UK Academic Publishing

British Journal of Language and Literary Research (BJLLR)

Vol. 1 | No. 1 | 2026 |pp. 26-33

Website: https://www.ukapub.com/journals/bjllr

Received: 2 November 2026 Revised: 12 November 2026 Accepted: 17 November 2026 Published: 22 November 2026

Original Research Article

A Corpus-based Comparative Analysis of Lexical Richness in EFL vs. ESL Undergraduate Students' Written Texts

Mohammed Sho'ee Mohammed Damom[1*]

[1] Department of English, Faculty of Education-Abs, Hajjah University, Yemen

*Corresponding author: mdmom@hau.edu.ye

Abstract: This study aims at investigating lexical richness in 40 texts of the written English texts produced by the United Arab Emirates undergraduate learners of English as a foreign language vs. Hong Kong learners of English as a second language. The data were taken from the ICNALE project and analysed using a free online software called "Lexical Complexity Analyzer". Lexical richness includes measures like lexical diversity, lexical sophistication, number of different words, token/type ratio and lexical variation. The findings showed that lexical diversity is a reliable and valid measure to discriminate between different proficiency levels and between the status of learning L2 whether it is a second or foreign language. Lexical Sophistication (LS) and Token/Type Ratio (TTR) proved that they are affected by topic of writing. However, Number of Different Words (NDW) and Lexical Variation (LV) proved that they are less affected by variables like Topic, Proficiency level and Nationality.

Keywords: Lexical Complexity, Lexical richness, Lexical Sophistication, Lexical Variation, Lexical Density

1. Introduction:

After the arrival of technology it became possible to store learners' data in electronic format. Many spoken and written corpora produced by students of English as a second or foreign language had been compiled and used for various research purposes adopting many approaches to study these corpora. One of those approaches makes use of corpora to explore the "interlanguage" of non-native learners focusing on the word choices they have and comparing the frequency of use in the texts of non-native to that of native speakers. The results of the studies that adopt this approach have been applied towards pedagogic purposes. For instance, annotated corpora were made available to students in the classroom to explore particular lexical or grammatical phenomena and to allow the students to discover the way they use language for themselves. Also, studies of overuse/underuse of some linguistic features were presented to students, making them aware of how their own writing differs from native writing (James, 1998). Another interesting application of these corpora involves using the corpus as a source for student exercises. This helps teachers produce exercises from a learner corpus, including cloze, comprehension questions, jumbled sentences, matching exercise and multiple choice questions.

© 2026 The Author(s). Published by UK Academic Publishing. This is an open access article distributed under the terms of the Creative Commons Attribution (CC BY 4.0) License.

Those corpora have also been empirically studied for academic rather than pedagogic purposes. Lexical richness is one of the topics that received attention in those corpora-based studies. There are many different aspects of lexical richness, such as lexical diversity (i.e., type-token ratio), lexical variation (e.g., type of different word type but focused on lexical word), lexical sophistication (i.e., the proportion of advanced words), average word length, and lexical density (i.e., the proportion of lexical words in the whole text) (Daller et al., 2007; Read, 2000). Nevertheless, the English corpora compiled in written form by Arab learners of English have not yet received enough attention for such analysis. Thus, this paper questions the written English texts produced by the United Arab Emirates undergraduate learners of English as a foreign language vs. Hong Kong learners of English as a second language. The targeted texts are 40 texts of the written texts included in the ICNALE Corpus project (20 texts for each group). The objective in mind is to analyze the students' lexical richness in those written texts.

2. Measuring lexical

Lexical richness has already been studied both in applied linguistics and in the context of word frequency distributions. In applied linguistics, a number of tests have been developed for measuring the lexical usage of language learners (for an overview, see for example Read, 2000). Lexical richness measures are used to assess the lexical proficiency level of students and those measures include Lexical Diversity, the Type/Token Ratio, Lexical Variation, Lexical Density and Lexical Sophistication and the selection between them depends on the type of data and purpose of study. In the context of word frequency distributions, lexical richness has also been studied through their lexical frequency spectrum: that is ranking the words in a text according to their frequency of occurrence (viz. the words that occur once, twice, tree times, and so on).

In this study, a set of measures are adopted to study lexical richness. These include:
1- Lexical diversity: The result of dividing the number of types (i.e., different words) by the number of tokens (i.e., all words).
2- Lexical sophistication: It is the proportion of advanced words or the percentage of words which are not among the 2,000 high-frequency words.
3- Type/Token Ratio (TTR): The ratio of new words (type) to the total number of all words (token).
4- Number of Different Words (NDW): The calculation of the number of words without repetition.
5- Lexical Variation: This measurement shows how the learner tends to repeat a particular lexical item.

3. Literature Review

Grant and Ginther (2000) examined a sample of 90 Test of Written English (TWE) essays, written at three levels of proficiency to analyze essay length, lexical specificity, lexical features, grammatical structures, and clause level features. They revealed that a combination of all of these characteristics determined L2 writing proficiency level. Douglas (2012) examined the papers written by a group of students who were non-native speakers of English. He came to the conclusion that lexical richness correlates highly with writing ability. In turn, the writing ability predicts academic success and lexical richness could be used to predict their chance of being

lecturers or used as a criterion for their admission to post graduate studies. **Olinghouse and Wilson (2013)** examined the students' use of different variety of words when writing three different genres: stories, persuasion, and informative. They indicated that students vary their words depending on the genre they wrote. They use most diverse words when writing stories, and use more varied words when writing persuasive genre than when writing informative texts. **Mazgutova and Kormos (2015)** analyzed argumentative academic writing. They found that although the lower-level proficiency group showed significant improvements in the measures of lexical diversity than the higher proficiency group, the changes in lexical features (e.g., lexical variability) can be observed in the text of all students. **Gregori-Signes and Clavel-Arroitia (2015)** analyzed lexical density and lexical diversity in university students' written discourse. They reported that Lexical Frequency Profile (LFP) seems to correlate well with an independent measure of vocabulary size. **Djiwandono (2016)** conducted a study to identify the lexical richness of senior students by comparing them to academic papers written by their lecturers. This study attempted to determine the difference between the type-token ratio (TTR); the difference between the use of 2000-word level; the difference between the use of academic words; the difference between the students' essays and their lecturers' in terms of the use of words other than the 2000-word level and the academic words ("off-list words").The essays written by the respondents were submitted to a website for vocabulary profiling (http://www.lextutor/ca/vp). This study showed that the lecturers fare better in terms of TTR and academic words, but write slightly fewer 2000-word level and off-list words than their students. While the differences in TTR and academic words are significant, the differences in the use of 2000-word level and off-list are not significant. **Bestgen (2017)** conducted a study on formulaic measures, lexical features (i.e., lexical diversity, lexical sophistication) in two datasets of EFL learner texts. Bestgen showed that combining diversity and sophistication measures can explain a statistically significant component that predicts writing quality.

Based on what has been shown in prior works presented above, this study addresses the following research questions:

1. Is there any significant difference between lexical diversity, lexical sophistication, number of different words, token/type ratio and lexical variation in the EFL and ESL written texts?
2. Is there any significant effect of proficiency level, topic selection and nationality of learner on lexical diversity, density, sophistication, TTR and lexical variation?

4. Data and Methodology

4.1 Data Collection

The written module (Version 2.1) of the International Corpus Network of Asian Learners of English (ICNALE) was used in this study (cf. Ishikawa, 2013, 2014). This corpus is selected because it is readily available as an open-accessed corpora and the type of data it includes is interesting and representative of the chosen research topic. The ICNALE is an international learner corpus developed by Shin Ishikawa, Kobe University, Japan including more than 10,000 topic-controlled speeches and essays produced by college students (incl. grad students) in many countries/ regions in Asia as well as English native speakers. It enables users to study not only learner English but also varied Englishes in Asia. The topics are common both for essays and spoken monologues and learner are required to show their opinions about two statements: (a) It is important for college students to have a part-time job and (b) Smoking should be completely

banned at all the restaurants in the country. Dialogue tasks are also related to those topics. The written essays used in this study are 200-300-words essays about these two topics mentioned above. About 20 texts were selected for each group (i.e., 20 texts produced by United Arab Emirates students and 20 texts produced by Hong Kong students). The participants vary in their proficiency level between B1 and B2 level in each group and the topics selected include "part-time job and "smoking".

4.2 Instrumentation

To measure aspects of lexical richness, the current study uses a free online software called "Lexical Complexity Analyzer" that is available in https://aihaiyang.com/software/lca/single/. After computing the lexical richness indices, the second stage of analysis includes the use of the SPSS statistical analysis software. One of the advantages of this study is that it is almost completely computerized because all the profiles of data were done in the computer.

4.3 Procedures of analysis

The procedures of analysis include these steps:

1- Downloading the corpus from the ICNALE official site in (.txt) format files.

2- Selecting an equal number of the written texts produced by United Arab Emirates students and those produced by Hong Kong students (20 texts for each group). Those texts belong to the levels B1 and B2.

2-Manually copying and submitting the texts to the free online software "Lexical Complexity analyzer" to generate the automatic calculations of lexical richness indices.

3- Saving the automatically generated calculations and indices in an Excel sheet file.

4- Importing the data into SPSS statistical software to be statistically analysed.

5. Results and Discussion

The section below presents the answers to the first research question stated before (Is there any significant difference between the EFL and ESL written texts in terms of lexical diversity, lexical sophistication, number of different words, token/type ratio and lexical variation?).

5.1 Lexical Richness and Topic

An Independent Sample T Test was carried on and the results are shown below in Table (1) the comparison occurs between the indices of lexical richness when the grouping variable is topic (part-time job vs. smoking).

Table (1) Descriptive Statistics for the effect of Topic on Lexical Richness

	Topic	N	Mean	Std. Deviation	t	Sig.
LD	part job	20	0.50	0.04	-.475	.637
	smoking	20	0.50	0.07		
LS	part job	20	0.14	0.05	-7.570	.000
	smoking	20	0.30	0.08		
NDW	part job	20	105.65	18.82	-1.665	.104
	smoking	20	115.30	17.83		
TTR	part job	20	0.47	0.06	-2.372	.023
	smoking	20	0.52	0.06		
LV	part job	20	0.63	0.08	-1.502	.141
	smoking	20	0.67	0.07		

Table (1) above shows that there is a significant difference between the means of the LS and TTR. The p-value recorded is .000 in LS and .023 in TTR and this is too much lower than the

significance level .05. However, the difference is non-significant in LD (where the p-value recorded .637), NDW (the p-value was .104) and LV (the p-value was .141) and these records are higher than the significance level .05.

5.2 Lexical Richness and Level of proficiency

Again the results of an Independent Sample T Test are provided in Table (2) below. This table shows the calculations of Lexical Diversity (LD), Lexical Sophistication (LS), Number of Different Words (NDW), Token/Type Ration (TTR) and Lexical Variation (LV) when the grouping variable is level of proficiency i.e., B1 vs. B2).

Table(2) Descriptive Statistics for the effect of Level of Proficiency on Lexical Richness

	Level of proficiency	N	Mean	Std. Deviation	t	Sig.
LD	B1	20	0.48	0.06	-2.289	.028
	B2	20	0.52	0.04		
LS	B1	20	0.22	0.09	-.152	.880
	B2	20	0.22	0.12		
NDW	B1	20	107.55	16.64	-.987	.330
	B2	20	113.40	20.64		
TTR	B1	20	0.48	0.05	-1.236	.224
	B2	20	0.51	0.06		
LV	B1	20	0.65	0.08	-.385	.703
	B2	20	0.66	0.09		

The table above shows that there is a significant difference between the means of the LD only where the p-value resulted is .028 and this is lower than the significance level .05. The difference in the means of LS is .880, NDW is .330, TTR is .224 and the LV is .703. All of that shows that the difference is non-significant in these four lexical aspects when the dependent variable is Level of proficiency.

5.3 Lexical Richness and Nationality

In the same way the difference between the same lexical richness aspects is tested when the grouping variable is (Nationality of the respondents) i.e., United Arab Emirates Learners vs. Hong Kong Learners. The results are shown in Table (3) below.

Table(3) Descriptive Statistics for the effect of Nationality on Lexical Richness

	Nationality	**N**	**Mean**	**Std. Deviation**	**t**	**Sig.**
LD	UAE	20	0.48	0.07	-2.222	.032
	HONG KONG	20	0.52	0.03		
LS1	UAE	20	0.20	0.09	-1.471	.149
	HONG KONG	20	0.24	0.11		
NDW	UAE	20	105.95	13.88	-1.555	.128
	HONG KONG	20	115.00	22.03		
TTR	UAE	20	0.48	0.06	-1.070	.291
	HONG KONG	20	0.50	0.06		
LV	UAE	20	0.64	0.08	-1.092	.282
	HONG KONG	20	0.67	0.08		

It is clear from the table above that there is a significant difference between the means of the LD only which is .032. The difference in the means of LS1 is .149, NDW is .128, TTR is .291 and the LV is .282. This indicates that the difference is non-significant between these four lexical aspects when the dependent variable is Nationality.

The second section answers the second question of the current study (Is there any significant effect of proficiency level, topic selection and nationality of learners on lexical diversity, number of different words, lexical sophistication, TTR and lexical variation?). The analysis here presents the effect of all those independent variables on each one of the indices of lexical richness separately.

5.4 Lexical Diversity (LD)

Consider Table (4) below to see the effect of level of proficiency, topic and nationality on Lexical Diversity (LD) as the first one of the indices of lexical richness.

Table (4) Effect of Level of Proficiency, Topic and Nationality on Lexical Diversity (LD)

Lexical Aspect (Dependent variable)	Independent Variables	Sig.
LD	Level of proficiency	.028
	Nationality	.032
	Topic	.637

Table (4) above shows that the dependent variable (lexical Diversity) is significantly affected by Level of Proficiency (.028) and Nationality (.032). However, the effect of Topic is not significant (.637).

5.5 Lexical Sophistication (LS)

Now, consider table (5) below to see the effect of the same independent variables (Level of Proficiency, Topic and Nationality) on the dependent variable (Lexical Sophistication).

Table (5) Effect of Level of proficiency, Topic and Nationality on Lexical Sophistication (LS)

Lexical Aspect (Dependent variable)	Independent Variables	Sig.
LS	Topic	.000
	Nationality	.148
	Level of proficiency	.880

With reference to Table (5) above, it is shown that the dependent variable (Lexical sophistication) is significantly affected by Topic (.000). However, it is neither significantly affected by Nationality (.148) nor by Level of proficiency (.880).

5.6 Number of Different Words (NSW)

Table (6) below shows the effect of level of proficiency, topic and nationality on the (Number of Different Words (NSW)).

Table (6) Effect of Level of proficiency, Topic and Nationality on Number of Different Words (NDW)

Lexical Aspect (Dependent variable)	Independent Variables	Sig.
NDW	Topic	.104
	Nationality	.128
	Level of proficiency	.330

Table (6) above exhibits that the dependent variable (Number of Different Words (NDW)) is not significantly affected by any one of the three independent variables level of proficiency, topic and nationality. The p-value recorded for the difference in means in all the dependent variables is higher than the significance level .05.

5.7 Token/Type Ratio (TTR)

Table (7) below shows the effect of level of proficiency, topic and nationality on the Token/Type Ratio (TTR).

Table (7) Effect of Level of proficiency, Topic and Nationality on Token/Type Ratio (TTR)

Lexical Aspect (Dependent variable)	Independent Variables	Sig.
TTR	Topic	.023
	Level of proficiency	.224
	Nationality	.291

As inferred from Table (7) above, the dependent variable (Token/Type Ratio) is significantly affected by Topic (.023). However, it is not significantly affected by Nationality (.224) and Level of proficiency (.291).

5.8 Lexical Variation (LV)

Table (8) below shows the effect of level of proficiency, topic and nationality on Lexical Variation (LV).

Table (8) Effect of Level of proficiency, Topic and Nationality on Lexical Variation (LV)

Lexical Aspect (Dependent variable)	Independent Variables	Sig.
LV	Topic	.141
	Nationality	.282
	Level of proficiency	.703

Table (8) above exhibits that the dependent variable Lexical Variation (LV) is not significantly affected by any one of the three independent variables Level of proficiency, Topic and Nationality.

6. Conclusion

Lexical richness measures aim to reveal the richness of lexis used in a text, with the intention also of discovering the richness of the lexicon which produced the text. Some of the indices of lexical richness tested in this study showed that they are reliable and valid measures in L2 writing. Some of those indices (such as lexical diversity) showed discrimination between different proficiency levels and between the status of learning L2 whether it is a second or foreign language. Other indices of lexical richness such as Lexical Sophistication (LS1) and Token/Type Ration (TTR) proved that they are reliable measures between the groups of study when the topic of writing is different. However, the lexical indices (Number of Different Words (NDW) and Lexical Variation (LV) proved that they are less affected by the independent variables of Topic, Proficiency level and Nationality.

References

Bestgen, Y. (2017). Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. System, 69, 65-78.

Daller, H., Milton, J., & Treffers-Daller, J. (Eds.). (2007). Modelling 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, Vol. 5 No. 2, pp. 209-216

Douglas, R. S. (2012). Non-Native English Speaking Students at University: Lexical Richness and Academic Success. (Unpublished dissertation). Universitas Calgary.

Grabowski, E. and D. Mindt (1995). "A corpus-based learning list of irregular verbs in English". ICAME Journal 19: 5-22.

Granger, S. (1999). "Use of tenses by advanced EFL learners: evidence from an error-tagged computer corpus". In H. Hasselgard, S. Oksefjell (eds.), Out of Corpora- Studies in Honour of Stig Johannson. Amsterdam: Rodopi, pp.191-202.

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. Journal of Second Language Writing, 9(2), 123-145.

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

Ishikawa, S. (2014). Design of the ICNALE-Spoken: A new database for multi-modal contrastive interlanguage analysis. In S. Ishikawa (Ed.), Learner corpus studies in Asia and the world, 2 (pp. 63-76). Kobe, Japan: Kobe University.

James, C. (1998). Errors in Language Learning and Use: Exploring error analysis. Harlow: Longman.

Mazgutova, D., & Kormos, J. (2015). Syntactic and lexical development in an intensive English for academic purposes programme. Journal of Second Language Writing, 29, 3-15.

McCarthy, P.M. (2005). An Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical Diversity (MTLD). Unpublished PhD dissertation. The University of Memphis.

Olinghouse, N. G., & Wilson, J. (2013). The relationship between vocabulary and writing quality in three genres. Reading and Writing, 26(1), pp. 45 – 65.

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

Tweedie, F.J., & Baayen, R.H. (1998). How variable may a constant be? Measures in lexical richness in perspective. Computers and the Humanities 32, 323-352.