ATU PRESS

ISSUES IN LANGUAGE TEACHING (ILT)
VOL. 12, NO. 1, 271-294, June 2023
https://doi.org/10.22054/ilt.2023.67794.696 Document Type: Research Paper

Lexical Sophistication in the Discussion Section of MA Theses Authored by Iranian EFL vs. English Students: A Coh-Metrix Report on Similarities and Differences

Masoud Azadnia
Assistant Professor of TEFL,
Islamic Azad University, Isfahan Branch, Isfahan, Iran

Received: October 21, 2022**; Accepted:** October 19, 2023

Abstract
This study compared the Discussion section of theses written by M.A. English L1 and L2 students regarding lexical sophistication. Certain linguistic features were applied to investigate whether texts written in English L2 had any similarities or differences with native speakers' (NSs) texts. To achieve this, 20 English L2 theses authored by Iranian M.A. students of the Islamic Azad University of Isfahan (Khorasgan), were sampled. As such, 20 English L1 theses written by M.A. English NSs of the same major in the US and UK were randomly downloaded as a comparison corpus of English L1. The corpora were later uploaded to Coh-Matrix, a computational tool that processes text and discourse at different levels of language. Two main statistical procedures for data analysis were the MANOVA (Multivariate Analysis of Variance) and Discriminant Function Analysis. According to Coh-Metrix analysis, the results revealed certain similarities and differences between the corpora. Specifically, more CELEX content words were used in the NSs theses. However, the differences between the two corpora did not reach a statistical significance in terms of other indices of lexical sophistication (i.e., polysemy, concreteness, hypernymy, age-of-acquisition scores, and lexical diversity). Based on the findings, academic writing pedagogy and thesis writing abilities of Iranian English L2 learners can be improved by applying both word processing tools like Coh-Metrix and appropriate writing strategies and techniques.

Keywords: Coh-Metrix, lexical sophistication, L1 and L2 thesis writing, linguistic features

Author's email: masoudazadnia@gmail.com

INTRODUCTION

Over the last few decades, a large number of research studies based on learner-corpora have been conducted with the purpose of assessing the lexical characteristics of students' writing. There are a number of ways researchers can analyze text lexical sophistication like traditional hand counts, for instance. In particular, in the Persian EFL corpus and research, traces of traditional methods can be overtly seen, except for few recent studies like those of Nevisi and Hosseinpur (2019) and Ansarin et al. (2021). Nonetheless, significant advances in lexical analyses have been made possible by virtue of the developments of software programs and linguistic analysis tools, like Coh-Metrix (McNamara et al., 2014), TAALES (Kyle & Crossley, 2015; Kyle et al., 2018**)**, and VocabProfiler (Cobb, 2018). Investigating the lexical characteristics of student-produced texts and examining the way various lexical characteristics are connected to writing quality and language proficiency are the most important goals of these tools which have been applied to many studies (Crossley et al., 2011, 2014; Crossley & McNamara, 2012; Durrant et al., 2019; Guo et al., 2013; Karafkan & Hadidi, 2021; Kim & Crossley, 2018; Kyle & Crossley, 2016; McNamara et al., 2010; Vögelin et al., 2019; Yu, 2010). Taking the result of these studies into account, one can realize how important students' writing ability, lexical analysis, and vocabulary measures are in order to reflect their writing proficiency. As such, it can address how one's vocabulary domain affects the writing quality (Maamuujav, 2021).

Thesis writing analysis by automated tools is worth addressing insofar as EFL post graduates' writing ability is concerned. It, sure, outpaces traditional text investigation like solely word frequency hand counts as it spans and uncovers a number of other vital lexical features herein. Chances are certain identifiable text characteristics and/or difficulties can be tackled and then

academically taught a priori to enhance post-graduate thesis writing potential. To this end, the present study is attempting to bring in one of the powerful automated text analysis tools, CohMetrix, to the context of EFL thesis writing by comparing it with that of native speakers to detect similarities and differences in lexical characteristics hoping that this might come to Iranian students' assistance when they are about to begin writing their thesis.

LITERATURE REVIEW

Ever since the 1950s, it has been said that grammatical and lexical errors analysis are derived from contrastive/error analysis covering L2 learning research. Essentially, transferring syntactic and lexical patterns and language properties from L1 to L2 is assumed to be the main reason of many L2 errors (Keshavarz, 2007). As well, grammatical sophistication acts as an indicator of writing quality by contributing to the production and comprehension of writing (Ebrahimi & Imandar, 2021).

Recently the literature of contrastive corpus analysis has shown works on some areas like lexical bundles and the role they might have on academic research writing among native English L1 and English L2 writers (Esfandiari & Barbary, 2017). In another contrastive corpus study, lexical bundles were investigated in dissertation abstracts authored by Chinese and L1 English doctoral students (Xiaofei & Jinlei, 2019). Regarding lexical sophistication analysis and employing computer-mediated tools like Tool for the Automatic Analysis of Lexical Sophistication (TAALES) 2.2, written and oral L2 productions of Spanish and Japanese students have been assessed (Clavel & Speck, 2021).

Lexical and/or morphological sophistication, as one of the aspects of linguistic complexity (Esfandiari & Jafari, 2021) and as an important signifier of overall writing potential, has been defined as the frequency of reference-corpus of words within a text (Coh-

Metrix, 2012). In other words, if a word occurs less frequently, it is labeled sophisticated whereas more frequent words are regarded less so. According to Coh-Metrix team, when higher number of words are used a couple of instances throughout the text, lexical sophistication becomes lower (and, therefore, cohesion is probable to be higher). It should be pointed out that two important classes in containing different characteristics of lexical sophistication are: lexical diversity and word information based on Medical Research Council Psycholinguistic Database.

Coh-Metrix

At different discourse and language analysis levels, Coh-Metrix as an automated tool is capable of evaluating text difficulty and cohesion measures. Improvements in various disciplines as reported in Graesser et al. (2004), have paved the way to computationally measure various texts and languages, explore them more deeply, and discover more global attributes of language. Corpus linguistics, computational linguistics, information retrieval, information extraction, psycholinguistics and discourse processing are different disciplines and approaches that made this possible.

Drawn together, automating the analysis of many in-depth linguistic features of language and textual coherence, more precise and detailed analyses of languages have been made possible due to the above areas' improvements. Integration of the developments in these areas have occurred by means of using Coh-Metrix. Lexicons, taggers of part of speech, pattern classifiers, syntactic analyzers (parsers), and other components developed in computational linguistics are integrated into this system (Jurafsky & Martin, 2002).

Graesser et al. (2004) analyzed several lexical indices like word meaningfulness, polysemy, age-of-acquisition scores, word frequency, concreteness, hypernymy, word familiarity measures and word imageability by Coh-Metrix, which shares some identicalities

with the present study. Also, in English L1 studies (e.g., Louwerse, McCarthy, McNamara & Graesser, 2004; McCarthy, Lewis, Dufty & McNamara, 2006) distinguishing types of texts, investigating the texts in terms of linguistic structure, and examining textual constructs have been accomplished by applying Coh-Metrix. Moreover, regarding lexical indices (e.g., Crossley, Greenfield, & McNamara, 2008, Crossley & McNamara, 2011), several validation studies were performed by Coh-Metrix and its measures.

In summary, Coh-Metrix has made it possible to evaluate the role of linguistic characteristics and writing quality by expert readers in many recent studies. Generally, argumentative essays (and not dissertations or theses) written by English native and nonnative (NSs & NNSs) speakers of English have been examined through these investigations. In many other studies that distinguish text types, Coh-Metrix indices of lexical sophistication have been validated (e.g., Crossley et al., 2007; McCarthy et al., 2006, 2007). Therefore, there might be a great deal of confidence that the measures investigated can be reliably evaluated by Coh-Metrix indices. These indices were selected from 108 indices regarding different linguistic features of Coh-Metrix to better fulfill the goals of this study.

At this point, it should be pointed out that our research question is mainly seeking any similarities and/or differences within the existing corpora. Hypothetically, they can be pinpointed to reflect academic writing pedagogy and theses writing abilities of Iranian L2 learners of English. As such, incorporating web tools like Coh-Metrix can be anti-traditional to word hand-counts or other dated methods of text analysis.

METHOD

Corpus Selection Procedure

We chose Coh-Metrix as a computational, automated web tool for data analysis concerning lexical sophistication. Twenty L2 theses

files were randomly selected by a librarian at Islamic Azad University of Isfahan (Khorasgan) because Coh-Metrix had to be fed by the file of the theses. In other words, there was no bias or pre-determination of choice in theses selection and the researchers were given the files by the librarian. As such, twenty L1 theses written by M.A. English NSs of the same major in the US and UK were randomly downloaded and an English L1 comparison corpus was also collected. The researchers had to search and troll the universities' higher education websites (in English-speaking countries) to find downloadable versions of theses due to the limited number of theses soft copies available online. This, though, can be addressed as one of the research limitations and can, in turn, reflect the generalizability of the research results.

The thematic structure of the theses showed that they were divided into Chapters (Introduction, Literature Review, Methodology, etc.) and each Chapter into certain subheadings like Overview, Discussion, Research Implications, etc. Analyzing the cohesion of the Chapter as a unified text could be impossible due to this situation. Therefore, first, the Discussion and Conclusion sections (Chapter Five) of the M.A. theses were chosen since other Chapters showed evidence of plagiarism. Quite evidently, there were statements that students were unlikely to write on their own (especially Iranian students!), which could deviate research results. And secondly, treating each subheading as a single text, the researcher calculated the cohesion index and, then, calculated the average cohesion of all the texts in the Chapter. These brought the total number of 99 texts.

Later, the texts had to be formatted and cleaned. A clean corpus is the one which is as human-readable as possible. However, in our case, we could detect a few typos and errors since when the corpus passed from one computer to another, it is likely to develop a variety of weird things, such as weird Spanish letter, sets of

mathematical symbols, or maybe just a wingding or two. Each of these contaminants could significantly impair the credibility of the analysis. Thus, they had to be free from these contaminants. In addition, Text Pad was the software recommended by the Coh-Metrix team to convert texts into files with txt extension that can be read by Coh-Metrix.

As the basis for comparing the queried L2 texts, the lexical sophistication indices of 20 English L1 texts were used. Thus, by providing a starting point for comparing and contrasting the L2 theses, the L1 theses provided us with internal validity; and regarding the production of linguistic characteristics, it has made it possible to determine how different or similar the English L2 and L1 writers are. As with the methodology of Reid (1992) and Crossley and McNamara (2011), English L1 texts should not be viewed as an ideal but as the baseline for comparison.

Data Collection Procedure

The above indices, of all the lexical sophistication indices calculated by Coh-Metrix, were chosen to technically address the lexical sophistication of the texts. To not waste the potential model power, before the final variable selection, we evaluated the colinearity between them. When testing for colinearity, we confirmed that there were no pair of indices that correlated more than $r \geq 0.70$, and that each variable was verified against VIF and tolerance tests. To represent the two categories in terms of lexical sophistication, a total of seven indices were selected in this study.

Data Analysis

The researcher used two statistical tools to analyze the data and to see if there existed significant similarities and/or differences between the two corpora (L1 and L2) regarding lexical sophistication. In so doing, a Multivariate Analysis of Variance

(MANOVA) followed by a Discriminant Function Analysis (DFA), a statistical routine applied to many preceding analyses to examine the discriminating features in the corpus (e.g., Biber, 1993; Crossley & McNamara, 2009, 2011), were performed. In fact to determine the measures of lexical sophistication, DFA was used to distinguish texts authored by Iranian writers from texts written by English NSs.

RESULTS

First, the records were analyzed descriptively to capture the differences and similarities among the corpora regarding lexical sophistication. Table 1 below shows descriptive statistics of the seven matrices measuring lexical sophistication. Here, the unit of analysis has been calculated based upon the number of occurrences of each matrix, i.e. CELEX Word Frequency, Age of Acquisition, Familiarity, etc.

Table 1: Descriptive statistics of the lexical sophistication indices in the L1 and L2 texts

Variable	L1/ L2	Min	Max	Mean	Std. Deviation	Skewness	Kurtosis
CELEX Word Frequency	L2	1.42	2.39	2.02	.12	-.96	1.12
	L1	2.01	2.27	2.14	.073	.11	1.35
Age of Acquisition	L2	326	536	406.87	35.48	.61	1.43
	L1	371.82	447.75	413.24	19.79	-.62	1.96
Familiarity	L2	521.15	578.13	564.01	9.21	-1.32	1.64
	L1	559.85	580.34	570.32	5.46	-.21	1.18
Concreteness	L2	242.66	418.18	348.13	22.51	-.86	1.98
	L1	339.05	367.72	353.86	9.35	-.67	.17
Polysemy	L2	2.58	4.96	3.55	.31	.55	1.26
	L1	3.61	3.98	3.83	.13	-.30	-1.41
Hypernymy (Nouns & Verb)	L2	1.58	3.35	2.20	.30	1.15	1.67
	L1	1.74	2.06	1.94	.11	-.76	-.55
Lexical Diversity	L2	.46	1	.70	.11	.34	.00
	L1	.65	.91	.73	.07	1.30	1.63

Compared to the L2 theses (M = 2.025), on average, the L1 texts recorded the use of a slightly higher CELEX word frequency which

meant higher average of content words in the theses (M = 2.143), as shown in Table 1. As mentioned earlier, spoken words that children will learn later are indicated by words with higher values of age of acquisition. Therefore, the L1 texts' average age of acquisition (M = 413,249) was higher than the average age of acquisition of L2 texts (M = 406,876). The extent to which a word appears familiar to an adult is all about the familiarity matrix. As represented in Table 1, in comparison with the L2 texts (M = 564.016), the average of this index was greater in the L1 corpus (M = 570.324).

In contrast to the L2 (M = 348.137), on average, the writers of L1 used more concrete or non-abstract words (M = 353.866), as shown in Table 1. The L1 writers, also, used more polysemy, that is, multi-senses words (M = 3.830) than the Iranian L2 writers of English (M = 3.554). In terms of hypernymy of nouns and verbs, unlike L1 writers (M = 1.943) and according to Table 1, generally, the Iranian L2 authors used more specific nouns and verbs (M = 2.209). Considering lexical diversity, in comparison with the L2 texts (M = 0.709), the overall lexical diversity was greater for the L1 texts (M = 0.738), which means that more identical words (i.e., tokens) were used multiple times in texts by the Iranian L2 writers compared to the NSs. For all variables in both groups (i.e., L1 and L2), the skewness and kurtosis values range from +2 to -2, indicating low resulting clustering and very low flatness.

Additionlly, we performed a one-way MANOVA to find out if there existed any differences significantly between the two corpora on the various indices in terms of lexical sophistication. Moreover, it was guaranteed that no pair of indices correlated above r> 0.70, indicating no multicollinearity based on the results of the Pearson correlation matrix (between dependent variables) which can be seen in Table 2 below. They are the results of checking underlying assumption to conduct MANONA for the dependent variables regarding lexical sophistication.

Table 2: Pearson correlation matrix for the lexical sophistication indices

Index		CELEX Word Frequency	Age of Acquisition	Fam.	Con.	Pol.	Hyp.	LD
CELEX Word Frequency	*r*	1	-.23**	.65**	-.04	.57**	-.36**	.17**
	Sig.		.00	.00	.48	.00	.00	.00
	N	285	285	285	285	285	285	285
Age of Acquisition	*r*		1	-.45**	-.46**	-.12*	-.11	-.03
	Sig.			.00	.00	.3	.05	.58
	N		285	285	285	285	285	285
Familiarity (Fam.)	*r*			1	.15**	.42**	-.14*	-.14*
	Sig.				.00	.00	.01	.01
	N			285	285	285	285	285
Concretenes s (Con.)	*r*				1	-.1	.26**	.01
	Sig.					.06	.00	.75
	N				285	285	285	285
Polysemy (Pol.)	*r*					1	-.24**	.2**
	Sig.						.00	.00
	N					285	285	285
Hypernymy (Hyp.)	*r*						1	.04
	Sig.							.47
	N						285	285
Lexical Diversity (LD)	*r*							1
	Sig.							
	N							285

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

Finally, using the Box's M Test for covariance equivalence and the Levene's Test for Uniformity of Variance, the uniformity of the variance-covariance matrices was checked. The results proved that no significant differences were seen between the covariance matrices and that the Box's M was not significant (Box's M = 57.653, $p > 0.001$). Therefore, it did not violate the basic assumptions of MANOVA. These two Tables can be seen below.

Table 3: Box's test of equality of covariance matrices for the lexical sophistication indices

Statistics	Value
Box's M	57.653
F	1.336
df1	28
df2	664.219
Sig.	.117

Table 4: Levene's Test of Equality of Error Variances for the Lexical Sophistication Indices

Index	F	df1	df2	Sig.
CELEX Word Frequency	1.964	1	106	.164
Age of Acquisition	2.960	1	106	.088
Familiarity	3.532	1	106	.063
Concreteness	2.791	1	106	.098
Polysemy	3.288	1	106	.073
Hypernymy	4.092	1	106	.053
Lexical Diversity	2.400	1	106	.124

Therefore, the dependent variables were the selected Coh-Metrix indices regarding lexical sophistication and independent variables were the written corpora by L1 and L2 writers. Table 5 below shows the results of the indices of one-way MANOVA.

Table 5: Multivariate tests results for the lexical sophistication indices

Statistic	Value	*F*	Hypothesis *df*	*Sig*.	Partial Eta Squared
Wilks' Lambda	.852	2.478	7.000	.022	.148

Here, it can be concluded that the test was significant using an alpha level of 0.05. Wilks' Lambda = 0.852, F (7,100) = 2.478, $p < .05$, multivariate $\eta^2 = .148$.

Table 6 below illustrates how the dependent variables (i.e., the linguistic features that appeared in lexical sophistication)

differed in L1/L2 texts as it shows the results of testing Between-Subjects Effects.

Table 6: Tests of between-subjects effects on the lexical sophistication indices

Variable	*Df*	*F*	*Sig.*	Partial Eta Squared
CELEX Word Frequency	1	7.749	.006	.068
Age of Acquisition	1	.281	.597	.003
Familiarity	1	4.068	.046	.037
Concreteness	1	.570	.452	.005
Polysemy	1	6.593	.012	.059
Hypernymy	1	6.531	.012	.058
Lexical Diversity	1	.578	.449	.005

The only significant difference between the L1 and L2 texts was found in CELEX word frequency F (1,106) = 7.749, $p < .007$, η^2 = 0.068, according to the following univariate ANOVAs. It is worth noting that to create an alpha level that describes the multiple ANOVAs performed, the Bonferroni correction (0.05 divided by the number of ANOVAs performed) was applied. Therefore, the researcher accepted statistical significance at $p < .007$ (.05/7).

Finally, to explore the possibility of the index that strongly distinguished between L1 and L2 academic writing samples, we performed Discriminant Function Analysis (DFA). Therefore, the linguistic index, namely the CELEX Word Frequency, which showed significant differences between the two corpora was chosen to be analyzed by the DFA model. According to the mean of this index across the L1 and L2 samples, L1 writers used more content words in academic text (M = 2.135) compared to L2 Iranian writers (M = 2.017).

DISCUSSION

There has been a plethora of research on text and discourse analysis employing different tools and aiming at different research purposes.

This report is worth noting, most emphatically, because of using Coh-Metrix which allows us to better find out about the potential linguistic contrasts and/or similarities between Iranian EFL and English writers at the higher education levels. Moreover, in terms of lexical sophistication, unlike more traditional studies using hand counts, it can successfully analyze, identify and distinguish linguistic features of L1 and L2 texts for students' writing output at any level or in any genres. Thus, this report is aimed at providing the English L2 writing community in Iran with a robust text analysis web apparatus (i.e., Coh-Metrix).

As mentioned earlier, in order to give a complete image of the degree of lexical sophistication in the corpora, of all the indices measured by Coh-Metrix output, seven representative indices (i.e., CELEX word frequency, age of acquisition, familiarity, concreteness, polysemy, hypernymy, and lexical diversity) were adopted.

According to the results, CELEX word frequency was the only index (among the seven indices) that caused a difference significantly between the two corpora. The English L1 texts had a significantly higher average content word usage in the texts than L2 texts did. The discovery of Hinkel (2011), who examined the textual features used in L2 writings of speakers of various languages (including Persian) is inconsistent with this finding. It was reported that L2 writers repeat content words (adverbs, verbs, nouns, adjectives,) more often; therefore, the content words frequency appeared higher in L2 texts.

The average age of acquisition of the L1 texts turned out to be higher than that of the L2 texts. Spoken words that children will learn later are indicated by words that have higher age of acquisition scores. As a result of first language acquisition in a natural environment, especially in relation to lexical retrieval where L1 writers can reliably use the amount and type of linguistic knowledge,

this finding seems rational (Chenoweth & Hayes, 2001). Because of cognitive nature of the Age of Acquisition index, as Crossley and McNamara (2011) point out in their study, this result is rightly expected. They explained that the age of acquisition measures represent the lexical entities like intuited order of lexical acquisition, word associations, the evocation of mental and sensory images, word abstractness, and spoken word exposure. NSs often encounter language input, resulting in a cognitive knowledge and words' mental frame, according to the description of this index by Crossley and McNamara (2011). However, because the academic genre has a specific lexical structure and L2 writers can improve their English skills through practice and exposure to academic writing, it may be predictable that as L2 writers improve their proficiency, the index may show higher levels of distribution in academic writing. Because children start to, first, use verbs that are general semantically like *come, make, do, go,* etc. in first language acquisition, the same lexical items are common. However, these common words, over time, are replaced with more advanced words by L1 learners, which are productive but less common (Clark, 1978). As time goes on in second language learning, similar movements to more advanced words also happen among L2 learners (Ellis & Ferreira-Junior, 2009). Maybe the fact that they have not established many hierarchical links amongst neighboring words has caused the generation of present result and as such L2 writers may have fewer specific words (Crossley, Salsbury, & McNamara, 2009).

According to the research analysis, the same conditions can be considered for the Familiarity index which was higher in L1 than in L2 texts. An assessment of the extent of familiarity of a word to an adult represents the familiarity index; accordingly, a native speaker's exposure to English will allow them to become more familiar with more words than Iranian English L2 learners. As mentioned earlier, looking at the index of Concreteness, English L1

writers, on average, used more concrete or non-abstract words compared to their L2 counterparts.

Polysemy showed a higher prevalence in L1 texts according to the findings. The results of Crossley and McNamara (2011) which reported the existence of significant differences amongst all L2 groups of their study (Spanish, German, Finnish, Czech) and the L1 group of word polysemy scores, show consistency with the findings of the present study. English L1 writers attempted to carry the intended meaning with the aid of using a single word in the appropriate context and they also tried to use multi-sense words than their English counterparts (Crossley & McNamara, 2011). However, the goal of L2 writers is to avoid ambiguity and make their production as simple as possible. In addition, it might be said that access to the same conceptual organization is not equal for these writers; rather, L1 writers seem to have more access. Therefore, per lexical entry, the L2 writers have fewer senses and are able to establish weaker links between senses while the conceptual organization of L1 writers depicts a lexical entry with most of all senses associated with a word (Crossley, Salsbury, & McNamara, 2010). Due to this difference and most probably because the meaning of those words and the strength of the links between the words are more beyond reach, English L2 writers seem to generate words with fewer senses. Taken together, it might be concluded that, compared to the L1 texts, the type or amount of vocabulary knowledge accessible to the L2 writers seem to push them inevitably into making less ambiguous (polysemy) texts.

Regarding the Hypernymy index, the theses writers of English L2 applied more specific nouns and verbs to their statements. This is incompatible with the view that the writers of L2, due to the lack of automatic lexical processing, use fewer specific words (hypernymy) that involve different general concepts (Clark, 1978). Also, the results of Crossley and McNamara's (2011) which

showed that the writers of L2 (Czech, Finnish, German, Spanish) used significantly fewer specific words than the writers of L1, are opposed to the results of current study.

Lexical Diversity which was the final index of lexical sophistication, was reported to be nearly greater in L1 texts, showing that Persian L2 writers have used more identical words (i.e., token) multiple times in the text. The traditional and general view that lexical diversity has been connected to lexical knowledge is consistent with this result. Only more experienced L2 writers have proved, by previous studies, to create texts with a high degree of lexical diversity (Engber, 1995; Grant & Ginther, 2000; Jarvis, 2002; Crossley & McNamara, 2011).

CONCUSION AND IMPLICATIONS

As the title of the paper suggests, current study is a report on similarities and differences throughout a corpus while promoting text analyzers' potentials, like Coh-Metrix, to the researchers in Iran. Therefore, it is preferable to avoid being highly conclusive about the results except for a touch, maybe, on some very general uptakes. It is said that the use of unique words in writing can be influenced by the genre of writing. For example, frequent use of unique words and special terms in academic genres such as the current research corpus are absolutely necessary and NSs' rich lexicon as well as their past experiences can play an important role.

According to Coh-Metrix analyses, regarding the lexical sophistication of the L1 and L2 texts, it obviously seems that having mastery of a word choice as well as learning English in a natural environment will increase the frequency of CELEX content words in English NSs' texts. That is, the English NSs' higher lexical knowledge caused a higher presentation of CELEX words in their texts compared to the Iranian L2 writers. Similar distribution

between L1 and L2 texts are approximately shared by other indices of lexical sophistication (i.e., familiarity, concreteness, polysemy, hypernymy, and lexical diversity). The main goal of practically all L2 text analyses and comparative studies, similar to the study of L2 discourse, stems from the educational needs to teach L2 writing to college students and professionals and academically-bound language learners.

It should be noted that for L2 pedagogy, the results of this study may have some preliminary implications. The distribution of using different linguistic features in L2 academic writing, leading to a satisfactory text can be influenced by the practice of writing strategies as well as dealing with academic genre, according. By applying writing strategies, most language aspects can be supported and drawn upon. The idea that lower-skilled students need to learn the writing strategies of more experienced students, or strategies that recompense deficiencies should be the basis of strategic instruction in language pedagogy.

For both understanding and learning, providing guidance and practice on how to apply strategies have been shown to be very beneficial (e.g., McNamara et al., 2006; Palincsar & Brown, 1984). Strategy instruction, for those who have little knowledge and poor reading skills and the students who have difficulty most of the times, is especially necessary and effective (Magliano et al., 2005; O'Reilly & McNamara, 2006). Increase in working memory resources may be insured through applying strategies (e.g., McNamara & Scott, 2001). Accordingly, language proficiency of L2 students can be improved by helping them pick up strategies of writing which support the procedures of writing task in case they pay less attention to processes connected with planning, drafting, and revising their essays. Therefore, academic writing pedagogy, concerning the features reported by Coh-Metrix analyses, may benefit from a focus

on applying appropriate writing strategies and techniques that would improve the Iranian L2 learners' writing proficiency.

As there are always certain restrictions to conducting research and may prevent researchers from achieving their desired goals, no research can claim to be a perfect examination of its subject matter. This study is no exception. Some potential methodological limitations have been identified as follows. It should be taken into account that true representative of the larger population of EFL learners in Iran may not be M.A. students in Islamic Azad University of Isfahan (Khorasgan). Further empirical studies using much larger samples from different universities in Iran are needed.

The corpus contains Discussion and Conclusion section (Chapter Five) of all the theses investigated and that is another important limitation that can be addressed. Other chapters could show at least partial plagiarism, that is, they were not written by students themselves as the thesis composers. That was the case about English L2 theses. In the context of EFL considered in this study, the students' thesis-writing ability may be best represented in fifth chapters of their theses.

Finally, despite some minor generalizations, attempts have been made to avoid being technically judgmental about the results and, instead; promote more modern computer-assisted language analyzers in corpus studies in Iran. Regarding the fast pace technology, Iranian researchers are expected to be familiar with text examiner tools specifically if they work with oral and/or written corpora. So, what matters is that reports of the present kind hopefully trigger Iranian EFL researcher's mind to depart from more traditional methods and pursue their studies with cyber assistance on the terrain ahead. Researchers may investigate potential web tools, texts analyzers, and computational softwares or applications to help identify flaws within paragraphs written by Persian learners of English. Alternatively, a joint-venture of students studying English

and computer programmers (IT professionals) can work upon developing a localized and standardized version of, say, Coh-Metrix for internal research purposes.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Masoud Azadnia http://orcid.org/0000-0001-5425-7862

References

Ansarin, A., Karafkan, M., & Hadidi, Y. (2021). The effects of task type on Iranian EFL learners' use of lexical diversity and sophistication. *Applied Research on English Language*, *10*(4), 39-70. doi: 10.22108/are.2021.126660.1673

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database* (CD-ROM). Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8,* 243-257. DOI: https://doi.org/10.1093/llc/8.4.243

Chenoweth, N. A., & Hayes, J. R. (2001). Fluency in writing generating text in L1 and L2. *Written Communication*, *18*(1), 80-98. DOI: https://doi.org/10.1177/0741088301018001004

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,*14, 35-47.

Clavel-Arroitia, B., & Pennock-Speck, B. (2021). Analysing lexical density, diversity, and sophistication in written and spoken telecollaborative exchanges. *Computer Assisted Language Learning Electronic Journal (CALL-EJ), 22*(3), 230-250

Cobb, T. (2018). *Compleat lexical tutor*. Retrieved from http://www.lextutor.ca.

Coh-Metrix (2012). *Cohmetrix.com*. Retrieved 16 September 2019, from the World Wide Web: http://cohmetrix.com/documentation_indices.html

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, *33*(4), 497-505.

Crossley, S. A., & McNamara, D. S. (2008). Assessing second language reading texts at the intermediate level: An approximate replication of Crossley, Louwerse, McCarthy, and McNamara (2007). *Language Teaching, 41*, 409-429. DOI: https://doi.org/10.1017/S0261444808005077

Crossley, S. A., & McNamara, D. S. (2009). Computationally assessing lexical differences in L2 writing. *Journal of Second Language Writing, 17*(2), 119-135. DOI: https://doi.org/10.1016/j.jslw.2009.02.002

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*(2 & 3), 170-191.

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing, 26*, 66-79. DOI: https://doi.org/10.1016/j.jslw.2014.09.006

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly, 42*, 475-493. DOI:

Crossley, S. A., Louwerse, M. M., McCarthy, P. M., & McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *Modern Language Journal, 91*(2), 15-30. DOI: https://doi.org/10.1111/j.1540-4781.2007.00507.x

Crossley, S. A., McCarthy, P. M., & McNamara, D. S. (2007). Discriminating between second language learning text-types. In D. Wilson & G. Sutcliffe (Eds.), *Proceedings of the 20th international*

Florida Artificial Intelligence Research Society international conference (pp. 205-210). Menlo Park, California: AAAI Press.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning, 59*(2), 307-334. DOI:

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2010). The development of Polysemy and frequency use in English second language speakers. *Language Learning, 60,* 573-605. https://doi.org/10.1111/j.1467-9922.2010.00568.x

Dufty, D. F., Graesser, A. C., Louwerse, M., & McNamara, D. S. (2006). Is it just readability, or does cohesion play a role? In *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1251-1256).

Durrant, P., Moxley, J., & McCallum, L. (2019). Vocabulary sophistication in first-year composition assignments. *International Journal of Corpus Linguistics, 24*, 33-66.

Ebrahimi, S. F., & Imandar, S. (2021). Grammatical complexity in research articles: Iranian local journals and international journals. *Issues in Language Teaching*, *10*(2), 301-323. doi: 10.22054/ilt.2022.61187.600

Ellis, N. C., & Ferreira-Junior, F. (2009). Construction learning as a function of frequency, frequency distribution, and function. *The Modern Language Journal*, *93*(3), 370-385. DOI: https://doi.org/10.1111/j.1540-4781.2009.00896.x

Engber, C. A. (1995) The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4*, 139-155.

Esfandiari, R., & Barbary, F. (2017). A contrastive corpus-driven study of lexical bundles between English writers and Persian writers in psychology research articles. *Journal of English for Academic Purposes, 29*, 21-42. 10.1016/j.jeap.2017.09.002.

Esfandiari, R., & Jafari, H. (2021). Morphological complexity across descriptive, expository, and Nnarrative text types in Iranian lower-intermediate language learners. *Issues in Language Teaching*, *10*(1), 237-267. doi: 10.22054/ilt.2021.59736.580

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Gilhooly, K. J., & Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods, Instruments, & Computers, 12,* 395-427.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, & Computers, 36,* 193-202.

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing, 9,* 123-145. DOI: https://doi.org/10.1016/S1060-3743(00)00019-9

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing, 18*(3), 218-238.

Hinkel, E. (2011). What research on second language writing tells us and what it doesn't. *Handbook of Research in Second Language Teaching and Learning, 2,* 523-538. https://doi.org/10.1002/j.1545-7249.2008.tb00142.x

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing, 19,* 57-8410. DOI: https://doi.org/10.1191/0265532202lt220oa

Jurafsky, D., & Martin, J. H. (2002). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* Upper Saddle River, NJ: Prentice-Hall.

Keshavarz, M. H. (2007). *Contrastive analysis and error analysis.* Tehran: Rahnama.

Kim, M., & Crossley, S. A. (2018). Modeling second language writing quality: A structural equation investigation of lexical, syntactic, and cohesive features in source-based and independent writing. *Assessing Writing, 37,* 39-56.

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing, 34*, 12–24.

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

Kyle, K., Crossley, S. A., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication version 2.0. *Behavior Research Methods, 50*(3), 1030-1046.

Louwerse, M. M., McCarthy, P. M., McNamara, D. S., & Graesser, A. C. (2004). Variation in language and cohesion across written and spoken registers. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th annual conference of the Cognitive Science Society* (pp. 843-848). Mahwah, NJ: Erlbaum.

Maamuujav, U. (2021). Examining lexical features and academic vocabulary use in adolescent L2 students text-based analytical essays. *Assessing Writing, 49*.

Maas, H. D. (1972). Zusammenhang zwischen Wortschatzumfang und Länge eines Textes. *Zeitschrift für Literaturwissenschaft und Linguistik, 8*, 73-79.

Magliano, J. P., Todaro, S., Millis, K., Wiemer-Hastings, K., Kim, H. J., & McNamara, D. S. (2005). Changes in reading strategies as a function of reading training: A comparison of live and computerized training. *Journal of Educational Computing Research, 32*(2), 185-208.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development*. New York, NY: Palgrave Macmillan.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavioral Research Methods, Instruments, & Computers, 42*, 381-392. DOI: 10.3758/BRM.42.2.381

McCarthy, P. M., Lehenbauer, B. M., Hall, C., Duran, N. D., Fujiwara, Y., & McNamara, D. S. (2007). A Coh-Metrix analysis of discourse variation in the texts of Japanese, American, and British Scientists. *Foreign Languages for Specific Purposes, 6*, 46-77.

McCarthy, P. M., Lewis, G. A., Dufty, D. F., & McNamara, D. S. (2006). Analyzing writing styles with Coh-Metrix. In G. Sutcliffe & R. Goeble (Eds.), *Proceedings of the Florida Artificial Intelligence Research Society International Conference* (pp. 764-769).

McNamara, D. S., & Scott, J. L. (2001). Working memory capacity and strategy use. *Memory & Cognition, 29*(1), 10-17.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. New York, NY: Cambridge University Press.

O'Reilly, T., & McNamara, D. S. (2006). Reversing the reverse cohesion effect: Good of text structures. In A. Kao & S. Poteet (Eds.), *Natural language processing of the 19th Annual Florida Artificial Intelligence Research Society International Conference.*

Paivio, A. (1965). Abstractness, imagery, and meaningfulness in paired-associate learning. *Journal of Verbal Learning and Verbal Behavior*, *4*(1), 32-38.

Palincsar, A.S., & Brown, A.L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 1,* 117-17

Reid, J. R. (1992). A computer text analysis of four cohesion device in English discourse by native and nonnative writers. *Journal of Second Language Writing, 1,* 79-107. DOI: https://doi.org/10.1016/1060-3743(92)90010-M

Templin, M. C. (1957). *Certain language skills in children*. Minneapolis: University of Minnesota Press.

Toglia, M. P., & Battig, W. R. (1978). *Handbook of semantic word norms*. Hillsdale, NJ: Erlbaum.

Vo¨gelin, C., Jansen, T., Keller, S. D., Machts, N., & Mo¨ller, J. (2019). The influence of lexical features on teacher judgments of ESL argumentative essays. *Assessing Writing, 39,* 50-63.

Xiaofei Lu, Jinlei Deng, (2019). With the rapid development: A contrastive analysis of lexical bundles in dissertation abstracts by Chinese and L1 English doctoral students,

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics, 31*(2), 236-259.