Teaching English as a Second Language Quarterly (TESLQ)
(Formerly Journal of Teaching Language Skills)
40(4), Fall 2021, pp. 61-90
DOI: 10.22099/JTLS.2021.40043.2960

Online ISSN: 2717-1604
Print ISSN: 2008-8191

Research Paper

A Corpus-based Analysis of Lexical Richness in EAP Texts Written by Iranian TEFL Students

Masoud Azadnia *

Department of English Language Teaching, Central Tehran Branch, Islamic Azad University, Tehran, Iran

© Shiraz University

Abstract

The literature on second/foreign language (L2/FL) discourse is replete with corpus-based studies into the use of various features representing lexical proficiency. Nonetheless, the lexical construct of English for academic purposes (EAP) texts developed by postgraduates majoring in teaching English as a foreign language (TEFL) still sounds like a relatively unexplored domain that merits further multi-dimensional investigation. To narrow the gap, the authors in the current study set out to evaluate the lexical richness of a corpus containing doctoral dissertations written by Iranian TEFL students in terms of lexical density, diversity, and sophistication. Taking advantage of the computational tool Coh-Metrix to analyze the lexical features, the corpus was analyzed in comparison with a first language (L1) baseline containing doctoral dissertations written by English native speakers. The comparative analysis of the L1 and L2 corpora revealed that the texts written by Iranian TEFL learners were lexically less diverse but more sophisticated. Additionally, the lexical density of the L2 corpus exceeded that of the L1 one in terms of nouns and adjectives. Based on the results drawn from a discriminant function analysis (DFA), the features representing lexical sophistication and density were found to be the best predictors of lexical richness since they could significantly discriminate between the two sub-corpora. The findings may

Received: 28/03/2021 Accepted: 17/07/2021

* Instructor, Email: masoudazadnia@gmail.com

provide new insights into the ways of evaluating and enhancing the lexical richness of FL/L2 written discourse.

***Keywords**:* Corpus-based Linguistic Analysis, Lexical Density, Lexical Diversity, Lexical Richness, Lexical Sophistication

© Shiraz University

As the cornerstone of every effective language-mediated communication (Schmitt, 2000), vocabulary has acquired a great deal of significance in the realm of second language acquisition (SLA). Aside from the intense scrutiny of effective methods of enhancing vocabulary learning and retrieval, the literature on SLA includes a plethora of investigations (e.g., Sasaki, 2007; Schmitt, 2010; Storch & Tapper, 2009) into the most workable techniques for enriching the lexical construct of L2/FL learners' discourse. These techniques are intended to help learners use their productive vocabulary knowledge while producing a written/oral piece of discourse. In spite of the abundance of vocabulary enhancement and discourse enrichment techniques employed in different language learning contexts, unrestricted access to a native-like lexical database still sounds too far-fetched for L2/FL learners (Laufer & Nation, 1995; Muncie, 2002). This may depict why the creation of either oral or written discourse approximates the lexical construct of native speakers' productions could pose severe difficulties on language learners' shoulders.

In spite of the fact that non-native speakers of a language could deploy a variety of coping skills such as simplification and paraphrasing to compensate for the lack of a native-like lexical repertoire, problems aroused by a deficient lexical database are to be compounded while going through the process of writing academic texts. As stated by Breeze (2008), the major intricacies of academic writing include "to be exact, to be sophisticated, to express complex ideas in complex sentences, to master the techniques of written cohesion rather

© Shiraz University

than to repeat the same basic words, and to cultivate a high, academic register in both vocabulary and syntax" (p. 52). Although a writer's awareness of the more frequent content-relevant technical vocabulary, as well as the words widely used across all academic disciplines, may pave the way for making sure of the lexical appropriateness of an academic-genre text (Cobb & Horst, 2004; Vongpumivitch, Huang, & Chang, 2009), the generation of a lexically well-structured EAP text requires proper regard for more detailed criteria such as word depth/breadth, uniqueness, and accessibility, all known as lexical richness qualities (Meara, 2005).

Providing a vivid picture of how well linguistic features are manipulated to express a subject-specific interpretation, lexically and structurally rich EAP texts not only embody the author's high level of English proficiency (Douglas, 2013; Lavallée & McDonough, 2015) but also open up a golden opportunity to join the international content-specific scientific community (Hirvela, 2011). The need for satisfying lexical richness requirements takes on a special significance in the EFL academic landscape, where FL university students experience serious difficulties going through the laborious process of writing EAP texts such as master's/doctoral theses and scientific articles. The issue is of special importance for Iranian postgraduates involved in TEFL, as members of an academic cluster who are obliged to develop EAP texts in English. Setting an L1 baseline for comparison, the authors in this corpus-based comparative analysis intended to evaluate the current level of lexical richness in the written style of Iranian TEFL students and determine the areas in need of either enhancement or modification.

© Shiraz University

Review of Literature

Lexical Richness: Operational Definitions and Evaluative Measures

As a linguistic feature reflecting the quality of words used in a specific context (Read, 2000), Lexical richness has been conceptualized differently by the broad range of scholars involved in applied linguistics. The underpinnings of such a wide-ranging conceptualization range from a simplistic view on lexical richness as the frequency of various lexical items used in a text (O'Loughlin, 1995) to a balanced focus on a comprehensive list of lexical features including word originality, variety, specificity, simplicity/difficulty, and so on (Laufer & Nation, 1995; Read, 2000). The variety of definitions proposed for lexical richness has excited considerable controversy over the measures that could ideally portray it. While a significant number of scholars (e.g., Engber, 1995; Grobe, 1981; Vermeer, 2004) presumed a single quality to be a useful measure of lexical richness, there are others (e.g., Bulte´ & Housen, 2014; Malvern, Richards, Chipere, & Durán, 2004; Read, 2000) who corroborated the usefulness of a multi-dimensional system for extrapolating lexical richness of written productions.

In spite of the great enthusiasm engendered in linguists for lexical richness analysis, no clear consensus has still been made on how best to conceptualize lexical richness so as to avoid conceptual confusion and facilitate cross-analytical comparisons (Bulte´ & Housen, 2014). Nonetheless, the micro-features widely used to operationalize lexical richness include lexical diversity (i.e., the proportion of various word types in a text), lexical variation (the proportion of individual lexical words in a text), lexical sophistication (the proportion of advanced/sophisticated words in a text), lexical density (the proportion of different sorts of lexical items in a text), lexical originality/individuality (the proportion of words unique to the writer

in the target group), lexical fluency (the number of words used in a text in a given time span), lexical errors, and average word length (Daller, Milton, & Treffers-Daller, 2007; Laufer & Nation, 1995; Read, 2000). However, some of these micro-features (e.g., originality and density) have received severe criticism for being dependent upon either text length or changes in the target group (Douglas, 2010; Gregori-Signes & Clavel-Arroitia, 2015). Additionally, as pinpointed by Šišková (2012), lexical errors and fluency are two features peculiar to an evaluative system aimed at gauging oral/written discourse produced in a timely fashion. Having evaluated a multiplicity of lexical richness measurement systems, Šišková (2012) concluded that a three-component system including lexical sophistication, diversity, and density could successfully measure lexical richness.

© Shiraz University

Lexical Sophistication

Lexical sophistication, also known as lexical rareness, has been widely approved as a central component of various lexical richness evaluation schemes. Lexical sophistication is mainly concerned with the ratio of advanced/sophisticated words to the total number of words used in a part of writing/speech (Crossely & Kyle, 2018; Nation & Meara, 2010; Read, 2000). The operational definitions proposed to conceptualize lexical sophistication are basically grounded in response to the key question: What feature/features does/do a sophisticated word/expression enjoy?. Since the majority of the responses provided to address this leading question were concerned with word frequency, the bulk of definitions proposed heretofore refer to lexical sophistication as the proper use of low-frequency vocabulary items in a text (Kyle & Crossley, 2015; Laufer & Nation, 1995; Malvern et al., 2004; Meara & Bell, 2001; Vermeer, 2004). Nonetheless, the advent of automatic corpus

analysis tools in the two recent decades has facilitated the evaluation of lexical sophistication based on other conceptual criteria such as word familiarity, imageability, concreteness, and so on (Crossley & McNamara, 2011).

Lexical Diversity

In spite of the conceptual and priority-based differences between lexical diversity and lexical richness, the two terms are mainly regarded as interchangeable concepts in the literature (Kim & Jeon, 2016; Kojima & Yamashita, 2014). In an attempt to clear up the distinction between the two variables, Malvern et al. (2004) referred to the definition provided by Laufer and Nation (1995) whereby lexical diversity has been defined as "the ratio in percent between the different words in the text and the total number of running words" (p. 310). Lexical diversity is also defined by Johansson (2009) as the variety rate of the words used in an/a oral/written discourse with a given length. Taking such definitions into account, one can easily infer that a written discourse enjoying a high level of lexical diversity includes a broad range of unique words and, as a result, few instances of word repetition. Accordingly, the ratio of unique words (types) that occur in a text to the total number of words used in a text (tokens) could ideally portray the degree of lexical variation in written discourse. This measure, called Type-Token Ratio (TTR), has been validated widely (e.g., Crossley & McNamara, 2011; Šišková, 2012; Read, 2000) as an index of lexical richness.

Lexical Density

Lexical density, generally defined as the ratio of the total number of content words (i.e., nouns, verbs, adjectives, and adverbs) to the total tokens used in a piece of writing (Daller et al., 2007; Johansson, 2009; Read, 2000),

© Shiraz University

is another linguistic feature presumed to have the potential for describing the lexical richness of written discourse. As claimed by Gregori-Signes and Clavel-Arroitia (2015), the measures evaluating lexical density reflect the lexical construct of the text and provide an informative scheme for gaining an initial understanding of its overall linguistic structure (cohesive and syntactic) construct. The rationale behind including lexical density as a key component of many evaluative systems intended for lexical richness stems from the assumption that the use of more instances of content words facilitates the conveyance of a message denoting complex information through more sophisticated words. Given that a taxonomy of words includes both lexical and functional items, lexically dense writing includes a high proportion of lexical items (content words) of different types (Read, 2000).

Empirical Background to the Study

As the empirical data on the positive correlation between lexical richness and academic success continue to grow (e.g., Douglas, 2010; Ha, 2019; Kwon, 2009; Morris & Cobb, 2004; Šišková, 2012; Staehr, 2008), the study of the lexical construct of EAP texts written by L2/FL learners assumes even greater significance. Putting a central focus on the lexical construct of EAP texts, many researchers sought to provide a workable scheme for lexical enrichment in L2 academic texts. Notwithstanding the abundance of the studies into lexical richness development in L2 EAP texts (e.g., Chen & Baker, 2010; Crossley & McNamara, 2009, 2012; Crossley, Weston, McLain Sullivan, & McNamara, 2011; Djiwandono, 2016; Failasofah & Alkhrisheh, 2018; Gregori-Signes & Clavel-Arroitia, 2015; Ha, 2019; Higginbotham & Reid, 2019; Juanggo, 2018; Kalantari & Gholami, 2017; Storch & Tapper, 2009), few instances of scientific endeavor have been made in recent years (e.g.,

© Shiraz University

© Shiraz University

Breeze, 2008; Douglas, 2010; Kusumaningrum & Ardi, 2020; Kwon, 2009; Šišková ; 2012) so as to specifically compare/contrast the lexical richness features of EAP texts composed by non-native English speaking writers and the target style, operationalized as essays written by natives. Additionally, a detailed review of the literature corroborates the scant attention paid to the analysis of lexical richness in academically-bound texts (master's, thesis, doctoral dissertations, and research articles) written by TEFL students.

Acknowledging the claim made by the bulk of the previous studies that lexical richness is a multi-faceted concept (Bulté & Housen, 2014; Schmitt, 2010; Zheng, 2016), the current study focused on lexical sophistication, density, and diversity as the widely-approved descriptors of writing quality. What acts as the main incentive for the authors of the present study was the pedagogical need for lexical richness evaluation and enhancement in EAP texts written by Iranian TEFL students. The present comparative study also aimed to ascertain which lexical features could significantly account for the variety of lexical richness between the texts composed by Iranian TEFL learners and those written by native speakers of English. To pursue the objectives enumerated above, the following research questions guided the current study:

1. To what extent do EAP texts written by Iranian TEFL students approximate those written by native speakers of English in terms of lexical richness?
2. Which lexical richness features significantly discriminate between EAP texts written by Iranian TEFL students and those written by native speakers of English?

© Shiraz University

Method

Corpus

The corpus of the study was comprised of two sub-corpora, including an L2 and an L1 corpus. The L2 corpus contained 182 texts (46757 words) extracted from postgraduate dissertations written by Iranian TEFL students at the Ph.D. level. The L1 corpus included 103 texts (28188 words) extracted from doctoral dissertations developed native speakers of English. The cluster from which the L2 corpus was sampled included the official register of Islamic Azad University (IAU) of Isfahan, Khorasgan Branch, whereas the L1 corpus was chosen via the Internet. Although the two clusters were chosen due to their availability and accessibility to the current study's authors (convenience sampling), random sampling and purposive sampling were employed to decide on the L2 and L1 dissertations, respectively. While random sampling increased the representativeness of the main (L2) corpus, purposive sampling maximized the between-corpus homogeneity, facilitating the selection of a comparison (L1) corpus enjoying several properties identical to the main one. Aside from the authors' national background (Australian, Canadian, American, and British) and academic degree (Ph.D. student), the criteria taken into account while sampling the L1 dissertations included genre (academic), topic (English teaching and Applied linguistics), publication date (between 2000 and 2018), and text length (between 400 and 1000 words). Of all the sections included in the dissertations, Discussion was decided on for analysis.

Design

The current corpus-based comparative study adopted a descriptive approach to data collection and analysis. The lexical construct in the L1 and L2 corpora was analyzed and compared in terms of various features representing lexical richness. The descriptive design was deemed to ideally

suit the research objectives since limited information is available on the topic of inquiry in the current study. The appropriateness of a descriptive design in cases in which detailed information is required to build up a more vivid picture of the phenomenon under study has been validated empirically (Bickman & Rog, 1998).

Computational Tool

The automated web tool Coh-Metrix (version 3.0) was employed to analyze the two sub-corpora's lexical patterns. Coh-Metrix is a computational tool that analyzes a written discourse in terms of a total of 108 syntactic, cohesive, and lexical indices by integrating different linguistic components (e.g., lexicons, pattern classifiers, syntactic parsers, and semantic interpreters (Jurafsky & Martin, 2002). The rationale behind employing Coh-Metrix was its capability to measure both count-based and band-based lexical indices. This foresight for measuring a broad range of lexical indices well suited the multi-dimensional conceptual framework of the study, whereby lexical richness was referred to as a combination of lexical density, diversity, and sophistication. The validity and reliability of Coh-metrix have been established earlier through research (Crossley & McNamara, 2011; Crossley, Salsbury, McCarthy, & McNamara, 2008).

Analytical Procedure

As the preliminary stage of the analytical procedure, the soft copy of the two sub-corpora was fed into Coh-Metrix. The texts were then cleaned and formatted, removing oddities (i.e., non-English letters and strings of mathematical symbols), pictures, charts, and diagrams. The TextPad software then converted into the Coh-Metrix-readable format (txt-type files) and sent to the Coh-metrix team for text processing and lexical construct evaluation. The computational results reported by the Coh-Metrix team were used to

© Shiraz University

address the two research questions quantitatively. As a stage prerequisite to the quantitative data analysis, however, a total of 13 lexical indices were selected from the broad range of features computed by Coh-Metrix, taking account of the definitions that underpinned the three-component conceptual framework of the study. The indices are displayed in Table 1.

Table 1

Lexical Indices Included in the Analytical Process

Sub-component	Index	Description
Lexical Diversity	TTR for content words	the number of unique content words (types) divided by the number of tokens of content words
Lexical Density	Noun incidence	the incidence score (occurrence per 1000 words) of nouns in the text
	Verb incidence	the incidence score of verbs in the text
	Adjective incidence	the incidence score of adjectives in the text
	Adverb incidence	the incidence score of adverbs in the text
Lexical Sophistication	CELEX word frequency	the average word frequency for content words based on CELEX, the database from the Dutch Centre for Lexical Information
	Age of Acquisition	the average age of acquisition norms for content words based on MRC
	Familiarity	the average familiarity ratings for content words based on MRC
	Concreteness	the average concreteness ratings for content words in a text based on MRC
	Imageability	the average imageability (i.e., how easy it is to construct a mental image) ratings for content words in a text based on MRC
	Meaningfulness	the average meaningfulness (i.e., the extent to which a word is associated with other words) for content words in a text based on MRC
	Polysemy	the number of senses a word has computed by WordNet

© Shiraz University

Sub-component	Index	Description
	Hypernymy	a normalized scale within 0 and 1 reflecting an overall use of less/more specific nouns and verbs computed by WordNet

As the initial data analysis step, the scales computed by Coh-Metrix for each of the lexical features under investigation were used to estimate several descriptive statistics. Following the descriptive analysis of the data, the lexical richness indices that significantly differentiated the two sub-corpora were explored, conducting the first-step process (significance testing of discriminant functions) of a DFA. The multicollinearity between the indices was initially assessed to avoid wasting the power of the potential model,. In testing for multicollinearity, it was tried to assure that the correlation value for every pair of the lexical indices is lower than .70 ($r < .70$) and the variance inflation factor (VIF) values (VIF) fall between zero and 10 (see the Appendix).

Results

A descriptive analysis was performed to address the first research question, which explored the lexical richness similarities and differences between EAP texts written by TEFL students and those written by native speakers of English. Table 2 shows the descriptive statistics of the lexical features in terms of the three sub-components under investigation.

© Shiraz University

Table 2.

Descriptive Statistics of the Features Representing Lexical Richness in the L1 and L2 Corpora

Sub-component	Index	Corpus	Min.	Max.	Mean	*SD*
Diversity	TTR	L1	.397	.960	.717	.113
		L2	.139	.985	.671	.146
Density	Noun Incidence	L1	173.912	516.129	287.023	41.101
		L2	187.500	488.637	309.005	40.949
	Verb Incidence	L1	90.667	239.129	129.967	24.628
		L2	61.539	196.971	120.821	23.541
	Adjective Incidence	L1	.000	177.777	84.421	24.171
		L2	28.986	192.938	101.706	28.220
	Adverb Incidence	L1	.000	103.448	49.418	18.138
		L2	.000	187.500	39.385	23.448
Sophistication	CELEX Word Frequency	L1	1.788	2.551	2.135	.139
		L2	1.429	2.392	2.017	.127
	Age of Acquisition	L1	300.400	461.750	388.731	28.207
		L2	326.000	536.000	405.008	31.918
	Familiarity	L1	549.733	588.627	569.656	7.725
		L2	521.154	580.857	561.982	8.735
	Concreteness	L1	317.259	432.333	364.036	19.062
		L2	242.667	432.000	351.906	23.833
	Imageability	L1	356.793	448.188	396.106	17.750
		L2	320.333	451.667	380.282	20.987
	Meaningfulness	L1	384.444	492.571	430.433	16.512
		L2	325.000	487.333	417.132	25.327
	Polysemy	L1	2.532	4.825	3.613	.382
		L2	2.581	4.964	3.535	.358
	Hypernymy	L1	1.551	2.706	2.019	.218
		L2	1.580	3.352	2.157	.292

© Shiraz University

© Shiraz University

As the statistics (see Table 2) estimated for the only lexical diversity index (i.e., TTR) depict, the L1 corpus (M = .717, SD = .113) enjoyed a higher degree of diversity in comparison with the L2 one (M = .671, SD = .146). As to the lexical density results, the average incidence scores in the L2 corpus were higher in terms of nouns (L1: M = 287.023, SD = 41.101; L2: M = 309.005, SD = 40.949) and adjectives (L1: M = 84.421, SD = 24.171; L2: M = 101.706, SD = 28.220), whereas the average incidence scores for adverbs and verbs in the L1 corpus exceeded (verb: M = 129.967, SD = 24.628; adverb: M = 49.418, SD = 18.138) the corresponding scores in the L2 one (verb: M = 120.821, SD = 23.541; adverb: M = 39.385, SD = 23.448).

Concerning the lexical sophistication indices, the comparative results shown in Table 2 demonstrated higher degrees of CELEX word frequency (L1: M = 2.135, SD = .139; L2: M = 2.017, SD = .127), familiarity (L1: M = 569.656, SD = 7.725; L2: M = 561.982, SD = 8.735), concreteness (L1: M = 364.036, SD = 19.062; L2: M = 351.906, SD = 23.833), imageability (L1: M = 396.106, SD = 17.750; L2: M = 380.282, SD = 20.987), meaningfulness (L1: M = 430.433, SD = 16.512; L2: M = 417.132, SD = 25.327), and polysemy (L1: M = 3.613, SD = .382; L2: M = 3.535, SD = .358) in the L1 corpus. The only two lexical sophistication indices were found to be, on average, higher in the L2 corpus included age of acquisition (L1: M = 2.109, SD = .218; L2: M = 2.157, SD = .292) and hypernymy (L1: M = 388.731, SD = 28.207; L2: M = 405.008, SD = 31.918).

To find a clear answer to the second research question, which focused on the lexical features differentiating between EAP texts written by Iranian TEFL students and those written by their native English-speaking counterparts, a DFA was conducted. Before conducting the DFA, however, the preliminary

assumptions (i.e., non-multicollinearity, no-outliers, homogeneity of variance/covariance matrix, and normality of the lexical indices included in the model) were analyzed in detail. Based on the results, a model including 10 (out of 13) of the initially selected indices met the broad range of assumptions underlying a DFA model (see the Appendix). Table 3 shows the results of the univariate analysis of variance (ANOVA) on the ten indices included in the DFA model.

Table 3.
Tests of Equality of Group Means in terms of the Lexical Indices Included in the DFA Model

Index	Wilks' Lambda	*F*	*df1*	*df2*	Sig.
TTR	.973	7.821	1	283	.006
Verb Incidence	.967	9.601	1	283	.002
Adjective Incidence	.912	27.298	1	283	.000
Adverb Incidence	.953	14.081	1	283	.000
CELEX Word Frequency	.841	53.435	1	283	.000
Age of Acquisition	.938	18.572	1	283	.000
Familiarity	.837	55.102	1	283	.000
Concreteness	.935	19.581	1	283	.000
Polysemy	.989	3.023	1	283	.083
Hypernymy	.941	17.761	1	283	.000

As shown in Table 3, with the exclusion of polysemy (*Wilk's* Λ = .989, $F(1, 283) = 3.023$, $p > .05$), the average values of the other indices differentiated between the L1 and L2 corpora. The eigenvalue (.472) and the canonical correlation (.566) estimated based on the DFA model (see Table 4), however, were found to be moderate (1.00 is perfect), indicating a moderately

© Shiraz University

strong function on the basis of the independent variables that acceptably discriminate between the two sub-corpora.

Table 4.
Eigenvalue and Canonical Correlation Estimated based on the DFA Model

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.472[a]	100.0	100.0	.566

Table 5 provides the relative importance of the lexical indices included in the DFA model on the basis of the correlations of each variable with each discriminate function, known as structure coefficients or discriminant loadings. According to the results, familiarity (r = .642), CELEX word frequency (r = .632), and adjective incidence (r = .452) were the most important predictors suggesting a label of lexical richness as the function that discriminates between the L1 and L2 EAP texts, whereas polysemy (r = .150), TTR (r = .242), and verb incidence (r = .268) were found to hardly be capable of discriminating between the two sub-corpora.

Table 5.
The Structure Matrix

Index	Structure Coefficients
Familiarity	.642
CELEX Word Frequency	.632
Adjective Incidence	-.452
Concreteness	.383
Age of Acquisition	-.373
Hypernymy	-.365

© Shiraz University

Index	Structure Coefficients
Adverb Incidence	.325
Verb Incidence	.268
TTR	.242
Polysemy	.150

Discussion

As its main objective, the present research study sought to explore the lexical differences and/or similarities between academic texts produced by Iranian Ph.D. level TEFL students and those written by their native English-speaking counterparts. To this end, the L1 and L2 corpora of the study were scrutinized in terms of a total of 13 indices representing lexical diversity, density, and sophistication by means of a computational tool, namely Coh-Metrix. The interval scales evaluated with respect to every particular lexical index were then analyzed descriptively. The comparative analysis of the descriptive statistics revealed that the L1 corpus enjoyed higher degrees of TTR, CELEX word frequency, familiarity, concreteness, imageability, meaningfulness, and polysemy compared to the L2 one. In other terms, the texts written by native speakers of English at Ph.D. level, on average, included higher proportions of unique, high frequency, familiar (easy to process), concrete (non-abstract), imageable (easy to image), meaningful (associated with other words used in the text), and multi-sense (ambiguous) content words. On the other hand, the L2 corpus enjoyed higher levels of the age of acquisition and hypernymy, denoting the use of more specific (content) words learned later by children.

The micro-findings enumerated above revealed that the texts written by Ph.D. level Iranian trainee teachers enjoyed a partially low level of diversity

© Shiraz University

© Shiraz University

but a high level of sophistication, as demonstrated by lower degrees of lexical uniqueness, meaningfulness, concreteness, learnability, imageability, and familiarity. The more diverse lexical construct in the L2 texts written by Iranian TEFL learners, compared with that of the L1 baseline, is totally in line with the empirically-approved (e.g., Breeze, 2008; Djiwandono, 2016; Grant & Ginther, 2000; Jarvis, 2002; Kwon, 2009) notion that lexical diversity correlates positively with the writer's level of lexical proficiency. To realize the justification for making such a widely-approved claim, one can refer to the contention made by Breeze (2008) that "Whereas good writers make an effort to find synonyms rather than repeat the same words, less proficient writers tend to be satisfied when communication is achieved, and are less concerned with questions of style" (p. 55). Admitting the self-evident idea that L2 learners have lower lexical knowledge compared to their native counterparts who experience automatic language acquisition in a natural environment, it seems quite reasonable that the L1 corpus produced by more proficient users of English enjoyed higher lexical diversity in comparison with those of the TEFL students.

The lower lexical diversity in the academic-genre texts written by TEFL students from an Iranian academic context may lend complementary support to the finding of Kwon's (2009) study whereby the lexical construct of a corpus including L2 academic texts written by intermediate and advanced South Korean university students was found to be less diverse than that of an L1 comparison corpus written by native speakers of English. Nonetheless, the higher lexical diversity in the L1 corpus contradicted the results drawn from the contrastive analysis by Crossley and McNamara (2011). Employing the same computational tool (Coh-Metrix) and lexical diversity index (TTR), Crossley and McNamara (2011) compared texts written by L1 speakers of

English and L2 texts written by high-intermediate and advanced writers from Czech, Finland, Germany, and Spain and came to a conclusion that the diversity of the L1 corpus was significantly lower than that of the L2 one, irrespective of the writers' first language background. Having called such finding 'a counter to expectations', Crossley and McNamara (2011) attributed the oddity to the disparity between the L1 and L2 texts in the stylistic and structural choices.

As a revealing finding, the average scales estimated for the majority of the lexical sophistication features were found to be higher in the L1 corpus. Accordingly, the lexical construct of the EAP texts written by Iranian TEFL students was more sophisticated than the benchmark of the study (the L1 Corpus). This result bore a remarkable similarity with the descriptive results revealed by Crossley and McNamara's (2011) linguistic analysis. In spite of the heterogeneous nature of the L2 corpus explored by Crossley & McNamara (2011) in terms of the L2 writers' nationality, almost all of the lexical sophistication features (i.e., meaningfulness, hypernymy, polysemy, imageability, and familiarity) were found to be higher in the L1 corpus. As the only matter of difference, contrary to what has been found in the study by Crossley and McNamara (2011), the L2 corpus explored in the current study contained more specific (less generalizable) words compared to the L1 one.

Notwithstanding the similarities on a descriptive level, the current study and that of Crossley and McNamara (2011) did diverge in interpreting the overall lexical sophistication level of the L2 texts. While the L2 corpus of the current study was found to be generally more sophisticated in comparison with the texts written by English natives, Crossley and McNamara (2011)

© Shiraz University

came to the conclusion that L2 writers produce words that are less sophisticated. This disparity in interpretation stems from the differences in the significance of testing results. While polysemy and hypernymy were regarded as the distinguishing indices in Crossley and McNamara's (2011) study, familiarity and CELEX word frequency were found to be the best predictors of lexical sophistication in the current study. Relying upon dissimilar lexical sophistication features, the two studies achieved different interpretative results.

The higher levels of lexical sophistication in the writing style favored by Iranian TEFL students seem in direct contradiction with the non-automated process of word retrieval that, as a salient feature of L2 writing (Schoonen, Snellings, Stevenson, & van Gelderen, 2009), could potentially yield a low incidence of using specific low-frequency words (Crossley & McNamara, 2011; Clark, 1978). A possible explanation for such a revealing finding may be the lengthy process of dissertation development. Although no one denies the L1 writers' access to an automatically retrievable lexical database as a result of language learning in a natural setting (Chenoweth & Hayes, 2001), it needs to be noted that such lexical ascendancy is more likely to draw a distinction between L1 and L2 writers while writing in a timely fashion (i.e., writing under time pressure). Having ample time and opportunity to go through an iterative process of lexical choices optimization by virtue of widely accessible databases and dictionaries, the L2 writers of the study were very likely to gain access to a lexical repertoire that either approximates or outclassed the naturally-occurred lexical organization possessed by the L1 ones.

© Shiraz University

© Shiraz University

Another important explanation for the sophisticated nature of the L2 corpus may be the specific genre (academic) explored in the current study. Many L2 writers involved in academic writing consciously embark on following the lexical and syntactic construct of well-structured academic texts to improve writing quality. Such structural and lexical benchmarking may have resulted in the emergence of a sophisticated written style in the Iranian postgraduate academic discourse. The superiority of the L2 texts over the L1 ones in lexical sophistication is in harmony with the empirical data (Failasofah & Alkhrisheh, 2018; Juanggo, 2018), suggesting a negative correlation between English proficiency and the use of less-frequency words (lexical sophistication) by EFL learners in the Indonesian EFL context.

Given the results relevant to lexical density, in comparison with the L1 corpus, the L2 texts included a higher proportion of nouns and adjectives but a smaller proportion of verbs and adverbs. This finding is hardly comparable with the previously-drawn empirical data since most of the previous studies on lexical density (e.g., Hinkel, 2011; Kwon, 2009) adopted a holistic view to measuring lexical density which entails estimating the proportion of the content words in a text, irrespective of their type. Nonetheless, the heavier use of nouns and adjectives in the L2 texts, which is indicative of L2 learners' tendency to frequently use nouns/noun phrases in conjunction with adjectival phrases, has been previously validated by Schleppegrell and Go's (2007) study. Such tendency may be regarded as a deliberate attempt to compensate for the lack of explicitness and clarity, widely recognized (e.g., Hinkel, 2005; Silva, 1993) as the salient drawback of L2 texts (Silva, 1993).

© Shiraz University

As another area of inquiry, the study aimed to ascertain the lexical features (qualities) that significantly differentiate between EAP texts written by Iranian TEFL students and those written by their native counterparts. To this end, a DFA model including 10 (out of the 13) lexical richness indices was fit on the data computed by Coh-Metrix. As shown by the DFA results, with the exclusion of polysemy, the other lexical sophistication indices included in the DFA model (i.e., familiarity, CELEX word frequency, concreteness, age of acquisition, and hypernymy) were found to be capable of discriminating between the two sub-corpora. Aside from the lexical sophistication indices enumerated above, adjective and adverb incidence scores representing lexical density were found to be successful in predicting the differences between the two sub-corpora. Consequently, lexical sophistication and lexical density were found to be the best lexical richness predictors whereby one can distinguish whether or not a text is written by Iranian Ph.D. level TEFL students. Among these predictors, familiarity, CELEX word frequency, and adjective incidence were found to be the most successful ones.

The heavier use of familiar high-frequency words in the L1 texts, as revealed by significantly higher average values estimated based on CELEX word frequency and familiarity indices, was consistent with the findings of Kwon's (2009) study, reflecting greater proportion of high-frequency words in an L1 corpus written by native speakers of English in comparison with the L2 essays written by Korean university students as well as L2 sample essays written in a TOEFL written test. This finding accounts for the conclusion made by Kwon (2009) that "good writing may not necessarily require exceptionally difficult or sophisticated words" (p. 169).

© Shiraz University

Relying upon D value -as the only measure of lexical diversity- and the proportion of content words in a text -as a single descriptor of lexical density- and five band-based measures of lexical sophistication, the findings of Kwon's (2009) study revealed that lexical sophistication and diversity are much more successful in predicting how the lexical richness differs between L1 and L2 texts. Although lexical sophistication was found to be the common ground between the current study and the study by Kwon (2009), the discrepancy between the two studies in terms of lexical density may be attributed to the disparities in the approaches (holistic vs. specific) adopted to measure lexical density.

To sum up, unlike writing tasks intended to be accomplished in a timely fashion, which compel L2 learners to consent to the least text development requirements (i.e., accuracy and comprehensiveness), pick familiar high-frequency words immediately accessible to them, and avoid the attendant risks of producing a sophisticated lexical construct (Hasselgren, 1994), extended writing tasks provide writers with ample time and room for the heavy use of advanced/sophisticated words chosen from either academically-approved well-structured exemplars or specific/general scope dictionaries/databases. This could yield a sophisticated lexical construct in L2 discourse, as witnessed in Iranian TEFL students' dissertations. Nonetheless, the laborious, time-consuming, and temporal nature of such a word retrieval system, in comparison with the permanent immediately-accessible retrieval system possessed by L1 writers, could potentially hinder unrestricted access to a variety of unique words. Accordingly, the advanced words chosen laboriously from either dictionaries or texts regarded as the epitome of academic writing are very likely to be used repeatedly by L2 writers. This may account for the

disequilibrium between lexical sophistication and diversity in the L2 corpus. The asymmetrical pattern of lexical density in the two sub-corpora may also be rooted in the different retrieval systems the two groups of writers were accustomed to.

Conclusion

The concluding remarks made by the findings suggested a highly sophisticated, normally diverse, and asymmetrically dense lexical construct in the academic writing of Iranian TEFL students. In agreement with the result of many other studies, the findings may be the proof of a focal emphasis placed by L2/FL learners on the use of advanced/sophisticated (low-frequency) words enjoying lower levels of meaningfulness, concreteness, and familiarity to satisfy the lexical richness requirements of academic writing. In parallel with the findings of the previous studies on L2 discourse, the current study's findings may bring about a change in the widely-held view that a lexically sophisticated text necessarily reflects the high lexical proficiency of its writer.

To develop a lexically rich academic discourse, considerations for maintaining sophistication need to be accompanied by a proper regard for lexical diversity and balanced use of lexical items of different types. The findings could also directly impact the local pedagogy aimed at enhancing academic writing quality among Iranian postgraduates, specifically those majoring in TEFL. With a detailed understanding of the lexical properties that dominate a specific academic community's written discourse, more workable techniques are likely to be proposed to enrich the prevalent writing style. As the findings of the current study may partially be flawed by several limitations and delimitations such as non-random selection of the clusters, the limited size

© Shiraz University

of the two sub-corpora, and the specific type of EAP texts (doctoral dissertation), there is an apparent need to replicate the study using larger, more various (including master's theses, doctoral dissertations, and scientific papers), and randomly-selected main and comparison corpora. Such replication increases the generalizability of the findings and investigates the authenticity of the results obtained in the current study.

Declaration of Conflicting Interests

The author declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

© Shiraz University

References

Bickman, L., & Rog, D. J. (1998). *Handbook of applied social research methods*. London: Sage Publications.

Breeze, R. (2008). Researching simplicity and sophistication in student writing. *International Journal of English Studies, 8*(1), 51–66.

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, *26*(4), 42–65.

Chen, Y. H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology*, *14*(2), 30–49.

Chenoweth, N. A., & Hayes, J. R. (2001). Fluency in writing. Generating text in L1 and L2. *Written Communication*, *18*(1), 80-98.

Clark, E. V. (1978). Discovering what words can do. In D. Farkas, W. M. Jacobsen, & K. W. Todrys (Eds.), *Papers from the parasession on the lexicon* (pp. 34–57). Chicago: Chicago Linguistic Society.

Cobb, T., & Horst, M. (2004). Is there room for an AWL in French? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 15-38). Amsterdam: John Benjamins Publishing.

Crossley, S. A. & McNamara, D. S. (2009). Computationally assessing lexical differences in second language writing. *Journal of Second Language Writing*, *17*(2), 119–135.

Crossley, S. A., & Kyle, K. (2018). Assessing writing with the tool for the automatic analysis of lexical sophistication (TAALES). *Assessing Writing*, *38*, 46–50.

Crossley, S. A., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life Long Learning*, *21*(3), 170–191.

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, *35*(2), 115–135.

Crossley, S. A., Salsbury, T., McCarthy, P. M., & McNamara, D. S. (2008). Using latent semantic analysis to explore second language lexical development. In D. Wilson & G. Sutcliffe (Eds.), *Proceedings of the 21st international Florida artificial intelligence research society* (pp.136–141). Menlo Park, California: AAAI Press.

Crossley, S., Weston, J., McLain Sullivan, S., & McNamara, D. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication, 28*, 282–311.

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

Djiwandono, P. (2016). Lexical richness in academic papers: A comparison between students' and lecturers' essays. *Indonesian Journal of Applied Linguistics*, *5*(2), 209–216.

Douglas, S. R. (2010). *Non-native English speaking students at university: Lexical richness and academic success* (Unpublished doctoral thesis). University of Calgary, Calgary, AB.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4*, 138–155.

Douglas, S. R. (2013). The lexical breadth of undergraduate novice level writing competency. *The Canadian Journal of Applied Linguistics*, *16*(1), 152–170.

Failasofah, F., & Alkhrisheh, H. T. D. (2018). Measuring Indonesian students' lexical diversity and lexical sophistication. *Indonesian Research Journal in Education*, *2*(2), 97–107.

© Shiraz University

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, *9*, 123–145.

Gregori-Signes, C., & Clavel-Arroitia, B. (2015). Analyzing lexical density and lexical diversity in university students' written discourse. *Procedia-Social and Behavioral Sciences*, *198*, 546–556.

Grobe, C. (1981). Syntactic maturity, mechanics, and vocabulary as predictors of quality ratings. *Research in the Teaching of English*, *15*(1), 75–85.

Ha. H. S. (2019). Lexical richness in EFL undergraduate students' academic writing. *English Teaching*, *74*(3), 3–28.

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics*, *4*, 237–258.

Higginbotham, G., & Reid, J. (2019). The lexical sophistication of second language learners' academic essays. *Journal of English for Academic Purposes*, *35*, 127-140.

Hinkel, E. (2005). Analyses of L2 text and what can be learned from them. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 615–628). Mahwah, NJ: Lawrence Erlbaum Associates.

Hinkel, E. (2011). What research on second language writing tells us and what it doesn't. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 523-538). New York: Routledge.

Hirvela, A. (2011). Writing to learn in content areas: Research insights. In R. M. Manchón (Ed.). *Learning-to-write and writing-to-learn in an additional language* (pp. 37-59). Amsterdam: John Benjamins.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, *19*, 57–84.

Johansson, V. (2009). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working Papers in Linguistics*, *53*, 61–79.

Juanggo, W. (2018). Investigating lexical diversity and lexical sophistication of productive vocabulary in the written discourse of Indonesian EFL learners. *Indonesian Journal of Applied Linguistics*, *8*(1), 38–48.

© Shiraz University

Jurafsky, D., & Martin, J. (2002). *Speech and language processing*. New Delhi: Pearson Education Inc.

Kalantari, R., & Gholami, J. (2017). Lexical complexity development from dynamic systems theory perspective: Lexical density, diversity, and sophistication. *International Journal of Instruction*, *10*(4), 1–18.

Kim, S., & Jeon, M. (2016). An analysis study of English writing of elementary school 6th grade English languagelearners using Coh-Metrix. *Modern English Education*, *17*(3), 263–287.

Kojima, M., & Yamashita, J. (2014). Reliability of lexical richness measures based on word lists in short second language productions. *System*, *42*, 23–33.

Kusumaningrum, M. V., & Ardi, P. (2020). A corpus analysis of lexical sophistication in LLT journal: A journal on language and language teaching. *ELTR Journal*, *4*(1), 53–75.

Kwon, S. (2009). Lexical richness in L2 writing: How much vocabulary do L2 learners need to use? *English Teaching*, *64*(3), 155–174.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*, 757–786.

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

Lavallée, M., & McDonough, K. (2015). Comparing the lexical features of EAP students' essays by Prompt and Rating. *Revue TESL du Canada*, *32*(2), 30–44.

Malvern, D., Richards, J. B., Chipere, N., & and Durán. P. (2004). *Lexical richness and language development: Quantification and assessment.* Houndmills, Basingstoke: Palgrave MacMillan.

Meara, P. (2005). Lexical frequency profiles: A Monte Carlo analysis. *Applied Linguistics*, *2*(1), 32–47.

Meara, P., & Bell, H. (2001). P Lex: a simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect*, *16*(3), 5–19.

Morris, L., & Cobb, T. (2004). Vocabulary profiles as predictors of the academic performance of teaching English as a second language trainees. *System*, 32, 75–87.

Muncie, J. (2002). Process writing and vocabulary development: comparing lexical frequent profiles across drafts. *System*, *30*, 225–235.

Nation, I. S. P., & Meara, P. (2010). Vocabulary. In N. Schmitt (Ed.), *An introduction to applied linguistics* (pp. 34 –52). London, Hodder Education.

© Shiraz University

O'Loughlin, K. (1995). Lexical density in candidate output on direct and semi-indirect versions of an oral proficiency test. *Language Testing 12*(2), 217–237.

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

Sasaki, M. (2007). Effects of study-abroad experience on EFL writers: A multiple-data analysis. *The Modern Language Journal*, *91* (4), 602-620.

Schleppegrell, M. J., & Go, A. L. (2007). Analyzing the writing of English learners: A functional approach. *Language Arts, 84*(6), 529−538.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Basingstoke: Palgrave Macmillan.

Schoonen, R., Snellings, P., Stevenson, M., & Van Gelderen, A. (2009). Towards a blueprint of the foreign language writer: The linguistic and cognitive demands of foreign language writing. In R. M. Manchón (Ed.), *Writing in foreign language contexts. Learning, teaching and research* (pp. 77−101). Bristol, UK: Multilingual Matters.

Silva, T. (1993). Toward an understanding of the distinct nature of second language writing: The ESL research and its implications. *TESOL Quarterly*, *27*, 657–677.

Šišková, Z. (2012). Lexical richness in EFL students' narratives. *Language Studies Working Papers*, *4*, 26–36.

Staehr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, *36*(2), 139–152.

Storch, N., & Tapper, J. (2009). The impact of an EAP course on postgraduate writing. *Journal of English for Academic Purposes*, *8*(3), 207–223.

Vermeer, A. (2004). The relation between lexical richness and vocabulary size in Dutch L1 and L2 children. In P. Boogards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 173 –189). Amsterdam: Benjamins.

Vongpumivitch, V., Huang, J., & Chang, Y. (2009). Frequency analysis of the words in the academic word list (AWL) and non-AWL content words in applied linguistics research papers. *English for Specific Purposes*, *28*, 33 –41.

Zheng, Y. (2016). The complex, dynamic development of L2 lexical use: A longitudinal study on Chinese learners of English. *System*, *56*, 40–53.

© Shiraz University

© Shiraz University

Appendix

Assumptions Checked before Running the DFA Model

Table A1
Results Relevant to the Normality of the Predictors (Lexical Indices)

Index	Kolmogorov-Smirnov[a]		
	Statistic	*df*	*Sig.*
TTR	.043	285	.200
Verb Incidence (VI)	.041	285	.200
Adjective Incidence (AdjI)	.048	285	.092
Adverb Incidence (AdvI)	.052	285	.056
CELEX Word Frequency (CELEX)	.036	285	.200*
Age of Acquisition (AOA)	.050	285	.086
Familiarity (Fam.)	.040	285	.200*
Concreteness (Con.)	.051	285	.070
Polysemy (Pol.)	.044	285	.200*
Hypernymy (Hyp.)	.050	285	.080

Table A2
Results Relevant to the Homogeneity of the variance/Covariance Matrix

Box's M		163.745
F	Approx.	2.854
	df1	55
	df2	150053.362
	Sig.	.004

Table A3
Results Relevant to the Multicolliniarity between the Predictors (Lexical Indices)

Index		TTR	VI	AdjI	AdvI	CELEX	AOA	Fam.	Con	Pol.	Hyp.
Correlation	TTR	1.000	.028	-.173	.015	.117	.008	.085	-.024	.189	.064
	VI	.028	1.000	-.153	.005	.188	-.256	.327	.137	.313	-.037
	AdjI	-.173	-.153	1.000	.033	-.215	.157	-.253	-.151	-.062	-.138
	AdvI	.015	.005	.033	1.000	.316	.046	.039	-.168	.168	-.437
	CELEX	.117	.188	-.215	.316	1.000	-.149	.584	-.161	.582	-.434
	AOA	.008	-.256	.157	.046	-.149	1.000	-.401	-.429	-.107	-.170
	Fam.	.085	.327	-.253	.039	.584	-.401	1.000	.058	.423	-.071
	Con.	-.024	.137	-.151	-.168	-.161	-.429	.058	1.000	-.141	.347
	Pol.	.189	.313	-.062	.168	.582	.107	.423	-.141	1.000	-.186
	Hyp.	.064	-.037	-.138	-.437	-.434	-.170	-.071	.347	-.186	1.000

Table A4
VIF values Relevant to the Predictors

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	198.978	122.881		1.619	.107		
TTR	-14.796	10.001	-.048	-1.479	.140	.913	1.096
VI	-.404	.061	-.232	-6.610	.000	.774	1.291
AdjI	-.210	.053	-.139	-3.943	.000	.767	1.305
AdvI	-.361	.068	-.190	-5.296	.000	.746	1.341
CELEX	-17.887	16.205	-.061	-1.104	.271	.318	3.145
AOA	.038	.053	.029	.729	.467	.619	1.617
Fam.	.021	.216	.005	.099	.921	.437	2.291
Con.	.163	.068	.088	2.383	.018	.697	1.435
Pol.	-11.887	4.659	-.104	-2.551	.011	.581	1.722
Hyp.	91.041	6.279	.594	14.498	.000	.570	1.753