e.g., O'Loughlin (1995)), lexical sophistication or rareness (Read, 2000), linguistic variation or variety (e.g., Granger and Wynne (2000)), lexical range and balance (Crystal, 1987), and vocabulary density or lexical diversity, have been

used frequently and interchangeably with lexical complexity. This has led to conceptual confusion while interpreting and comparing the results of individual studies (Bulté & Housen, 2012; Norris & Ortega, 2009). Read (2000) defined lexical diversity, sophistication, and variation as different aspects of lexical complexity. Therefore, I conceptualize this study as a multidimensional approach to language use, which explores these interrelated components.

Lexical complexity in second language acquisition has recently received considerable attention in applied linguistics research. Lexical complexity measures have been developed and tested for assessing learners' language production. They have been associated with characteristics of language use, such as the diversity, sophistication, and density of the lexical items in language production (Read, 2000). Most of these measures gauge lexical complexity by quantifying one of the following: lexical density, lexical sophistication, degree of verb sophistication, range of word types, degree of variety of nouns, or type-token ratio (TTR). The correlation between lexical complexity measures in L2 writing and measures of accuracy and fluency has been the focus of a major branch of research (e.g., (Ai & Lu, 2010, 2013; Crossley, Salsbury, & McNamara, 2015; Daller et al., 2007; Lahmann, Steinkrauss, & Schmid, 2016; Laufer & Nation, 1995; Malvern et al., 2004; Milton, 2009; Read, 2000; Zareva, Schwanenflugel, & Nikolova, 2005)).

When investigating L2 lexical development, the extensive use and range of sophisticated vocabulary in the learners' production can provide clear indications of L2 proficiency. Several studies have looked into the relationship between lexical complexity and the quality of L2 learners' writing and speaking. They found that lexical richness measures used as indices of English as a second language (ESL) learners' speaking task performance had a significant impact (Yu, 2010), as did the control of genre-producing close correspondence in intermediate learners' vocabulary size, as measured by their writings and direct measures of vocabulary size (Laufer & Nation, 1995). Research has suggested that the lexical complexity of spoken and written texts differs (Halliday, 1985; Yu, 2010). Furthermore, some longitudinal studies have looked at changes in lexical complexity in L2 writing over time (e.g., Li and Schmitt (2009)). Others, on the other hand, believe that lexical complexity increases during lexical development (e.g., (Milton, 2013; Wolfe-Quintero et al., 1998; Yu, 2010)). Researchers have sought to investigate lexical complexity by looking at metrics of verb sophistication or noun variation rather than total vocabulary. The goal is to investigate learners' lexical behavior and its link to language proficiency. For example, Lu (2011) did a thorough analysis of a large range of complexity measures and concluded that more particular complexity measures and certain generic measures can be utilized to assess proficiency. He also argued that different characteristics per clause ratios are the greatest predictors of different L2 proficiency levels.

However, few studies have examined lexical complexity from an explicitly cross-linguistic perspective (e.g., (De Clercq, 2015; Harley & King, 1989; Vedder & Benigno, 2016)). Furthermore, researchers investigating the measurement of lexical complexity experience challenges involving the small number of computational tools in automating lexical complexity analysis in L2 writing and the labor-intensiveness of manual analysis. Consequently, there are few studies that have examined lexical complexity in L1 and L2 students' writing, with a limited number of lexical complexity measures applied to relatively small amounts of data. More corpus-based studies are required to advance our understanding of language-specific and universal trends in lexical development and ascertain whether the relation between lexical complexity and linguistic proficiency is the same for L2 learners.

The literature reveals considerable inconsistencies and variability among studies in defining lexical complexity measures, language tasks used, sample size, corpus length, and so on (Wolfe-Quintero et al., 1998). Read (2000) provided a comprehensive review of these measures in book-length research synthesis. Lu (2011) posited that the set of measures reviewed in this research synthesis, about 100 linguistic complexity measures in L2 writing development studies, represented the complete picture of the range of measures investigated in L2 writing research. This study investigates lexical complexity, a multidimensional feature of a learner's language use that consists of three interrelated components: lexical density, lexical sophistication, and linguistic variation.

Lexical density refers to the ratio of semantically lexical (as opposed to function) words to the total number of words in a text. Lexical words are open-class content words, such as nouns, verbs, adjectives, and adverbs, whereas function words denote grammatical words, such as prepositions, determiners, and auxiliaries. Lexical density reflects the information packaging of a text (Malvern et al., 2004). Written texts display a remarkably higher lexical density than spoken ones (Halliday, 1985), and a higher lexical density reflects a higher degree of concentration while presenting ideas in a text (Read, 2000). The present study applied the classification suggested by Lu (2012) for lexical words. He included verbs (excluding modal verbs), nouns, adjectives, adverbs with an adjectival base, and auxiliary verbs ("be" and "have") as lexical words. The higher percentage of content words in a text, the greater its lexical density.

The relative proportion of sophisticated words in a learner's text is measured as lexical sophistication. According to Read (2000), sophisticated words are uncommon, difficult, or advanced. Low-frequency words, for example, are widely regarded as sophisticated (Laufer & Nation, 1995; Vermeer, 2000); thus, it is sometimes referred to as lexical rarity. Laufer and Nation (1995) proposed the lexical frequency profile (LFP) approach, which assesses the percentages of word types in a text—the 2000 most frequent words are considered the "basic 2000," while those that are among the "beyond 2000") are considered sophisticated words. They proposed this model as a reliable measure of lexical richness that can also provide a measure of lexical sophistication, calculated as the ratio of sophisticated words (the "beyond 2000" words) in a text to the overall number of lexical words (Wolfe-Quintero et al., 1998). This study looks at five different lexical sophistication metrics, listed in Table 1. If a lexical word, or verb is not among the 2000 most frequent words generated by the British National Corpus (BNC; Leech, Rayson, and Wilson (2001)), it is deemed sophisticated. The BNC word list was used since it is based on a 100 million-word corpus.

Lexical variation, also known as lexical diversity or lexical range (Crystal, 1987), refers to "the range and variety of vocabulary deployed in a text by either a speaker or a writer" (McCarthy & Jarvis, 2007). Lexical diversity can also be utilized to predict learners' overall language proficiency and to measure the quality of their writing (e.g., Laufer and Nation (1995)). Several lexical variety metrics have been proposed in the literature (for a list, see (Malvern et al., 2004; Yu, 2010)). The number of different words (NDW) used in a text and the simple TTR (the ratio of various word types to all word tokens used in a corpus) are perhaps the most straightforward and prominent metrics of lexical variation. However, NDW is affected by the length of the linguistic sample since samples with different lengths require some type of normalization. Therefore, three revised measures have been proposed such as number of different words in the first 50 words (NDWZ), number of different words random 50 words (NDWERZ), number of different words expected sequence 50 words (NDWESZ).

TTR has lately been called into question due to its sensitivity to the length of the analyzed text, as the ratio tends to get smaller when the sample size increases (Arnaud, 1992). According to McCarthy and Jarvis (2007), "the more words (tokens) a text has, the less likely it is that new words (types) will occur." If a text is excessively long, high-frequency words will be repeated more frequently than low-frequency words, and this tendency will increase with the length of the text. Such sample size effects on lexical variation metrics may produce misleading results. As a result, various revised lexical variation measures have been proposed, including the mean segmental type/token ratio (MSTTR), corrected type/token ratio (CTTR), root type/token ratio (RTTR), the bi-logarithmic type/token ratio (LogTTR), and the Uber Index (see Table 1). TTR and some of its transformations have also been used to assess the variation of specific classes of words, such as the number of verb types (verb variation measure, computed as the ratio of the number of verb types to the total number of verbs in a text), noun types, adjective types, adverb types, and modifier (adjective and adverb) types.

Read (2000) added another component—the number of lexical errors that affect accuracy. However, this study did not investigate this component, as the corpus used is not tagged in terms of errors. Instead, I took advantage of the newly developed *L2 Lexical Complexity Analyzer* (LCA; Ai and Lu (2010)—a computational system designed to

automate the analysis of lexical complexity of writing samples produced by college-level L2 English learners using 25 distinct metrics proposed in the literature. In total, 3 types of lexical complexity measures are assessed in the analyzer: 1 lexical diversity measure, 5 lexical sophistication measures, and 19 measures of lexical variation, as shown in Table 1.

Table 1. Lexical complexity measures.

Measure	Code	Measure	Code
Lexical density	LD	Corrected TTR	CTTR
Measures of lexical sophistication:	LS	Root TTR	RTTR
Lexical sophistication-I	LS1	Bi-logarithmic TTR	LogTTR
Lexical sophistication-II	LS2	Uber index	Uber
Verb sophistication-I	VS1	Lexical word variation	LV
Corrected VS1	CVS1	Verb variation-I	VV1
Verb sophistication-II	VS2	Squared VVI	SVV1
Measures of lexical variation:	LV	Corrected VVI	CVV1
Number of different words	NDW	Verb variation-II	VV2
NDW (First 50 words)	NDW-50	Noun variation	NV
NDW (Expected random 50)	NDW-ER50	Adjective variation	AdjV
NDW (Expected sequence 50)	NDW-ES50	Adverb variation	AdvV
Type-token ratio	TTR	Modifier variation	ModV
Mean segmental TTR (50)	MSTTR-50		

Note: Lu (2012).

This study systematically investigates the extent to which L1 and L2 postgraduate researchers' writing differs in terms of lexical complexity, conceptualized here as a multifaceted construct encompassing lexical density, lexical sophistication, TTR, number of different words, degree of verb sophistication, and other measures. To achieve this objective, the following research questions are posed.

2.1. Research Questions

1. Are there any significant differences in the postgraduate academic writing of dissertations between the CAPUE and CENS corpora in terms of lexical density?
2. Are there any significant differences between the two corpora in any of the measures of lexical sophistication, and if so, in which aspects, and to what degree?
3. Are there any significant differences between the two corpora in any of the measures of lexical variation, and if so, in which aspects, and to what degree?

3. METHOD

3.1. Data Collection

This study adopts a cross-sectional research design when drawing English writing samples of native and non-native applied linguistics postgraduate students from the CAPUE and CENS (for details, see AlNafjan (2015) to conduct a contrastive analysis. CAPUE consists of 20 dissertations written by Saudi EFL postgraduate students of applied linguistics at King Saud University, who have passed an English proficiency exam and undergone three semesters of postgraduate coursework before writing their dissertations. Similarly, CENS comprises of 20 dissertations written by native English speakers in the same field. The corpus is compiled from ProQuest Dissertations and Theses Database from major universities in North America. Each thesis in the corpus is annotated with a header that encodes information about the type of the thesis (MA or PhD). This was originally collected as a control corpus for comparing work produced by L2 postgraduate students.

3.2. Data Analyses

The dissertations in the CENS corpus are considerably longer than those by Saudi postgraduates in the CAPUE. This difference, however, should not affect the comparison pursued here, as all the lexical complexity measures employed are computed as ratios of one lexical measure to other incomplete texts.

Both corpora are examined using the LCA—a computational system for the automatic analysis of 25 metrics of lexical complexity (see Lu (2012)). The analyzer can process a single text or a folder's worth of texts. This system accepts a cleaned text file as input and generates a number score for each of the 25 measures, as seen below: First, the cleaned text file is POS tagged, which assigns a label to each token in the text that identifies its POS (e.g., as a verb, adjective, adverb, etc.). Then, the POS-tagged sample is lemmatized. The lemmatized sample is processed by a Python script that computes the values of the 25 measures. Finally, the script counts the number of types and tokens to compute one or more of the metrics. These are the number of word types (T), sophisticated words (Ts), lexical words (Tlex), sophisticated lexical words (Tslex), verbs (Tverb), sophisticated verbs (Tsverb), nouns (Tnoun), adjectives (Tadj), and adverbs (Tadv), the number of tokens of words (N), lexical words (Nlex), sophisticated lexical words (Nslex), and verbs (Nverb).

Tables 2 and 3 present the descriptive statistics of each corpus length in terms of the number of different word types written per dissertation. For each measure, the table contains its minimum and maximum values, sum, and average across all samples in our dataset. For example, N in the first corpus ranges from 12,925 to 63,882; T ranges from 2,516 to 8,293; T_S was 60,368 and sophisticated word tokens are 193,582, as shown in Table 2.

Table 2. Descriptive statistics summary of the CAPUE dataset.

Type	Sum	Average	Min.	Max.
Sentences	8,126	406.3	263	1,068
Wordtypes	75,391	3,769.55	2,516	8,293
Swordtypes	60,368	3,018.4	1,923	7,291
Lextypes	3,161	158.05	97	317
Slextypes	2,428	121.4	70	263
Wordtoken	460,957	23,047.85	12,925	63,882
Sword token	193,582	9,679.1	5,611	27,797
Lextoken	13,997	699.85	264	1,738
Slextoken	7,629	381.45	151	959

Table 3 reveals that N in the CENS corpus ranges from 13,629 to 97,313; the sums of T and T_S are 126,754 and 107,252, compared to 75,391 and 60,368 in the CAPUE, respectively. Furthermore, the CENS contains 4,820 T_{slex} and 17,014 N_{slex}, compared to 2,428 T_{slex} and 7,629 N_{slex} in CAPUE, respectively.

Table 3. Descriptive statistics summary of the CENS dataset.

Type	Sum	Average	Min.	Max.
Sentences	11,881	594.05	1000	1214
Wordtypes	126,754	6,337.7	2,473	15,411
Swordtypes	107,252	5,362.6	1,877	13,931
Lextypes	5,831	291.55	108	730
Slextypes	4,820	241	81	649
Wordtoken	922,482	46,124.1	13,629	97,313
Sword token	379,007	18,950.35	5,835	40,811
Lextoken	31,974	1,598.7	363	3,579
Slextoken	17,014	850.7	214	1,527

In general, the number of sentences, T, Ts, Tlex, Tslex, N, sophisticated word tokens, Nlex, and Nslex of L2 postgraduate students are shorter and fewer than those of the L1 group.

The computational system used in this study considers a word, lexical word, or sophisticated verb if it is not among the 2,000 most frequent words in the BNC word list. The results of the 25 lexical complexity measures

generated by the system are imported to R for further statistical analyses. Lu (2014) illustrated a detailed, step-by-step procedure for using the LCA. The analyzer requires the input text to be POS-tagged using the PennTree Bank Tagset as well as lemmatized. The first step is to tokenize the text into individual words or tokens to identify word boundaries. Once the text is tokenized, the analyzer calculates the frequency of each word in the text. The analyzer defines lexical words as nouns, adjectives, verbs (excluding modal verbs and the verbs be and have), and adverbs with an adjectival base. This information is used to determine the prevalence of certain words and to identify frequently used or common words. To measure lexical complexity, the analyzer often relies on external resources such as dictionaries or word lists that provide information about the difficulty level or complexity of individual words. These resources may include measures like word frequency rankings, syllable counts, or readability scores. For the lexical sophistication measures, the analyzer defines sophisticated words as those that are not among the first 2,000 most frequent lemmas found in either the British National Corpus (for texts with British spelling) or the American National Corpus (for texts with American spelling). Based on the information gathered from word frequency analysis and linguistic features, the analyzer assigns a complexity metric to the text. This score can be a numerical value or a qualitative assessment, indicating the level of complexity or difficulty of the text.

4. RESULTS

This study used the LCA to calculate 25 lexical complexity indexes, representing three lexical complexity measures, for each text in the two corpora. After calculating the three measures for each text, statistical tests were run to investigate whether there were any significant differences in the measures between the two corpora using R. The significance level was set at $p < .05$. Information on the statistical significance of the differences in the scores is presented below.

4.1. Research Question 1

The first question concerns identifying significant differences in lexical density between the two corpora. For the lexical density measure, the mean value of the L2 corpus (0.0315) is lower than that of the L1 (0.0355). A Welch two-sample t-test was run to determine whether the difference was significant.

Table 4. Welch two sample t-test of lexical density.

T	df	P-value
-1.682	33.779	0.102

Table 4 shows that there is no significant difference ($p > .005$) and the degree of freedom is 33.779. Thus, the proportion of lexical words in the two corpora does not appear to be significantly different (i.e., L1 and L2 corpora yield similar levels of lexical density). Both corpora contain content words at a moderately high level.

4.2. Research Question 2

After identifying no statistically significant differences in the lexical density of L1 and L2 postgraduate researchers' academic writing, it was investigated whether the writing of Saudi postgraduate students approximates that of native speaker (NS) postgraduates in terms of lexical sophistication more closely. Multivariate analysis of variance (MANOVA) was employed for this purpose because more than one dependent variable was to be analyzed simultaneously.

Five lexical sophistication measures were investigated. The lexical sophistication value of the samples in the CAPUE ranged from 0.5 to 0.68, whereas in the CENS it ranged from 0.4 to 0.78. The verb sophistication measure was computed as the ratio of the number of sophisticated verb types (T_{sverb}) to the total number of verbs (N_{verb}) in a text. The average of the verb sophistication is 7.587 in the CAPUE and 14.20 in the CENS. The corrected verb

sophistication in the CAPUE ranged from 0.01 to 2.49, while in CENS it ranged from 1.56 to 4.77. Comparing the two corpora revealed that the L1 corpus contained more sophisticated words than the L2 corpus.

Furthermore, the results indicate a statistically significant difference between the two corpora in terms of lexical sophistication (F=4.34, P=0.003), as shown in Table 5.

Table 5. Fstat of lexical sophistication.

Fstat	P-value
4.336	0.003

As the mean difference is significant, simultaneous confidence intervals (SCIs) for the mean differences of the five measures of lexical sophistication were computed, as shown in Table 6. These simultaneous intervals were used for linear combinations among the variables.

Table 6. Simultaneous confidence intervals (SCI).

a	SCI	
a=c(1,0,0,0,0)	-0.107	a=c(1,0,0,0,0)
a=c(0,1,0,0,0)	-0.081	a=c(0,1,0,0,0)
a=c(0,0,1,0,0)	-0.081	a=c(0,0,1,0,0)
a=c(0,0,0,1,0)	-14.171	a=c(0,0,0,1,0)
a=c(0,0,0,0,1)	-1.316	a=c(0,0,0,0,1)

Next, the mean differences for each individual variable were focused upon. The output provided the lower and upper bounds for the SCI of the mean differences. As Table 6 reveals, the SCI of A with a=c (1,0,0,0,0) shows that one can be 99% confident that the mean difference of LS1 lies between -0.107 and 0.088; that of LS2 lies between -0.081 and -0.003; that of VS1 lies between -0.081 and 0.103; that of VS2 lies between -14.171 and 0.937; and that of CVS1 lies between -1.316 and -0.006.

Thus, three (LS1, VS1, and CVS1) out of the five SCIs include "0" (Table 6), which implies that the mean differences between the NNS and NS groups for these measures do not differ significantly. Interestingly, a statistically significant difference is revealed in the mean values of two out of the five measures of lexical sophistication that gauge the degree of word sophistication between the NNS and the NS groups. Therefore, this significant difference appears to be due to these two lexical sophistication measures involved, namely, LS2 and VS2. These results can be interpreted as follows: the NNS postgraduate researchers produced significantly fewer complex lexical sophisticated words and sophisticated verbs than the NS researchers did. These results appear to be consistent with the findings of Harley and King (1989) and Linnarud (1986), who demonstrated that L1 writers used significantly more sophisticated words than L2 writers.

4.3. Research Question 3

The third question focuses on determining whether there was any significant difference in lexical variation between the two corpora involved in this study. Again, MANOVA was used to determine whether the mean complexity values for the L1 and L2 groups differ significantly. The third section of the analyzer was concerned with 19 measures of lexical variation—NDW, NDWZ, NDWERZ, NDWESZ, TTR, MSTTR, CTTR, RTTR, LogTTR, Uber, LV, VV1, SVV1, CVV1, VV2, NV, ADJV, ADVV, and MODV (see Table 1).

The lexical variation complexity scores computed by the system are as follows: the NDW, TTR, MSTTR, CTTR, RTTR, LogTTR, and the Uber Index. In addition, some TTR transformations were used to evaluate the variation of specific classes of words, such as several verb types, noun types, adjective types, adverb types, and modifiers (adjective and adverb) types.

Data from the first corpus were compared with those of the second corpus, which suggest significant differences in most aspects of lexical variation between NNS and NS postgraduate students' writing, as shown in Tables 7 and 8.

Table 7. Fstat of lexical variation.

Fstat	P-value
3.085	0.002

Fstat test (the Hotelling's T square) revealed statistically significant difference ($p < .005$) in the mean value of the lexical variation measures between the two corpora. Again, as the difference is significant, SCIs for the mean differences of the 19 measures of lexical variation between the two groups were computed (Table 8).

Table 8. Simultaneous confidence intervals (SCI).

a	SCI	
a=c(1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)	-8271.957	a=c(1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
a=c(0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)	-10.245	a=c(0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
a=c(0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)	-3.000	a=c(0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
a=c(0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)	-4.122	a=c(0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
a=c(0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)	-0.041	a=c(0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
a=c(0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0	-0.056	a=c(0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0
a=c(0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0)	-11.692	a=c(0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0)
a=c(0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0)	-16.535	a=c(0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0)
a=c(0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0)	-0.024	a=c(0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0)
a=c(0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0)	-4.936	a=c(0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0)
a=c(0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0)	-0.095	a=c(0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0)
a=c(0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0)	-0.230	a=c(0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0)
a=c(0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0)	-26.851	a=c(0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0)
a=c(0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0)	-2.149	a=c(0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0)
a=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0)	-0.032	a=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0)
a=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0)	-0.105	a=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0)
a=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0)	-0.021	a=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0)
a=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0)	-0.023	a=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0)
a=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0)	-0.033	a=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0)
a=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1)		a=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1)

The SCI of A with a=c(1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0) shows that one can be 99% confident that the mean of "NDW" will lie between -8271.57 and 3135.657 and that the mean of "NDWZ" will lie between -10.245 and 5.045. It follows all other lexical variations. Contrary to expectations, all the above SCIs cover zero, although the *Fstat* test shows significant differences between the two corpora in lexical variation. I attribute this to the combined contributions of the various variables. These results suggest that L2 postgraduate learners produced significantly fewer different words, lower TTRs, and a lower ratio of the number of verb, noun, adjective, and adverb types, and modifiers compared to L1 researchers when writing their dissertations. Thus, statistical analyses reveal a significant difference between the two corpora for two of the three measures investigated. To examine the effect size or the strength of the significant differences, SCIs were computed for the significant differences of each measure. The results reveal similar levels of lexical density between the two corpora, but considerable differences in terms of sophistication and diversity. Taken together, these results provide important insights about the lexical complexity of L2 compared to native researchers. The lexical density measure used in this study is not shown to be a feature that distinguishes how the L1 and L2 corpora differ lexically.

5. DISCUSSION

Evidence suggests that the higher-proficiency L2 group significantly approximate the L1 group in areas of lexical density, thus implying that postgraduate English learners have largely mastered this dimension of lexical

complexity. This is not too surprising, as the difference between lexical and grammatical words is usually introduced at the early stages of English instruction. These results are consistent with those of De Clercq (2015). However, the latter highlighted similar developmental tendencies for lexical diversity in L2 French and English, but considerably different developmental tendencies in terms of sophistication and density. In this study, the L2 and L1 groups differ because the native writers produce a significantly higher proportion of advanced and sophisticated vocabulary than L2 writers. Thus, L2 learners, as their English proficiency level develops, may be expected to use rarer, more complex words in their written output; subsequently, their lexical complexity and profiles may become more native-like (Lindqvist, Bardel, & Gudmundson, 2011). It is widely acknowledged that lexical complexity in L2 learners will develop very slowly and sometimes unevenly. Advanced and competent learners are no exception, as they often experience problems using appropriate vocabulary and collocations (Li & Schmitt, 2009; Nesselhauf, 2003).

The common practice is to set the general NS standard as a target for EFL learners, with little focus on these students' motivation and usage requirements. However, the notion of "advanced learners" must be redefined, given the current status of English as the global lingua franca in academia. English is becoming the preferred language for cross-cultural and cross-national communications in many fields, such as business, politics, academia, and the media. Recent advances in learner corpora analysis suggest that the learners' L1 approach (customizing pedagogy to learners' needs) should be considered and adopted (Gilquin & Granger, 2015). EFL pedagogy should consider the specific linguistic needs of the students and customize the material according to their L1 and professional community of practice. Therefore, a more insightful and comprehensive approach that is targeted to the specific needs of learners is required.

Some advanced learners prefer to have accurate English, maintain a distinct style, and do not aim to sound native-like (Jenkins, 2011; Prodromou, 2008). They target communicative efficiency more than native-like attainment and selection. Therefore, the less sophisticated vocabulary selection of L2 learners should be viewed more positively, as an instance of implementation of distinct communication strategies by interlanguage users, such as experimentation, transfer, analogy, and repetition.

However, the lack of lexical complexity can detriment L2 performance, both productively and receptively, which can lead to misunderstandings. More importantly, advanced L2 lexical deviations may also signal a lack of academic expertise. Perhaps the low percentage of international publications among Saudi-applied linguistics researchers is one of the consequences of these lexical complexity differences between L1 and L2 researchers. Compared to the medical sciences and engineering, social sciences have the lowest number of international publications in Saudi Arabia (Al-Ohali & Shin, 2013; Smith & Abouammoh, 2013). L2 researchers must become aware of the linguistic and academic requirements and constraints of their respective fields to enhance their chances of international publication.

Moreover, classroom materials can be designed to improve L2 learners' linguistic diversity. L2 learners should be aware that their lexical complexity can be enhanced by knowing how and where certain words are used together, what patterns they are used in, and how to arrange them to express clear and organized ideas. To achieve complete knowledge regarding a word's usage, they need to grasp the following aspects—word form, word structure, syntactic pattern of the word in a phrase and sentence, meaning, lexical relations of the word with other words (e.g., synonymy, antonymy, and hyponymy), collocations, and idiomatic expressions.

The study findings highlight the importance of educators' awareness of L2 writing regarding the significant gap in certain aspects of lexical complexity between L2 postgraduates and L1 researchers. This gap calls for the design of relevant pedagogical interventions to enhance L2 postgraduates' lexical development. Although sophisticated words are highly recommended for proficient writing, there is also a need to raise L2 learners' awareness about not implementing a large proportion of rare and low-frequency words, as these words alone do not guarantee high-quality writing. Complete mastery of the most frequent words leads to proficient command over the

language. Using diverse words can also enhance the quality of their writing; therefore, it is recommended that L2 learners practice synonyms and expressions with similar meanings.

The interpretation of such quantitative results taken from small datasets should be presented with caution. Because of the considerable variations among the definitions of lexical complexity measures in related literature, it is difficult to pool consistent results to examine the relative performance of different measures (Ortega, 2003; Wolfe-Quintero et al., 1998).

Several important topics were not addressed in this study, due to the scope of the research and the information available in the CAPUE and CENS corpora. First, the postgraduate learner corpus includes writings produced by only L1 Arabic learners. Future research should look at whether similar discrepancies exist between NS and NNS postgraduates from other L1 backgrounds. Thus, the effect of L1 on the lexical development of NNS researchers can be explored. Second, it is recommended to systematically investigate the effects of individual factors on L1 versus L2 variations in lexical complexity. Third, it would be useful to examine the circumstances that lead to highly advanced L2 writers achieving the same level of lexical complexity as L1 writers. Fourth, it is possible to investigate the impact of their lexical complexity on their research productivity. Future research can also look at the various dimensions of lexical complexity and the interrelationships among them.

Research on thesis and dissertation writing should investigate issues such as lexical complexity measures and lexico-grammatical features of postgraduate student writers, to identify specific factors that contribute to the differences between L1 and L2 academic writing. Types of underuse and overuse in postgraduate learners' writing may be identified to help students improve their academic writing.

6. CONCLUSION

This explorative study on the lexical complexity of L2 English learners at advanced levels of proficiency examines whether NNS and NS postgraduate students' academic writing differs in three aspects of lexical complexity: lexical density, sophistication, and variation. While a similar pattern of lexical density was observed between the two corpora, significantly different tendencies in lexical complexity dimensions of linguistic sophistication and variation were noted. On average, Saudi postgraduate researchers produced shorter sentences, fewer words, and a smaller proportion of sophisticated lexical words than NS postgraduate researchers. The findings have pedagogical implications for university writing instructors, centers, and EAP programs that serve large Arab populations globally. They highlight areas with inappropriate lexical choices so that they can be addressed more efficiently.

Funding: This research is supported by the Deanship of Scientific Research, Princess Nourah bint Abdulrahman University, Saudi Arabia (Grant number: 299/(21) 1436-1437).
Institutional Review Board Statement: The Ethical Committee of the Princess Nourah bint Abdulrahman University, Saudi Arabia has granted approval for this study on 28 July 2017 (Ref. No. 299/21).
Transparency: The author states that the manuscript is honest, truthful, and transparent, that no key aspects of the investigation have been omitted, and that any differences from the study as planned have been clarified. This study followed all writing ethics.
Competing Interests: The author declares that there are no conflicts of interests regarding the publication of this paper.

REFERENCES

Ai, H., & Lu, X. (2010). *A web-based system for automatic measurement of linguistic complexity.* Paper presented at the 27th Annual Symposium of the Computer-Assisted Language Instruction Consortium. Amherst, MA.

Ai, H., & Lu, X. (2013). A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. *Automatic Treatment and Analysis of Learner Corpus Data*, 249-264. https://doi.org/10.1075/scl.59.15ai

Al-Ohali, M., & Shin, J. C. (2013). Knowledge-based innovation and research productivity in Saudi Arabia. *Higher Education in Saudi Arabia: Achievements, Challenges and Opportunities*, 95-102. https://doi.org/10.1007/978-94-007-6321-0_9

AlNafjan, E. (2015). *A contrastive corpus analysis of Arabic and English native language speakers' academic English written discourse.* Unpublished Doctoral Dissertation, King Saud University.

Arnaud, P. J. (1992). Objective lexical and grammatical characteristics of L2 written compositions and the validity of separate-component tests. *Vocabulary and Applied Linguistics*, 133-145. https://doi.org/10.1007/978-1-349-12396-4_13

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, and I. Vedder (Eds.). Dimensions of L2 performance and proficiency: Complexity, accuracy, and fluency in SLA. In (pp. 21–46). Amsterdam: John Benjamins.

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, *26*, 42-65. https://doi.org/10.1016/j.jslw.2014.09.005

Crossley, S. A., & McNamara, D. S. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing*, *18*(2), 119-135. https://doi.org/10.1016/j.jslw.2009.02.002

Crossley, S. A., Salsbury, T., & McNamara, D. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*, *36*(5), 570–590.

Crystal, D. (1987). Towards a 'bucket'theory of language disability: Taking account of interaction between linguistic levels. *Clinical Linguistics & Phonetics*, *1*(1), 7-22. https://doi.org/10.1080/02699208708985001

Daller, H., Milton, J., & Treffers-Daller, J. (2007). *Modelling and assessing vocabulary knowledge* (Vol. 10). Cambridge: Cambridge University Press.

De Clercq, B. (2015). The development of lexical complexity in second language acquisition: A cross-linguistic study of L2 French and English. *Eurosla Yearbook*, *15*(1), 69-94. https://doi.org/10.1075/eurosla.15.03dec

ElMalik, A. T., & Nesi, H. (2008). Publishing research in a second language: The case of Sudanese contributors to international medical journals. *Journal of English for Academic Purposes*, *7*(2), 87-96. https://doi.org/10.1016/j.jeap.2008.02.007

Foster, P., & Tavakoli, P. (2009). Native speakers and task performance: Comparing effects on complexity, fluency, and lexical diversity. *Language Learning*, *59*(4), 866-896. https://doi.org/10.1111/j.1467-9922.2009.00528.x

Gilquin, G., & Granger, S. (2015). Learner language. In D. Biber & R. Reppen (Eds.), The Cambridge Handbook of English Corpus Linguistics (Cambridge Handbooks in Language and Linguistics). In (pp. 418-436). Cambridge: Cambridge University Press.

Granger, S., & Wynne, M. (2000). Optimising measures of lexical variation in EFL learner corpora. In (pp. 249-257). Rodopi: Corpora Galore.

Halliday, M. A. K. (1985). *Spoken and written language.* Geelong: Deakin University Press.

Harley, B., & King, M. L. (1989). Verb lexis in the written compositions of young L2 learners. *Studies in Second Language Acquisition*, *11*(4), 415–439.

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency. *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*, *32*, 1-20. https://doi.org/10.1075/lllt.32.01hou

Jenkins, J. (2011). Accommodating (to) ELF in the international university. *Journal of Pragmatics*, *43*(4), 926-936. https://doi.org/10.1016/j.pragma.2010.05.011

Lahmann, C., Steinkrauss, R., & Schmid, M. S. (2016). Factors affecting grammatical and lexical complexity of long-term L2 speakers' oral proficiency. *Language Learning*, *66*(2), 354-385. https://doi.org/10.1111/lang.12151

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307-322. https://doi.org/10.1093/applin/16.3.307

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British national corpus.* London: Longman.

Li, J., & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: A longitudinal case study. *Journal of Second Language Writing*, *18*(2), 85-102. https://doi.org/10.1016/j.jslw.2009.02.001

Lindqvist, C., Bardel, C., & Gudmundson, A. (2011). Lexical richness in the advanced learner's oral production of French and Italian L2. *International Review of Applied Linguistics in Language Teaching*, *49*(3), 221–240.

Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish learners' written English (No. 74)*. CWK Gleerup.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly, 45*(1), 36-62. https://doi.org/10.5054/tq.2011.240859

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal, 96*(2), 190-208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x

Lu, X. (2014). *Computational methods for corpus annotation and analysis*. Dordrecht: Springer.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). Lexical diversity and language development. In (pp. 16-30). New York: Palgrave Macmillan.

McCarthy, P. M., & Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing, 24*(4), 459-488. https://doi.org/10.1177/0265532207080767

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol: Multilingual Matters.

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.). *Eurosla Monographs Series, 2*, 57-78.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics, 24*(2), 223-242. https://doi.org/10.1093/applin/24.2.223

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555-578. https://doi.org/10.1093/applin/amp044

O'Loughlin, K. (1995). Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing, 12*(2), 217-237. https://doi.org/10.1177/026553229501200205

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24*(4), 492-518. https://doi.org/10.1093/applin/24.4.492

Ortega, L. (2012). Interlanguage complexity: A construct in search of theoretical renewal. In: B. Kortmann and B. Szmrecsanyi (Eds.). Linguistic complexity: Second language acquisition, indigenization, contact. In (pp. 127–155). Berlin: Mouton De Gruyter.

Prodromou, L. (2008). English as a lingua Franca: A corpus-based analysis, Continuum. *Applied Linguistics, 29*(3), 521-524.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Smith, L., & Abouammoh, A. (2013). Higher education in Saudi Arabia: Reforms, challenges and priorities. *Higher Education in Saudi Arabia: Achievements, Challenges and Opportunities*, 1-12. https://doi.org/10.1007/978-94-007-6321-0_1

Vedder, I., & Benigno, V. (2016). Lexical richness and collocational competence in second-language writing. *International Review of Applied Linguistics in Language Teaching, 54*(1), 23-42. https://doi.org/10.1515/iral-2016-0015

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing, 17*(1), 65-83. https://doi.org/10.1191/026553200676636328

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity (No. 17)*. University of Hawaii Press.

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics, 31*(2), 236-259. https://doi.org/10.1093/applin/amp024

Zareva, A., Schwanenflugel, P., & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency: Variable sensitivity. *Studies in Second Language Acquisition, 27*(4), 567-595. https://doi.org/10.1017/s0272263105050254