

Research paper

Lexical Complexity in Texts Written by Native and Non-native English Writers: A Case of Conclusion Sections of Social and Natural Science Papers

Aly Javady

M.A in Applied Linguistics (TEFL), Science and Research Branch, Islamic Azad University, Iran

Javady, A. (2024). Lexical Complexity in Texts Written by Native and Non-native English Writers: A Case of Conclusion Sections of Social and Natural Science Papers, *Journal of new advances in English Language Teaching and Applied Linguistics*, 6(2), 195-214.

Doi: 10.22034/Jeltal.2024.6.2.9

*corresponding author's email: Javady.ELT@gmail.com

Abstract

Lexical complexity of a text is the extent to which it includes difficult words. Higher lexical complexity index in a piece of writing indicates the quality of that text and lexical knowledge of that writer. The goal of the current study was a comparison between the academic text quality (difficulty) of non-native and native English writers. Thereby, the researcher collected 72 conclusion sections of research papers in natural sciences (Biology, Physics Environmental science), as well as social sciences (Psychology, Linguistics & Economics) then analyzed them based on writing quality indexes. The papers were obtained from leading journals (JCR-Scopus) later were divided into two groups; namely, those written by the English native and non-native writers. According to the findings, texts written by natives in economics are more complex and have more diverse words. In addition, natural science conclusions based on all three measures were richer in lexicon than social science. The results of the study will help improve the writing quality of none-native writers of English.

Keywords: Academic texts, text readability, lexical complexity, conclusion sections of social science papers, conclusion sections of natural science papers

Received: 2024-01-12 **Revised** 2024-04-20 **Accepted** 2024-04-25

Introduction

Language and its features, such as discourse markers and lexical complexity, can serve as reliable predictors of the text's value to academicians, including researchers, professors, and students. Besides being a highly valued skill, required for academic credentials worldwide, academic writing in English as a second language (L2) is an intriguing area of study because it is a very difficult form of language production that pushes a learner to the edge of his/her abilities

(O'Leary & Steinkrauss, 2022). Well-written language with a high level of lexical richness in academic writing significantly helps readers understand the content better (Müngen, 2023). On the other hand, text comprehension is a sophisticated cognitive process that is impacted by the text's linguistic elements as well as the reader's characteristics (McNamara & Magliano, 2009 as cited in Strohmaier et. al, 2023). Learning and comprehension of a written text can be impacted by its linguistic features (Strohmaier et. al, 2023).

For many years, linguists or writing scholars have looked for methods to investigate the relationship between language features in texts and the writing development quality of first-language (L1) and second-language (L2) writers (Crossley, 2020). As Crossley (2020) states, many studies have confirmed the idea that writers' word choice and the way they are arranged, along with their general patterns of coherence, are the main indicators of the quality of their writing. Hence, it is crucial to analyze academic texts to find out their features to eventually provide guidelines for future writers in order to present a more qualified academic text. Regarding the importance of lexical complexity studies, Alsahlanee and Jaganathan (2022) state that:

In L2 writing studies, lexical complexity is a crucial concept that is frequently used to assess writers' skill and writing quality. Lexical complexity measures, serve as indicators of writing quality meaning that the higher the score, the higher the writing quality rating. (P. 1105)

Therefore, the ultimate aim of the current study is to provide a deeper insight into using more efficient and effective language in academic writing, thus the scholars may provide rich contexts. In addition, studies like O'Leary and Steinkrauss (2022) have confirmed that lexical complexity of an academic writing can reveal its quality.

A scientific paper is composed of many sections including Abstract, introduction, literature review, methodology, results and conclusion. Readers of papers in order to understand the ultimate results and aim of study, reach to the conclusion section; hence this part is one of the most crucial sections of a paper. Also, composing this section unlike other parts of the paper, solely depends on the special writing skills of the researchers. Therefore, special care should be taken in putting together the conclusion section of scientific papers. Unfortunately, this level of expertise is only reachable by practice and no association coaches this skill. The complexity of a text could be regarded both a positive and negative aspect. If the writer chooses totally unfamiliar and difficult words that have no association with the coherence of the text, the quality of that writing will be low. As Schmied et. al (2007) asserts if complexity is organized to meet the reader's or listener's expectations and demands for coherence, then a "good" text can be complex; in other words, the art of creating appropriate texts is to balance the complexity with appropriate coherence. Lexical complexity is a rarely studied research area that the researcher attempted to investigate in this study. For this reason, this study aimed at improving the writing skill of conclusion sections by providing the results. This study was designed to compare the writing ability of non-native scholars and provide some suggestions based on the findings of the study to improve their writing ability.

The innovative aspect of this study is comparing the lexical complexity of academic texts written by English native writers to those of non-natives. The following research questions were addressed in this study:

1. Is there any significant difference in lexical complexity of texts written by native and non-native English writers?
2. Is there any significant difference in lexical complexity of social and natural science papers in conclusion sections?

Literature Review

Lexical complexity is the degree of difficulty of a text in the process of reading and understanding by a reader. Of course, this difficulty level can be defined in many aspects depending on the reader, being an English native speaker or a non-native speaker. The reader also could be an academic or non-academic person. Even the English proficiency level, especially the lexical competence can be different from person to person; therefore, difficulty of a text could be claimed to be a relative concept. Therefore, in order to be able to provide a precise comparison, a sample text should be regarded as the standard English written text. In this study the difficulty of a text is compared to the texts composed by native writers; therefore, the researcher perceived these texts as the standard academic English writing.

Adel and Ghorbani Moghadam (2015) inspected the rhetorical move structure of 30 conclusion sections in three distinct fields—Persian literature, psychology, and applied linguistics—in a genre-based study. The corpus was obtained from top Iranian and international journals. The move structure of this corpus was examined using the framework developed by Yang and Allison (2003). The main moves from each article were taken out after analysis, then the frequencies were computed and contrasted. The conclusion sections of the articles on psychology and applied linguistics did not show any notable differences, despite the fact that the cross-disciplinary and cross-linguistic analyses revealed significant variations. Persian literature articles, on the other hand, showed greater diversity, which may indicate that authors adhere to their own standards when crafting conclusion sections.

In terms of lexical density, diversity, and sophistication, Azadnia (2021) assessed the lexical richness of a corpus of doctoral dissertations authored by Iranian TEFL students. The

corpus was compared to a first language (L1) baseline that included doctoral dissertations authored by native English speakers in order to examine the lexical features using the computational tool Coh-Metrix. The texts produced by Iranian TEFL learners were more sophisticated but had a less varied vocabulary, according to a comparison of the L1 and L2 corpora. Furthermore, the L2 corpus had a higher lexical density of nouns and adjectives than the L1 corpus. The best predictors of lexical richness, according to the results of a discriminant function analysis (DFA), were the features that represented lexical sophistication and density because they were able to distinguish between the two sub-corpora with significant accuracy.

Shahmoradi et. al (2021) attempted to identify 4-word lexical bundles through analysis of corpus accumulated from the abstract and conclusion sections of 1000 papers authored by L1 Farsi and L2 English writers in applied linguistics and information technology. After analyzing the data collected from roughly 600,000 words, using Antconc software, the researchers performed a functional analysis using Hyland's (2008, a,b) framework. The analysis demonstrated that texts in applied linguistics had more lexical bundles than information technology. Additionally, the abstract sections contained more lexical bundles used by L1-Persian writers. In concluding sections, both authors employed the same quantity of bundles. In terms of the distinctions between native and non-native writers, the latter typically employed a greater number of abstract lexical bundles. Regarding the distinctions between the two fields, applied linguistic writers tended to employ more bundles in both native and non-native corpora due to their increased lexicon and formula knowledge. Lexical bundles were used more frequently by applied linguistics writers in both sections. The proportions of functional distributions of bundles, with respect to both types and tokens, did not exhibit many differences between the two disciplines in the abstract section analysis. However, because they used these bundles as a professional starting point, information

technology writers used fewer text-oriented bundles than their applied linguistics counterparts, demonstrating their proficiency in English.

Nasseri and Thompson (2021) used LCA, TAALED, and Coh-Metrix language processing tools to analyze the abstracts of 210 dissertations written by EFL, ESL and English L1 postgraduate students with topic and text length effects being controlled. They investigated the connection between the construct-distinctiveness of lexical density, diversity and 15 lexical indices. The measure-testing process also evaluated how well each measure captured the lexical diversity differences of these texts in a pair or group of closely related measures (in terms of the quantification methods). With so many measures available, the goal of this process was to help future writing researchers choose the best ones while also obtaining a limited number of distinct measures that capture lexical diversity as an indicator of lexical proficiency. The EFL group's texts had the lowest lexical density, according to the results.

Mayangsari et.al (2021) investigated lexical complexity and readability in English journal texts specifically at introduction sections through a quantitative method. The outcome of the study demonstrated that the lexical density of text was greater than 0.50, meaning the higher usage of content words than function words. In addition, the majority of corpus selected from the journals had lexical diversity over 0.30. The highest level of lexical diversity was attributed to EEJ and Edulife journals both having the index of 0.35. The introduction sections contained advanced words from Academic Word Lists (AWL). The selected texts were readable for the selected groups based on “Flesch Kincaid Reading Ease, Flesch Kincaid Grade Level and Miyazaki readability index” (Mayangsari, et al, 2021. p. 298)

Alsahlanee and Jaganathan (2022) tried to describe the current stance of Iraqi academic writers through analyzing their publications in linguistics and comparing them to the works of L1

and other L2 writers. The corpus of the study was selected from Scopus listed papers. The lexical complexity was analyzed by Complexity Analyzer LCA. The findings suggested that the writings of Iraqi writers had significantly lower lexical complexity in comparison to English L1 writers.

North and Zampieri (2023) conducted a study titling “Features of lexical complexity: insights from L1 and L2 speakers”. This study tried to find sizable differences in the lexical complexity assignments of L1 and L2 English speakers. Three LCP datasets were used to extract and compare the complexity assignments of 940 shared tokens: the CompLex dataset, the CERF-J wordlist, and the WCL dataset. For L2 English speakers, perceived lexical complexity was found to be more influenced by word frequency, length, and syllable count than for L1 English speakers. L2 English speakers' lexical complexity was higher than that of L1 English speakers due to the use of "-ness" and "-tion.". In general, root words were found to be simpler than their derivational counterparts. Lexical complexity was influenced by both prevalence and familiarity for L2 English speakers, but only prevalence affected L1 English speakers. For L2 English speakers in particular, concreteness was found to predict lexical complexity. For L1 English speakers, however, a less strong correlation was also noted.

Müngen (2023) studied the differences in lexical richness in English language theses. The corpus included 320 theses in different scientific fields. The texts were analyzed by natural language processing techniques, word diversity, unique word count and a number of other tools. The results suggested that social science texts had higher lexical richness in comparison to natural science theses. In terms of word richness, no difference was found between social and natural science texts.

Alaei et. al (2024) in their study asked 46 Iranian EFL learners to write argumentative and narrative essays in L2, and L1 on different topics. The aim was to compare the syntactic complexity

and lexical diversity in L1/L2 writing of Iranian EFLs. Based on the results the participants had used more complex words in their L1 rather than their L2.

Ayadi (2024) carried out a study with the following title “Exploring lexical complexity in scientific writing: a corpus-based study across disciplines”. She analyzed the lexical complexity of 60 results and discussion sections of research papers written by Tunisian scholars. The papers were in linguistics and engineering fields. The researcher used Lu (2010) text analyzing tool to scrutinize the corpus. After conducting a number of statistical analyses, she found out that engineering texts were written in a laconic manner with higher lexical density whereas linguistics texts had broader use of linguistics forms, producing more varied and richer expressions. According to these results, texts in engineering were more complex than linguistics. In other words, the soft sciences place more value on diversity and adaptability in language use, whereas the hard sciences place more value on accuracy and technicality in their writing.

Method

The researcher adopted a quantitative corpus-based approach to collect and analyze data in this study. The corpus was 72 conclusion sections extracted from three groups that is to say, native English writers and non-native writers. The papers were selected from 2010 onward publications. The researcher collected the conclusions of each field in two major groups namely L1 and L2, thus he had 12 long files that were imported to the platform in each process. Then the texts were analyzed using the specified tools. All of the measures were calculated in an online web-based platform (Reuneker, 2017).

The Type-Token ratio and lexical density were used to quantify lexical complexities. The percentage of words with lexical rather than grammatical qualities relative to the total number of words in a text is known as the lexical density, also known as the type-token ratio. Lexical density

gauges how challenging a passage or text is. It is calculated by multiplying the ratio of lexical words to total words in a text by 100. Lexical density was computed using the formula $\text{lexical density} = (\text{lexical words}/\text{total words}) * 100$. Table 1 demonstrates the number of papers that the researcher collected for each field and L1/L2 groups.

Table 1
The corpus of the study

		English L1	English L2	Total	
Social Science	Psychology	6	6	12	Personality and social psychology review
	Linguistics	6	6	12	Annual review of linguistics
	Economics	6	6	12	The quarterly journal of economics
Natural Science	Biology	6	6	12	Annual review of biochemistry
	Physics	6	6	12	Living reviews in relativity
	Environmental science	6	6	12	Annual review of environment and resources
Total				72	

After extracting the conclusion sections from the papers, the analyses were carried out. In Table 2, the meaning of each measure that was used in the analyses process of the study is explained.

Table 2
The measures and explanations

Labels	Explanations
Tokens	Total number of words
Types	Number of unique words
Type-token-ratio (TTR)	Number of types divided by number of tokens
Mean word frequency (MWF)	Number of tokens divided by number of types
Guiraud's Index	Guiraud (1954)
Measure of textual lexical diversity (MTLD)	McCarthy & Jarvis (2010)

Once the analyses of the texts were over the researcher classified the results and compared them to find out their differences. Then the results were demonstrated in tables followed by their interpretation.

Results

The researcher used measures provided by Reuneker (2017) in order to analyze the texts. The first step to carry out the data was collecting conclusion sections from aforementioned journals. Then they were divided based on the field and being written by native and non-native English writers. The researcher checked the affiliations of the writers in order to find out whether they are native or non-native. Thus, writers from Canada, United States, Britain, Australia and south Africa were perceived as native English writers and authors from the rest countries were regarded as non-native. Finally, using Reuneker (2017) tool the analyses were carried out. This tool provided a number of measures including tokens, types, type-token-ratio (TTR), mean word frequency (MWF), Guiraud's Index, measure of textual lexical diversity (MTLD). The results of the study are provided based on each field in Table 3. This table provides the overall outcome of the analyses.

Table 3
The analyses of measures

	Natural Sciences						Social Sciences					
	Physics		Environmental science		Biology		Psychology		Economics		Linguistics	
	L1	L2	L1	L2	L1	L2	L1	L2	L1	L2	L1	L2
Tokens	2593	2242	1466	2896	1930	2023	1371	1519	2916	2608	2317	1395
Types	1092	925	699	1206	839	903	675	738	1084	1051	971	658
(TTR)	0.42	0.41	0.48	0.42	0.43	0.45	0.49	0.49	0.37	0.48	0.42	0.47
(MWF)	2.37	2.42	2.1	2.4	2.3	2.24	2.03	2.06	2.69	2.48	2.39	2.12
Guiraud's Index	21.44	19.54	18.26	22.41	19.1	20.08	18.23	18.94	20.07	20.58	20.17	17.62
(MTLD)	136.47	102.37	122.57	122.01	107.47	112.38	117.74	117.01	97.44	109.62	116.52	93.59

In Table 4, the type-token ratios of fields are shown. The results of this measure vary between 0 and 1. If the number is close to 0, the text has more function words and is lexically dense. Conclusions in Economics written by L1 writers (0.37) are the most lexical dense among six academic fields and among social science branches. On the other hand, conclusions in economics written by non-natives and environmental science texts written by natives were the least complex texts (0.48).

Table 4
Type-token ratio analyses results

		TTR Index	
		L1	L2
Social Science	Psychology	0.49	0.49
	Linguistics	0.42	0.47
	Economics	0.37	0.48
Natural Science		0.42	0.48
	Biology	0.43	0.45
	Physics	0.42	0.41
	Environmental science	0.48	0.42
		0.44	0.42

In order to provide a vivid image of the results the researcher demonstrated them on figure 1. According to the figure based on type-token ratio texts in social sciences written by L2 writers have more lexical complexity.

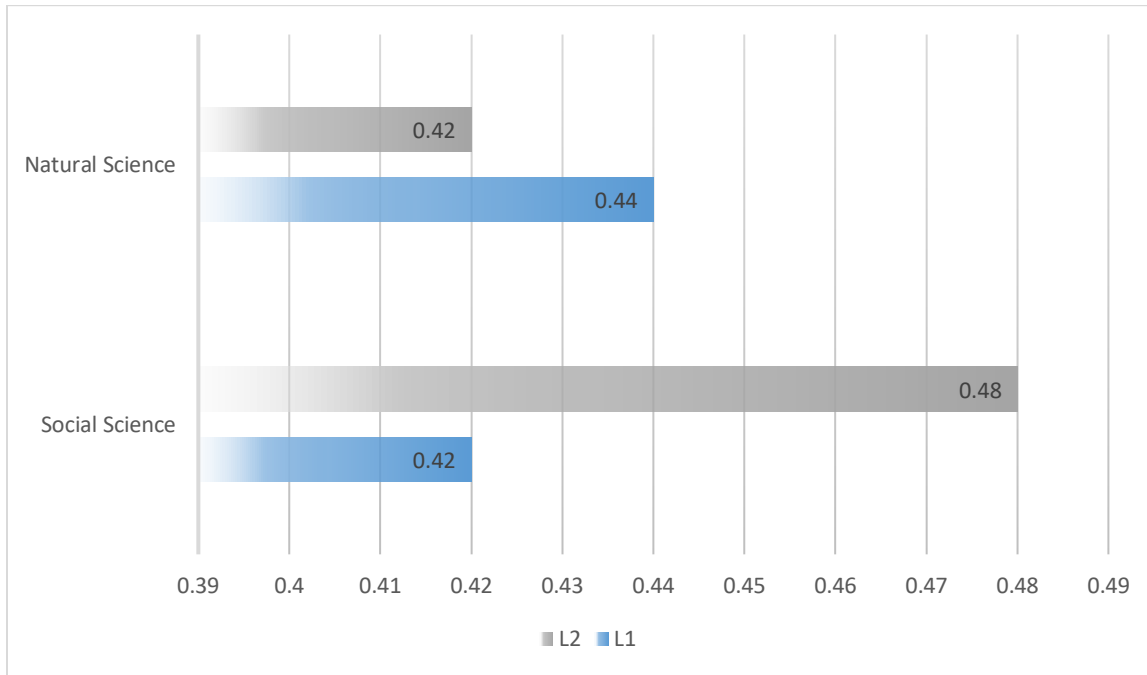


Figure 1: The results of Type-token ratio analyses

Based on TTR conclusions written by L2 writers in social science are more complex than the ones written by L1. Table 5 illustrates the results according to Guiraud's (1954) Index. Based on this index the results differ in comparison to TTR measure. The most complex conclusions are environmental science according this measure (22.41). The least complex texts on the other hand were psychology texts authored by L1s.

Table 5
 Guiraud's (1954) Index results

		Guiraud's (1954) Index		
		L1	L2	
Social Science	Psychology	18.23	18.94	18.58
	Linguistics	20.17	17.62	18.89
	Economics	20.07	20.58	20.32
		19.49	19.04	19.26
Natural Science	Biology	19.1	20.08	19.59
	Physics	21.44	19.54	20.49
	Environmental science	18.26	22.41	20.33
		19.54	19.86	20.13

Overall based on Guiraud's (1954) Index, Natural science texts were more complex than social science texts. It is clear as can be seen in figure 2 texts written by L2 writers in natural science are more lexically dense.

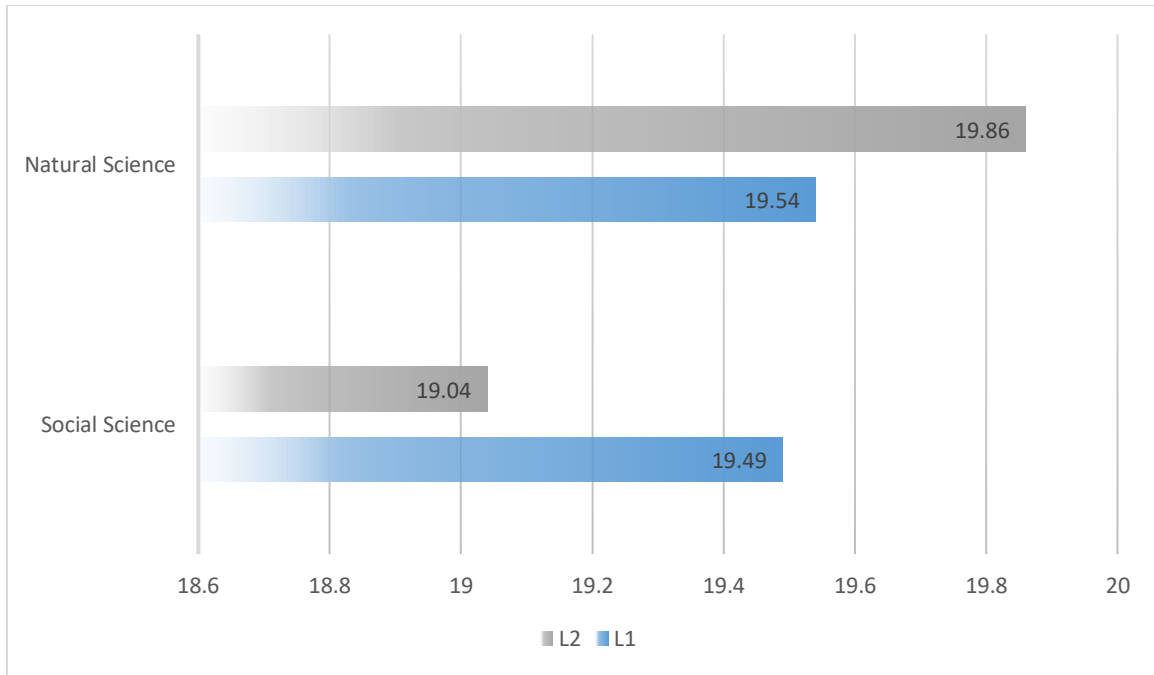


Figure 2: Guiraud's (1954) Index results

The researcher also conducted the analyses by the index of lexical diversity (MTLD) by McCarthy & Jarvis (2010) which is shown in Table 6.

Table 6
McCarthy & Jarvis's (2010) index results

Measure of textual lexical diversity (MTLD) McCarthy & Jarvis (2010)				
		L1	L2	
Social Science	Psychology	117.74	117.01	117.37
	Linguistics	116.52	93.59	105.05
	Economics	97.44	109.62	103.53
Natural Science		110.56	106.74	108.65
	Biology	107.47	112.38	109.92
	Physics	136.47	102.37	119.42
	Environmental science	122.57	122.01	122.29
		116.36	109.49	117.212

Figure 3 shows the results of McCarthy & Jarvis's (2010) index. The mean score of McCarthy & Jarvis's (2010) index for natural science conclusions written by L1s were the most complex text groups.

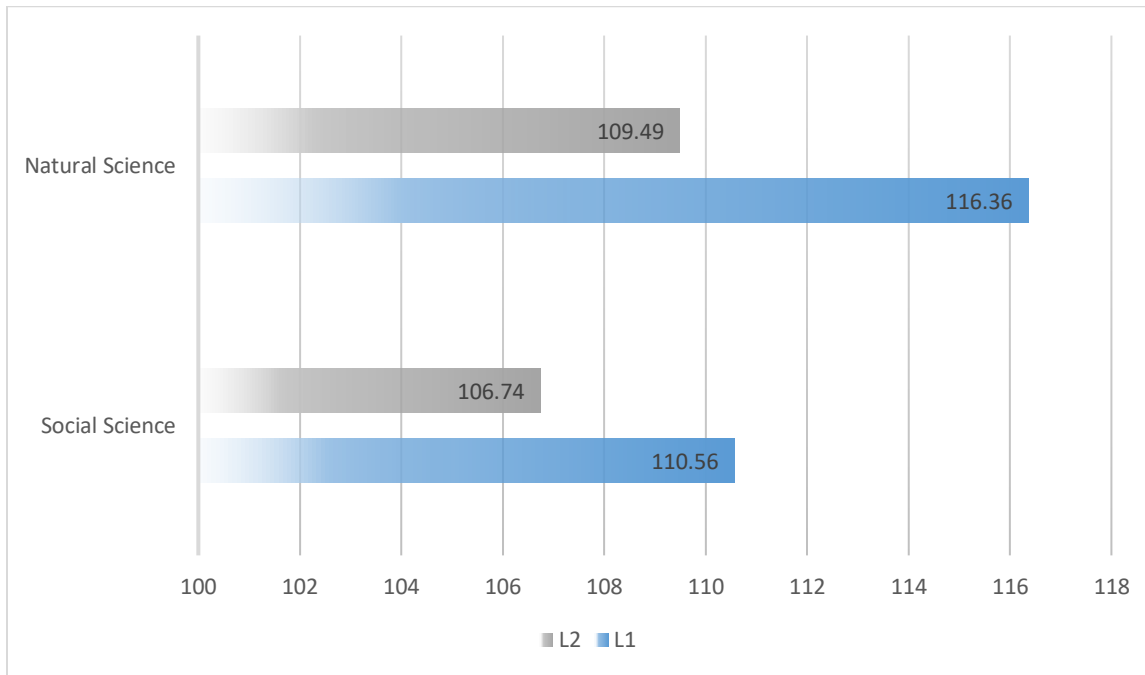


Figure 3: The results of McCarthy & Jarvis's (2010) index

The researcher analyzed the text through three different measures of lexical density. As it is clear from the tables and figures, each index provides contradicting results. Therefore, these indexes need more analysis and validity to be used by future studies. Based on the results texts written by native writers in economics are more complex based on TTR. Based on Guiraud's (1954) Index texts in environmental science authored by non-native writers has more word diversity. Finally, based on the measure of textual lexical diversity (MTLD) McCarthy & Jarvis (2010) texts in environmental science had the most diver words.

In addition, natural science conclusions, based on all three measures were richer in lexicon than social science. The summary of the findings based on all three measures are mentioned in Table 7.

Table 7
Comparing the results based on the three indexes of the study

	L1(Social Science)	L1(Natural Science)	L2(Social Science)	L2(Natural Science)
TTR Index	0.42	0.44	0.48	0.42
Guiraud's (1954) Index	19.49	19.54	19.04	19.86
McCarthy & Jarvis (2010)	107.47	116.36	112.38	109.49

As the Table 7 suggests each index suggests different results. This could be due to the file size or the criteria that each method uses to analyze the text. But as it is clear in Table 7, natural science texts based on all three measures had higher lexical richness. Nevertheless, this difference with non-native speakers of English and other branches of science is not large.

Discussion

The present study intended to find out the level of lexical complexity in different fields of social and natural sciences. Since the papers were extracted from finest journals, the results of the analysis could be a template for non-native English writers. As the results suggests there is a minor difference between native and non-native English writers in lexical difficulty and diversity. Based on the TTR index (the numbers suggest revers measures, meaning the lower the number the higher the TTR is) texts in social science written by English native writers and natural science texts written by non-native writers had the highest difficulty in word choices. Based on TTR value texts in social science written by non-native speakers were the easiest texts to read. According to Guiraud's (1954) Index, texts written by non-native writers in natural sciences were the most

difficult texts whereas social science texts written by the same non-native writers were the easiest texts. Finally, following McCarthy & Jarvis (2010) index the results were as follows, natural science texts written by native English writers were the most difficult text. The difference in the results of each index should be a topic for future investigation.

The results of the study contradict some earlier studies like Mungen (2023) who claimed that social sciences had higher lexical richness in comparison to natural sciences. But the results obtained through Guiraud's (1954) and McCarthy & Jarvis's (2010) Index in this study are in line with the findings of Ayadi (2024), who concluded that natural science texts are more complex than other fields. As a result, while some of the current study's findings were consistent with those of earlier research, others were not. Curriculum developers along with units in research method should include units in text quality so that the writers could adjust their text difficulty level based on their specific readers. In addition, this adjustment should be in alignment with their text cohesion. Finally, this study suggests that in order for a text to be comprehensible for a wide variety of readers, its lexical complexity should not be exceedingly high.

Conclusion

The current study concluded that there are differences in lexical complexity of academic texts in social and natural science. All three metrics showed that the lexical richness of natural science texts was higher. This difference is not very noticeable, though, when compared to non-native English speakers and other scientific fields. Also, there are contrasts between native and non-native writers. This study also innovatively utilized multiple measuring device to compare their results. The inconsistencies in the results should be investigated by the providers of these tools to find out whether that is due to the text length or there are other discrepancies. But in some

cases, there were higher usage of lexically complex words. Also, regarding the writers of the texts, the intention of an Academic writer should not solely be using sophisticated words but to create a balance between technical words, sophisticated words and general words so that the text could be easily read.

The results of this study are hoped to increase the concern of scholars in computational linguistics to improve the validity of such tools. The results are also valuable for academic writers and researchers in the area of lexical complexity. Novice writers can utilize the findings of this research to have a better choice of lexicons. Further studies can investigate other text analyzing indexes to uncover the quality of scientific papers. This study found out which disciplines had the highest lexical density. The frequency, accuracy, and complexity of lexical words may be the main subjects of future research projects. Numerous academic fields as well as different sections of theses and research articles (abstract, introduction, methodology, etc.) can be used to investigate these topics. Other facets of lexical complexity, such as the proportion of low-frequency to high-frequency words or the proportion of technical terms (vocabulary items) to non-technical words, can be investigated in future research. To find out if the same results apply in two different written and spoken form contexts, researchers can also investigate whether the study's findings hold true in spoken settings, like seminars. The results of the study can also be viewed more thoroughly by looking at the various text genres, lexical complexity, and the way in which discourse markers are used. There were a number of limitations in this study. The main issue with text analyzers were the validity of them which changed as the text length increased.

References

- Adel, S.M.R. & Ghorbani Moghadam, R. (2015). A comparison of moves in conclusion sections of research articles in psychology, Persian literature and applied linguistics. *Teaching English Language*, 9(2)167-19.
- Alaei, F., Kaivanpanah, Sh. & Mohammadi, M. (2024). Syntactic complexity and lexical diversity in L1/L2 writing of EFL learners. *Journal of English language teaching and learning university of Tabriz*, 16(33)97-115.
- Alsahlane, A. & Jaganathan, P. (2022). Lexical Complexity in The Writings of Iraqi, English L2, and English L1 Writers. *Asian Journal of University Education (AJUE)* 18(4)1105-1118.
- Azadnia, M. (2021). A corpus-based analysis of lexical richness in EAP texts written by Iranian TEFL students. *Teaching English as a Second Language Quarterly*, 40(4)61-90.
- Ayadi, M. (2024). Exploring lexical complexity in scientific writing: a corpus-based study across disciplines. *LLT Journal*, 27(1)421-433.
- Crossley, S.A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11(3), 415-443. <https://doi.org/10.17239/jowr-2020.11.03.01>
- Guiraud, P. (1954). *Les Caractères Statistiques du Vocabulaire. Essai de méthodologie*. Paris: Presses Universitaires de France.
- Hyland, K. (2008a). As Can Be Seen: Lexical Bundles and Disciplinary Variation. *English for Specific Purposes*, 27(1), 4-21.
- Hyland, K. (2008b). Academic Clusters: Text Patterning in Published and Postgraduate Writing. *International Journal of Applied Linguistics*, 18(1), 41-62.
- Nasseri, M. & Thompson, P. (2021). Lexical density and diversity in dissertation abstracts: Revisiting English L1 vs. L2 text differences. *Assessing Writing*, 47(2), 1-11.
- McNamara, D. S., & Magliano, J. P. (2009). Towards a comprehensive model of comprehension. In B. Ross (Ed.), *The psychology of learning and motivation* (pp. 297-384). New York, NY: Elsevier.
- McCarthy P. M., & Jarvis S. (2010). MTL, VocD, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42, 381-392.
- Mayangsari, A., Wuli Fitriati, S. & Sutopo, D. (2021). Lexical Complexity and Readability Realized in The Introduction Sections of Selected English Journals. *English Education Journal*, 11 (2) 298-298.
- Müngen, A.A. (2023). Exploring lexical richness in English-language theses across disciplines: a comparative analysis. *Journal of Scientometric Research*, 12(2)459-468.
- North, K., & Zampieri, M. (2023). Features of lexical complexity: insights from L1 and L2 speakers. *Frontiers in Artificial Intelligence*, 6(1236963)0-17. doi: 10.3389/frai.2023.1236963
- O'Leary, J.A., & Steinkrauss, R. (2022). Syntactic and lexical complexity in L2 English academic writing: Development and competition. *Ampersand*, 9(100096)1-10. <https://doi.org/10.1016/j.amper.2022.100096>
- Reuneker, A. (2017). *Lexical Diversity Measurements*. Retrieved 07 December, 2024, from <https://www.reuneker.nl/files/ld>.

- Shahmoradi, N., Jalali, H. & Ghadiri, M. (2021). Lexical bundles in the abstract and conclusion sections: The Case of Applied Linguistics and Information Technology. *Applied Research on English Language*, 10(3), 47-76. doi: 10.22108/are.2021.128024.1703
- Schmied, J., Al haase, Ch. & Povolná, R. (2007). Complexity and coherence: *Approaches to linguistic research and language teaching* (pp.1-12) Publisher: Göttingen: Cuvillier.
- Strohmaier, A.R., Ehmke, T., Härtig, H. & Leiss, D. (2023). On the role of linguistic features for comprehension and learning from STEM texts. A meta-analysis. *Educational Research Review*, 39. 1-23, doi:10.1016/j.edurev.2023.10053

Appendix A

https://www.reuneker.nl/files/ld/#

Lexical Diversity Measurements

On this page, you can calculate a large number of measures of lexical diversity. For each measurement, a reference is given. If you want to read about the *why* and *how* of these calculations, you should look up the references. If you use this tool for your research, please cite it as follows.

Reuneker, A. (2017). *Lexical Diversity Measurements*. Retrieved 07 December, 2024, from <https://www.reuneker.nl/files/ld>.

For other calculation tools (such as [ngrams](#), [wordlists](#) and [keyword analysis](#)), and for contact details, see <https://www.reuneker.nl>.

Input

Text (paste the text you want to analyze here):

The current meta-analysis demonstrated that people on average still clearly have traditional cognitions about SDS, in particular with regard to men and women having casual sex, having sex for the first time at an early age, and general sexual activity level. We also found clear evidence of traditional SDS in within- or between-subject experimental studies assessing differences in the evaluation, or expectation, of mens and womens sexual behavior. Nevertheless, SDS were less traditional in countries with higher levels of gender equality. This meta-analysis further demonstrated that both evolutionary theory and biosocial theory provide relevant and testable predictions with regard to the existence of SDS. It appears that a hybrid model including both evolutionary processes related to gender differences in parental investment and sexual strategies, as well as the societal division in

Options

- Use stemming (based on [Richard Heyes's \(2005\) implementation](#) of the [Porter algorithm](#))
 - Treat hyphenated words (e.g. writer-director) as one word
- No. of words to use: (min: 100, max: 1000000)

Output

Tokenization	View tokenization	<i>text split into words</i>
Frequency list	View frequencies	<i>top 40 of most frequent words</i>
Tokens	1519	<i>total number of words</i>
Types	738	<i>number unique of words</i>
Type-token-ratio (TTR)	0.49	<i>number of types divided by number of tokens</i>
Mean word frequency (MWF)	2.06	<i>number of tokens divided by number of types</i>
Mean segmental TTR (MSTTR)	1	<i>mean ttr for text segmented into chunks of 100 words; only for text longer than 100 words</i>
Guiraud's Index	18.94	<i>Guiraud (1954)</i>
Herdan's C	0.9	<i>Herdan (1960; 1964)</i>
Yule's I	129.73	<i>Yule (1944); Gries (2004)</i>
Yule's K	77.08	<i>Yule (1944); Oakes (2004, pp. 203-5)</i>
Maas's a²	0.03	<i>Maas (1972); Tweedie & Baayen (1998); Treffers-Daller (2013)</i>