International Journal of Education and Development using Information and Communication Technology (IJEDICT), 2023, Vol. 19, Issue 3, pp. 69-84

Student-Written Versus ChatGPT-Generated Discursive Essays: A Comparative Coh-Metrix Analysis of Lexical Diversity, Syntactic Complexity, and Referential Cohesion

Tlatso Nkhobo & Chaka Chaka
University of South Africa

ABSTRACT

This article reports on a comparative analysis of two sets of essays, student-discursive essays (SDEs) and ChatGPT-generated discursive essays (ChatGPT-GDEs) on the same essay topic using Coh-Metrix. It focused on three Coh-Metrix indices, lexical density, syntactic complexity, and referential cohesion as the basis for the comparative analysis. The authors also conducted a *t* test on the Coh-Metrix results, especially the mean scores, in relation to these three linguistic indices. Using convenience sampling, the study selected seven SDEs from the essays that were submitted as part of an assignment for an English Studies module in the second semester of 2020 at the University of South Africa. ChatGPT was prompted with the same essay topic that had been used for the SDEs. Overall, at raw mean score levels, the SDEs outperformed ChatGPT-GDEs in lexical density and referential cohesion, while ChatGPT-GDEs did so in syntactic complexity. Nonetheless, at a *t* test level, there was no statistically significant difference between the mean scores of the two essay sets in relation to the three Coh-Metrix linguistic indices investigated in this study.

Keywords: *Student discursive essays (SDEs); ChatGPT-generated discursive essays (ChatGPT-GDEs); lexical density; syntactic complexity; referential cohesion; Coh-Metrix*

INTRODUCTION

ChatGPT (**Chat G**enerative **P**re-trained **T**ransformer), which can be regarded as an instance of a paroxysm of machine-powered artificial intelligence (AI), has sparked widespread debates in the educational ecosystem. The rapid increase of AI-powered large language models (LLMs) such as ChatGPT is likely to simultaneously enhance and threaten classical academic writing as it is currently understood. These AI tools tend to challenge several theoretical constructs and practices that guide teaching and learning in higher education institutions (HEIs). For example, ChatGPT, which is also a generative AI chatbot or a generative LLM, can generate human-like responses based on prompts or on prompt engineering (Chaka, 2023a, 2023b, 2023c; Khare, 2023). It, thus, relies on zero-shot, one-shot or few-shot prompts (Chen, 2023; Tam, 2023) fed into it, from which it can generate different kinds of text such as reports, legal opinions, and essays. ChatGPT uses both natural language processing (Hariri, 2023; Mattas, 2023) and deep learning (Chen, 2023; Deng & Lin, 2022), and possesses the same capabilities as machine learning because of its ability to study large datasets through the employment of neural networks.

Recent research has explored the affordances ChatGPT has and the pitfalls it may pose to standardised and normative educational practices (Barrot, 2023; Deng & Lin, 2022; Kohnke, Moorhouse, & Zou, 2023; Su, Lin, & Lai, 2023). Against this backdrop, the current study investigated the occurrence of the three linguistic components, lexical diversity, syntactic complexity, and referential cohesion in the essay samples of first-year students registered for an English Studies module, at the University of South Africa (UNISA). The essay samples were analysed through the corpus software application, Coh-Metrix. In addition, the study sought to investigate the prevalence of lexical diversity, syntactic complexity, and referential cohesion in the essay samples generated by ChatGPT in response to the same essay topic that first-year students were given as part of their formative assessment in 2023.

Coh-Metrix is a corpus analysis tool that can display different linguistic components existing in text files uploaded into it (Alan, 2021; Chon & Shin, 2020; Howie, 2022; McNamara, Louwerse, McCarthy, & Graesser, 2010; Nkhobo & Chaka, 2023; Shipt, 2022; Tabassum, Mahmood, Mahmood, & Haider, 2022; Wang, Engelhard Jr., & Combs, 2023; Yildiz & Yeşilyurt, 2021). Coh-Metrix is capable of analysing and assessing large language texts subjected to it within few seconds due to its use of natural language processing algorithms (Dowell & Kovanovic, 2022; Gibson & Shibani, 2022; Reilly & Schneider, 2019) and machine learning capabilities (Ikram & Castle, 2020; Latifi & Gierl, 2021). The latter enable it to learn from existing linguistic features and make comparisons/patterns against texts subjected to it. All these capabilities allow researchers to study various linguistic features used in written texts. As a computational tool, Coh-Metrix is mostly used in the field of learning analytics to study students' learning trends as well as writing patterns (Dowell & Kovanovic, 2022; Fincham, Whitelock-Wainwright, Kovanović, Joksimović, van Staalduinen, & Gašević, 2019; Reilly & Schneider., 2019).

Considering the points highlighted above, the current study had the following research hypotheses:

- H_0: There will be no significant difference in the lexical diversity, syntactic complexity, lexical diversity, and referential cohesion of the student discursive essays (discursive essays written by students) and those generated by ChatGPT on the same essay topic.
- H_a: There will be significant difference in the lexical diversity, syntactic complexity, lexical diversity, and referential cohesion of the student discursive essays (discursive essays written by students) and those generated by ChatGPT on the same essay topic.

This pair of hypotheses relates to each linguistic feature in the null and alternative forms. This means that each linguistic feature is implied individually in each hypothesis.

RELATED LITERATURE

ChatGPT in Higher Education

Several scholarly papers have been written about the use of ChatGPT in the different fields of higher education (HE). Examples are Bishop (2023), Biswas (2023), Chaka (2023a), Cotton, Cotton & Shipway (2023), Kasneci et al. (2023), Kitamura (2023), Lund & Wang (2023), Mohammed, Al-Ghazali & Alqohfa (2023), and Wang et al. (2023). There are also scholarly papers that have analysed ChatGPT-generated text. Among these papers are Chaka (2023c), Chen (2023), Ifelebuegu, Kulume & Cherukut (2023), Rudolph, Tan & Tan (2023), Su et al. (2023), and Mattas (2023). In the current study, ChatGPT was used to generate discursive essays on the same topic that was given to first-year students. Thereafter, the researchers of the current paper examined the use of lexical diversity, syntactic complexity, and referential cohesion in student-discursive essays (SDEs) and in ChatGPT-generated discursive essays (ChatGPT-GDEs) as analysed by Coh-Metrix.

Zhou et al. (2023) used Coh-Metrix to analyse the written essays produced by ChatGPT and Chinese Intermediate English (CIE) students on a given narrative topic. The ChatGPT- and student-generated texts were analysed using five linguistic features: narrativity, word concreteness, syntactic simplicity, referential cohesion, and deep cohesion. The study comprised 40 participants (10 males and 30 females) from a university in China. The participants were requested to write a narrative essay and submit it within a week. They were required not to use grammar and spelling application tools. ChatGPT was also instructed to generate 40 English essays on the same topic that was given to the students. The data from both sets of essays were analysed using multivariate analysis of variance (MANOVA). The results showed that ChatGPT outperformed students in relation to narrativity, word concreteness, and referential cohesion but that the students performed better than ChatGPT in terms of syntactic simplicity and deep cohesion. This implies that ChatGPT, in this case, was able to perform better than the students in some of the linguistic components, while it did not do so in the other linguistic features. Human intervention is required to produce texts which are accessible and readable in terms of syntactic simplicity and deep cohesion. Therefore, the current study aims to study the use of lexical diversity, syntactic complexity, and referential cohesion in both SDEs and ChatGPT-GDEs.

In another study, Seals & Shalin (2023) employed Coh-Metrix to compare ChatGPT- and human-generated texts by evaluating their individual sentences from their long-form analogies in relation to biochemical concepts. The study conducted a supervised classification procedure to analyse 78 features available in Coh-Metrix with a view to investigating language, text cohesion, and readability. The participants in the study were registered in three biochemistry courses in one of the universities in the United States. The study focused on two biochemical processes, glycolysis and enzyme kinetics. The participants were provided with sample analogies that they used to discuss biochemical process and they were requested to create their own analogies to explain the biochemical process of the subject of their choice. The study employed a linear ridge classifier to analyse data. The results of the study indicated that the long-form analogies produced by ChatGPT were different from those written by human subjects. The difference was in terms of their descriptive linguistic properties and their underlying psycholinguistic properties. Human summaries had more variance in their word selection as compared to ChatGPT's word selection. Given that the prevalence of generative AI tools is almost ubiquitous and pervasive, it is important for teachers to learn how to identify AI-generated texts in terms of lexical diversity, syntactic complexity, and referential cohesion to protect the integrity of written essay assessments.

Linguistic Features in which Coh-Metrix Has Been Used

Studies have been conducted on the use of Coh-Metrix to investigate linguistic features displayed by student essays. Such studies include but are not limited to the following: Jafarie & Tabrizi (2022); Kim (2022); Leal et al. (2021); McCarthy et al. (2022); Mahadini, Setyaningsih & Sarosa (2021); Nasseri & Thompson (2021); islami et al. (2022); Shpit (2022); and Yildiz & Yeşilyurt (2021). All of these studies did not investigate lexical diversity, syntactic complexity, and referential cohesion together as linguistic features.

For example, on the one hand, Nasseri & Thompson (2021) examined lexical density and diversity differences in Master of Arts (MA) dissertations of English as a first language (L1) postgraduate students versus those of English as a foreign language (EFL) and English as a second language (ESL) postgraduate students in the United Kingdom (U.K.). So, this study wanted to compare English L1 MA dissertations versus ESL dissertations versus EFL dissertations. Coh-Metrix was one of the analysis tools used. The study's corpus included 210 texts and 70 abstracts, each ranging in length from 175 to 300 words. Overall, the EFL dissertations had abstracts that had the lowest lexical density and diversity than the abstracts of the MA dissertations of the other two groups. By contrast, the abstracts of the English L1 and ESL groups displayed a similar level of lexical diversity and dense representation.

On the other hand, Mahadini et al. (2021) focused on Coh-Metrix indices and used a traditional rubric to evaluate 20 essay samples of Indonesian EFL students. In this study, 81 EFL essays were chosen, and 20 were randomly selected for analysis by employing conventional and automatic scoring rubrics. Low narrativity, word concreteness, and high syntactic simplicity were evident in the students' essays. According to the study, using both conventional and Coh-Metrix scores allows for a more thorough understanding of student essays, but conventional scoring takes more time and provides more concrete qualitative examples.

In a different but related context, McCarthy et al. (2022) investigated students' counter and support arguments in argumentative writing. They collected 78 argumentative essays and divided their paragraphs into support, expostulation, counterargument, expostulation, and counterargument with expostulation and background. Coh-Metrix and Gramulator, two computational language analysis tools, were used in the study. The Coh-Metrix results showed that support paragraphs were substantially more prevalent than counterargument paragraphs as the most common types of paragraphs. Deep cohesion changed between paragraph types, suggesting a putative aspect of causation. Causal language can be tagged as a feature of counterarguments based on the analysis of this feature using Gramulator, which suggests that it may be a feature of counterarguments. The studies cited above did not examine the same linguistic components as the current study and its findings will fill this existing gap in the research that investigates the similarities and differences between student-written and AI-generated texts in terms of lexical diversity, syntactic complexity, and referential cohesion.

RESEARCH METHODOLOGY

As previously pointed out, the focus of the current study is to analyse SDEs and ChatGPT-GDEs in terms of their lexical diversity, syntactic complexity, and referential cohesion as exhibited by the corpus software application, Coh-Metrix. There is a dearth of research in the South African higher education (HE) context that has investigated these linguistic features from the point of view of comparing student-written and ChatGPT-generated discursive essays. To this end, the study employed an exploratory research design as described below.

Research Design

This was an exploratory study because it investigated a phenomenon that has not been extensively researched (Swedberg, 2020). Rahi (2017) defines an exploratory research design as a study that seeks new insights and new discoveries. The afore-mentioned definition is closely aligned with the focus of the current study in that it seeks to discover and compare linguistic features (lexical diversity, syntactic complexity, and referential cohesion) used in SDEs and ChatGPT-GDEs.

This type of research design restrains researchers from generalising their results because some of the studies in which exploratory research is mostly applicable are case studies which consist of only a certain number of participants from the larger group as is the case with the current study (Reiter, 2017). The current study adopted a mixed-method approach (Almalki, 2016; Almeida, 2018; Halcomb & Hickman, 2015; Khaldi, 2017) comprising qualitative and quantitative data. Qualitative data is presented through the descriptions of the linguistic features (lexical diversity, syntactic complexity, and referential cohesion), while quantitative data is presented through the mean scores in tabular forms as yielded by Coh-Metric and a *t* test in relation to the three linguistic features used in both SDEs and ChatGPT-GDEs.

Sampling Techniques

The current study adopted the convenience sampling technique in selecting student-written discursive essays. Convenience sampling involves a careful selection of participants or data that is accessible to the researcher (Etikan, Musa & Alkassim, 2016; Rahi, 2017). On this basis, seven student-written discursive essays were selected from those submitted by first-year students, who were registered for an undergraduate English Studies module in the second semester of 2020. ChatGPT was also considered as a participant in that it was instructed to generate seven discursive essays on the same topic that was given to the first-year students in 2020. The current study was granted ethical clearance in accordance with the College of Human Sciences Research Ethics Committee guidelines.

Data Collection

Seven student discursive essays that were submitted under Assessment 2 for an undergraduate English Studies module in the second semester of 2020, were downloaded from the module site hosted on Moodle. The length of the seven SDEs ranged from 419 words to 577 words. The topic for the discursive essay was: *Write an essay in which you discuss three negative effects of using drugs for mood or behaviour syndromes*. The same essay topic, as it is, was used as a long prompt for ChatGPT (ChatGPT-3.5) to generate seven different discursive essays. Each essay was generated separately at a time on 30 June 2023. The length of the seven ChatGPT-GDEs ranged from 492 words to 526 words.

Data Analysis

The two sets of discursive essays were analysed by means of a corpus software application, Coh-Metrix, in terms of their lexical diversity, syntactic complexity, and referential cohesion as mentioned earlier. The seven SDEs, which had been submitted as PDF files, were converted into text files and saved as Microsoft Word files. They were then uploaded into Coh-Metrix for analytical purposes. Similarly, the seven ChatGPT-GDEs were copied and saved as Microsoft Word files. They, too, were uploaded into Coh-Metrix for analysis. Coh-Metrix generated the analysis results for the two sets of text files in relation to their lexical diversity, syntactic complexity, and referential cohesion. Moreover, the two essay sets were subjected to unpaired *t* tests for the following elements: word length per essay, paragraphs per essay, sentences per essay, words per sentence; and lexical diversity, syntactic complexity, and referential coherence.

FINDINGS

The findings of this study are presented in keeping with the two sets of discursive essays mentioned above. The findings are divided into the three linguistic features being investigated in each essay set.

Student-Written Discursive Essay Findings

Table 1 below displays the results for lexical diversity in the seven student discursive essays (SDEs) as analysed by Coh-Metrix for each SDE.

Table 1: *Lexical diversity*

SDE1	SDE2	SDE3	SDE4	SDE5	SDE6	SDE7	Sub-categories
0.619	0.739	0.650	0.736	0.632	0.632	0.645	Lexical diversity, type-token ratio, and content word lemmas
0.421	0.540	0.470	0.525	0.443	0.443	0.447	Lexical diversity, type-token ratio, and all words
74.190	107.152	91.659	110.702	81.748	85.579	101.311	Lexical diversity, Measure of Textual Lexical Diversity (MTLD), and all words
82.677	115.078	96.802	135.258	106.048	107.495	122.791	Lexical diversity, VOCD, and all words

For each SDE, the following clusters of lexical features are illustrated for lexical diversity as an index: lexical diversity, type-token ratio, and content word lemmas; lexical diversity, type-token ratio, and all words; lexical diversity, measure of textual lexical diversity (MTLD), and all words; and lexical diversity, VOCD[1], and all words. For instance, there is a lower lexical diversity, type-token ratio, and content word lemmas for SDE1 (0.619), SDE5 (0.632), SDE6 (0.632), and SDE3 (0.650), while SDE7 (0.645) scored marginally higher in this lexical cluster. By contrast, SDE2 and SDE4 had slightly higher scores of 0.736 and 0.739, respectively. In relation to lexical diversity, type-token ratio, and all words as a cluster, low scores were recorded across all the SDEs. Here, the lowest score was 0.421. In terms of the third cluster (lexical diversity, MTLD, and all words), the highest score was 110.702, with 74.190 as the lowest score. Regarding the last cluster, the highest and lowest scores were 122.791 and 82.677 respectively.

Table 2: *Syntactic complexity*

SDE1	SDE2	SDE3	SDE4	SDE5	SDE6	SDE7	Sub-categories
0.979	0.810	1.013	0.776	0.842	0.830	0.689	Number of modifiers per noun phrase, and mean
0.876	0.896	0.879	0.923	0.880	0.877	0.896	Minimal edit distance, and all words
0.100	0.108	0.070	0.081	0.133	0.135	0.069	Sentence syntax similarity, adjacent sentences, and mean
0.085	0.066	0.067	0.074	0.125	0.125	0.062	Sentence syntax similarity, all combinations, across paragraphs, and mean

Table 2 shows the Coh-Metrix results for syntactic complexity as a linguistic index comprising four clusters for the seven SDEs. These four clusters are as follows: number of modifiers per noun phrase and mean; minimal edit distance and all words; sentence syntax similarity, adjacent sentences, and mean; and sentence syntax similarity, all combinations, across paragraphs, and mean. For the first cluster, the highest score is 1.013 and 0.689 is the lowest score. Pertaining to the second cluster, 0.923 is the highest score, with 0.876 as the lowest score. For the third cluster, 0.135 and 0.069, are the highest and lowest scores, respectively. Lastly, concerning the fourth cluster, 0.125 (a tie of two scores) and 0.062, are the highest and the lowest scores, respectively.

Table 3: *Referential cohesion*

SDE1	SDE2	SDE3	SDE4	SDE5	SDE6	SDE7	Sub-categories
0.607	0.400	0.429	0.320	0.429	0.417	0.346	Noun overlap, adjacent sentences, binary, and mean
0.75	0.5	0.607	0.440	0.486	0.5	0.577	Argument overlap, adjacent sentences, binary, and mean
0.626	0.227	0.430	0.473	0.403	0.400	0.340	Argument overlap, all sentences, binary, and mean
0.086	0.043	0.076	0.077	0.069	0.071	0.050	Content word overlap, all sentences, proportional, and mean

Table 3 displays the Coh-Metrix results for referential cohesion as a linguistic index with four clusters for the seven SDEs. For this linguistic feature, these four clusters are: noun overlap, adjacent sentences, binary, and mean; argument overlap, adjacent sentences, binary, and mean; argument overlap, all sentences, binary, and mean; and content word overlap, all sentences, proportional, and mean. As regards the first cluster, the highest score is 0.607, with 0.320 as the lowest score. 0.75 is the highest score, while 0.440 is the lowest score for the second cluster. Concerning the third and the fourth clusters, 0.626 and 0.227 and 0.086 and 0.043 are the highest and lowest pairs of scores in each case.

ChatGPT-Generated Discursive Essay Findings

Table 4 illustrates the results for lexical diversity in the seven ChatGPT-GDEs as analysed by Coh-Metrix for each ChatGPT-GDE. Each ChatGPT-GDE has four clusters of lexical features that are part of lexical diversity as a Coh-Metrix index. These four clusters are: lexical diversity, type-token ratio, and content word lemmas; lexical diversity, type-token ratio, and all words; lexical diversity, MTLD, and all words; and lexical diversity, VOCD, and all words. The highest and lowest scores in the first cluster are 0.417 and 0.390 respectively, while the second cluster has 0.182 and 0.170 (a tie for two scores) as its highest and lowest scores, each. 0.666 and 0.598 and 0.5 and 0.455 rank as a pair of the highest and lowest scores for the last two clusters respectively.

Table 4: *Lexical diversity*

ChatGPT-GDE1	ChatGPT-GDE2	ChatGPT-GDE3	ChatGPT-GDE4	ChatGPT-GDE5	ChatGPT-GDE6	ChatGPT-GDE7	Sub-categories
0.413	0.417	0.411	0.390	0.403	0.405	0.399	Lexical diversity, type-token ratio, and content word lemmas
0.182	0.180	0.170	0.178	0.173	0.172	0.170	Lexical diversity, type-token ratio, and all words
0.613	0.598	0.617	0.666	0.622	0.638	0.645	Lexical diversity, MTLD, and all words
0.458	0.455	0.462	0.5	0.476	0.485	0.485	Lexical diversity, VOCD, and all words

Table 5 shows the Coh-Metrix results for syntactic complexity as a linguistic index for the seven ChatGPT-GDEs. This linguistic index consists of the same four clusters into which the syntactic complexity of the SDEs were divided. For the first two clusters, the pairs of highest and lowest scores are 1.172 and 1.021 and 0.913 (a tie of two scores) and 0.881, correspondingly. The last two clusters have the pairs 0.145 and 0.123 (again, a tie of two scores) and 0.094 and 0.076 as sets of their highest and lowest scores.

Table 5: *Syntactic complexity*

ChatGPT-GDE1	ChatGPT-GDE2	ChatGPT-GDE3	ChatGPT-GDE4	ChatGPT-GDE5	ChatGPT-GDE6	ChatGPT-GDE7	Sub-categories
1.172	1.021	1.067	1.104	1.131	1.133	1.076	Number of modifiers per noun phrase, and mean
0.881	0.913	0.901	0.911	0.903	0.905	0.913	Minimal edit distance, and all words
0.123	0.123	0.135	0.145	0.137	0.132	0.132	Sentence syntax similarity, adjacent sentences, and mean
0.076	0.082	0.094	0.091	0.091	0.087	0.087	Sentence syntax similarity, all combinations, across paragraphs, and mean

Table 6 exhibits the four clusters of the features of referential cohesion as a Coh-Metrix index. For this linguistic feature, the highest and lowest scores for the first cluster are 0.478 and 0.208, whereas for the second cluster, 0.522 and 0.333 are the highest and lowest scores. Concerning the third cluster, 0.605 and 0.497 constitute the highest and lowest scores. With reference to the last and fourth cluster, 0.077 and 0.057 are the highest and lowest scores.

Table 6: *Referential cohesion*

ChatGPT-GDE 1	ChatGPT-GDE 2	ChatGPT-GDE 3	ChatGPT-GDE 4	ChatGPT-GDE 5	ChatGPT-GDE 6	ChatGPT-GDE 7	Sub-categories
0.478	0.348	0.320	0.333	0.375	0.25	0.208	Noun overlap, adjacent sentences, binary, and mean
0.522	0.478	0.400	0.458	0.458	0.333	0.375	Argument overlap, adjacent sentences, binary, and mean
0.605	0.578	0.546	0.497	0.564	0.508	0.528	Argument overlap, all sentences, binary, and mean
0.077	0.073	0.070	0.057	0.071	0.068	0.068	Content word overlap, all sentences, proportional, and mean

DISCUSSION

This section of the paper discusses the findings presented above in line with its two research hypotheses as stated earlier. The discussion is framed around the two sets of discursive essays analysed by Coh-Metrix as described above. First, it focuses on the two sets of essays in terms of their word length per essay, paragraphs per essay, sentences per essay, and words per sentence. These four aspects are discussed with reference to their *t* test results as well. The discussion, then, moves on to lexical diversity, syntactic complexity, and referential coherence, all which are explored against the backdrop of their *t* test results.

Word Length Per Essay, Paragraphs Per Essay, Sentences Per Essay, and Words Per Sentence

The two sets of essays (SDEs/ChatGPT-GDEs) had the following mean scores with regards to their word length per essay, paragraphs per essay, sentences per essay, and words per sentence, respectively: 531.57/511.57; 8.57/7; 30.71/24.85; and 17.59/20.92 as shown in Table 7.

Table 7: *Word length per essay, paragraphs per essay, sentences per essay, and words per sentence in the two sets of essays*

Source	Length (words/essay)	Paragraphs/essay	Sentences/essay	Words/sentence
SDEs	531.57	8.57	30.71	17.59
ChatGPT-GDEs	511.57	7	24.85	20.92

Against this background, in relation to the *t* test results for all the afore-mentioned features, the two essay sets' mean scores were 147.11 and 141.01, respectively. Their two-tailed *p* value was 0.974 as shown in Table 8. Reviewing the data in Table 7, it is clear that, except for the essay length (words per essay), the difference between the remaining three features of these two sets of essays is very minimal. In fact, their two-tailed *p* value is 0.974, which means that the overall difference between these four features of the two essay sets is not statistically significant (see Table 8).

Table 8: *Unpaired t test results for SDEs and ChatGPT-GDEs: Word length per essay, paragraphs per essay, sentences per essay, and words per sentence in the two sets of essays.*

Group	**SDEs**	**ChatGPT-GDEs**	**Two-tailed *p*-value at 0.05**
Mean	147.11	141.08	0.974
SD	256.46	247.11	
SEM	128.23	123.55	
N	4	4	

Lexical Diversity

In terms of raw mean scores, the student discursive essays (SDEs) had higher mean scores in relation to the four dimensions of lexical diversity or vocabulary richness (Bestgen, 2023; Herbold et al., 2023), especially so for the last two dimensions (see Tables 1 and 4), as opposed to ChatGPT-generated discursive essays (ChatGPT-GDEs). This means that, at a surface level, in these clusters of lexical diversity, SDEs outperformed their ChatGPT-GDE counterparts. In fact, the lowest mean score of 82.67 for SDE1 in the last cluster of lexical diversity is higher than 0.485, which is a tie mean score for ChatGPT-GDE6 and ChatGPT-GDE7. However, at a deeper, more nuanced level, the difference between the overall lexical diversity mean scores for both sets of essays is not statistically significant when calculated on a *t* test. See, in this case, a two-tailed *p* value of 0.374 versus *p* = 0.05 in Table 9. Table 10 provides the raw average mean scores for the three linguistic indices (lexical diversity, syntactic complexity, and referential coherence) for each essay set. Since the t test *p* value for both sets of essays for lexical diversity is greater than 0.05 as shown in Table 9) the null hypothesis part of this Coh-Metrix lexical index is retained, while its alternative hypothesis is rejected.

Table 9: *Unpaired t test results for SDEs and ChatGPT-GDEs*

Group	**SDEs**	**ChatGPT-GDEs**	**Two-tailed *p*-value at 0.05**
Mean	17.26	0.43	0.374
SD	29.17	0.11	
SEM	16.84	0.06	
N	3	3	

In a different but related context, Herbold et al. (2023) conducted a large-scale comparison of human-written (German high-school students) versus argumentative essays generated by the two versions of ChatGPT (ChatGPT-3 and ChatGPT-4), which found that the mean score of lexical diversity for human participants was 95.72, which was higher than that of ChatGPT-3 (75.68). Nonetheless, it was lower than that of ChatGPT-4 (108.91). Even though this study did not use any *t* test, it, nevertheless, has some parallels to the current study in that at a surface level of lexical diversity its SDEs had higher mean scores than those of ChatGPT-GDEs (ChatGPT-3.5) for four dimensions of lexical diversity.

***Table 10:** Average mean (M) and standard deviation (SD) for 3 Coh-Metric linguistic indices*

Linguistic indices	**SDEs**		**ChatGPT-GDEs**	
Lexical diversity	50.94	50.70	0.42	0.18
Syntactic complexity	0.48	0.44	0.55	0.52
Referential coherence	0.36	0.21	0.34	0.21

Syntactic Complexity

In contrast to lexical diversity, pertaining to syntactic complexity as a Coh-Metrix index, ChatGPT-GDEs had higher mean scores than its SDE counterparts, especially for the first three clusters of this linguistic index. This is so barring a few instances in which SDEs polled higher mean scores. The last cluster tends to be a mixed bag as each set of essays had higher mean scores in some aspects of this cluster but not in the other aspects. This is the observation made at the raw data level of the mean scores for both sets of essays (see Tables 2 and 5). In respect of their combined, deeper level of their *t* tests, though, the difference between the overall syntactic complexity mean scores for both sets of essays is not statistically significant at *p* = 0.374 (see Table 9; also cf. Table 10). Again, here, the null hypothesis part of the syntactic complexity as a Coh-Metrix linguistic index is retained, while its alternative hypothesis part is rejected.

The study by Herbold et al. (2023) has some relevance here given that there are not yet many studies on student-written versus ChatGPT-generated essays using Coh-Metrix. Its mean score for syntactic complexity (depth) for its human subjects was 5.72, while its mean scores for ChatGPT-3 and ChatGPT-4 were 6.18 and 5.94 respectively. Both versions of ChatGPT outperformed human subjects in this linguistic feature. The same was true of its mean scores for syntactically complex clauses generated by the two versions of ChatGPT, which were 2.31 and 2.08, apiece. That is, they were higher than that of the essays written by human participants (1.81). Again, it must be noted that this study did not employ a *t* test. As stated earlier, in Zhou et al. (2023) study that utilised Coh-Metrix to investigate narrativity, word concreteness, syntactic simplicity, referential cohesion, and deep cohesion between ChatGPT-generated and student-written essays, students performed better than ChatGPT in syntactic simplicity. In the study, the mean scores for syntactic simplicity for both ChatGPT-generated essays and student-written essays were 30.974 and 40.252, respectively. However, this study, too, did not administer a *t* test.

Referential Cohesion

With regards to referential coherence, SDEs had higher mean scores in the first two clusters of this linguistic index than ChatGPT-GDEs, save for one instance in each cluster. The last two clusters are, once more, a mixed bag for both sets of essays as they outdo each other in some aspects, while they fail to do so in the other aspects. Of course, this is at the level of their raw mean scores (see Tables 3 and 6). In terms of the *t* test mean scores, however, the difference between these two sets of essays is statistically insignificant with a *p* value of 0.374 (see Table 9; also cf. Table 10; also cf. Table 10). This means that, as in the previous two instances, the null hypothesis part of the Coh-Metrix linguistic index, referential coherence, is rejected in favour of its alternative hypothesis.

As mentioned earlier, in the Zhou et al. (2023) study, referential cohesion is one of the linguistic elements in which ChatGPT performed better than students. Even though this study did not conduct a *t* test, its mean scores for referential cohesion for both ChatGPT and students were 73.155 and 47.431, respectively. Seals & Shalin's (2023 study, which was cited earlier, found that ChatGPT and human subjects employed cohesive devices (referential, verb, and deep cohesion and connectivity and temporality) differently. The current study focused mainly on referential cohesion, whose results between SDEs and ChatGPT-GDEs have been discussed above.

CONCLUSION, LIMITATIONS, AND RECOMMENDATIONS

In terms of the four essay features investigated for the two sets of essays (SDEs and ChatGPT-GDEs) in this study, it was discovered that the overall difference between these features for both essay sets was very minimal. As such, this difference was not statistically significant. Concerning lexical diversity, at a raw data level, SDEs outperformed ChatGPT-GDEs in the four dimensions of this Coh-Metrix linguistic index. But from the point of view of *t* test results, the difference between the aggregated lexical diversity mean scores for both sets of essays was not statistically significant. Thus, the null hypothesis part of lexical diversity was accepted, while its alternative hypothesis counterpart was rejected.

As regards syntactic complexity as a Coh-Metrix linguistic index, barring a few cases, ChatGPT-GDEs performed better than SDEs at a raw mean score level. However, at a *t* test mean score level, there was no statistically significant difference between the mean scores of these two essay sets. As a result, the null hypothesis portion of syntactic complexity was retained, and its alternative hypothesis part was rejected. Lastly, pertaining to referential cohesion, again, SDEs outdid ChatGPT-GDEs in the first two clusters, notwithstanding one case in each cluster (see Tables 3 and 6). Nevertheless, there was no statistically significant difference between the mean scores of these two essay sets in their *t* test results. To this end, the null hypothesis section of referential cohesion was accepted, while its alternative hypothesis counterpart was rejected. Overall, at raw mean score levels, SDEs outperformed ChatGPT-GDEs in lexical density and referential cohesion, while ChatGPT-GDEs did so in syntactic complexity. Nonetheless, at a *t* test level, there was no statistically significant difference between the mean scores of these two essay sets in terms of the three linguistic indices. The two sets of findings have implications for academic essay writing. The first set implies that while students can outperform ChatGPT in the use of certain linguistic features of essay writing, ChatGPT, too, can outperform students in the production of the other linguistic features of essay writing. The second set has even dire implications: ChatGPT can handle lexical diversity, syntactic complexity, and referential cohesion in essay writing in almost the same way as students can.

The study has limitations. Firstly, its data pool is very small to be applicable to all instances of SDEs and ChatGPT-GDEs. In this case, the results of the study have a contextual applicability to the present study. However, the study has a value for comparing SDEs and ChatGPT-GDEs using Coh-Metrix as there are few studies conducted in this area of research at present. Therefore, it can serve as a reference point for future studies in this area. Secondly, the study focused on one set of student-written essays and on one set of ChatGPT-generated essays. Future studies need to investigate more than one set of essays for both cases.

Notes

1. VOCD is a computer programme for calculating lexical density (Bestgen 2023; deBoer, 2014).

REFERENCES

Alan, L. K. (2021). “Interlanguage issues in noun phrases and information flow”, *Modern English Education*, vol. 22, no. 4, pp. 1-11. http://doi.org/10.18095/meeso.2021.22.4.1

Almalki, S. (2016). “Integrating quantitative and qualitative data in mixed methods research – Challenges and benefits”, *Journal of Education and Learning*, vol. 5, no. 3, pp. 288-296. http://doi.org/ 10.5539/jel.v5n3p288

Almeida, F. (2018). “Strategies to perform a mixed methods study”, *European Journal of Education* *Studies*, vol. 5, no. 1, pp. 137-150.

Barrot, J. S. (2023). “Using ChatGPT for second language writing: Pitfalls and potentials”, *Assessing Writing*, vol. 57, 100745. https://doi.org/10.1016/j.asw.2023.100745

Bestgen, Y. (2023). “Measuring lexical diversity in texts: The twofold length problem”. https://arxiv.org/ftp/arxiv/papers/2307/2307.04626.pdf

Bishop, L. (2023). “A computer wrote this paper: *What ChatGPT means for education, research, and writing”*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4338981

Biswas, S. (2023). “ChatGPT and the future of medical writing”, *Radiology*, vol. 307, no. 2, e223312. https://doi.org/10.1148/radiol.223312

Chaka, C. (2023a). “Generative AI Chatbots - ChatGPT versus YouChat versus Chatsonic: Use cases of selected areas of applied English language studies”, *International Journal of Learning, Teaching and Educational Research*, vol. 22, no. 6, pp. 1-19. https://doi.org/10.26803/ijlter.22.6.1

Chaka, C. (2023b). “Stylised-facts view of fourth industrial revolution technologies impacting digital learning and workplace environments: ChatGPT and critical reflections”, *Frontiers in Education,* vol. 8, pp. 1-10. https://doi.org/10.3389/feduc.2023.1150499.

Chaka, C. (2023c). “Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools”, *Journal of Applied Learning and Teaching,* vol. 6, no. 2, pp. 1-11. https://doi.org/10.37074/jalt.2023.6.2.12

Chen, T-J. (2023). “ChatGPT and other artificial intelligence applications speed up scientific writing”, *Journal of the Chinese Medical Association*, vol. 86, no. 4, pp. 351-353. https://doi.org/10.1097/JCMA.0000000000000900

Chon, T. V., & Shin, D. (2020). “Direct writing, translated writing, and machine-translated writing: A text-level analysis with Coh-Metrix”, *English Teaching*, vol. 75, no. 1, pp. 25-48. https://doi.org/10.15858/engtea.75.1.202003.25.

Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2023). “Chatting and cheating: Ensuring academic integrity in the era of ChatGPT”, *Innovations in Education and Teaching International*, pp. 1–12. https:// doi.org/10.1080/14703297.2023.2190148

Davis, G., & Grierson, M. (2022). “Investigating attitudes of professional writers to GPT text generation AI based creative support tools”. https://ualresearchonline.arts.ac.uk/id/eprint/18621/1/Investigating%20attitudes%20of%20writers%20GPT%20text%20generation%20based%20creative%20support%20tools%20G%20Davis%20M%20Grierson%202021-22%20%281%29.pdf

deBoer, F. (2014. “Evaluating the comparability of two measures of lexical diversity”, *System*, vol. 47, pp. 139-145. https://doi.org/10.1016/j.system.2014.10.008

Deng, J., & Lin, Y. (2022). “The benefits and challenges of ChatGPT: An overview”, *Frontiers in Computing and Intelligent Systems*, vol. 2, no. 2, pp. 81-83. https://doi.org/10.54097/fcis.v2i2.4465

Dowell, N. M. M., & Kovanović, V. (2022). “Modeling educational discourse with natural language processing”, in Lang, C., Wise, A. F., Merceron, A., Gašević, D., & Siemens, G. (eds.), *The handbook of learning analytics*. 2nd edn. Society for Learning Analytics Research, pp. 105–119. https://doi.org/10.18608/ hla22.011

Etikan, I., Musa, S. A., & Alkassim, R. S. (2016). “Comparison of convenience sampling and purposive sampling”, *American Journal of Theoretical and Applied Statistics*, vol. 5, no. 1, pp. 1-4. http://www.sciencepublishinggroup.com/j/ajtas

Fincham, E., Whitelock-Wainwright, A., Kovanović, V., Joksimović, S., van Staalduinen, J .P., & Gašević, D. (2019) “Counting clicks is not enough: Validating a theorized model of engagement in learning analytics”, in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pp. 501-510. http://dx.doi.org/10.1145/3303772.3303775

Gibson, A., & Shibani, A. (2022). “Natural language processing-writing analytics”, in Lang, C., Wise, A. F., Merceron, A., Gašević, D., & Siemens, G. (eds.), *The handbook of learning analytics*. 2nd edn. Vancouver: pp. 96–103. https://opus.lib.uts.edu.au/handle/10453/156746

Halcomb, E. J., & Hickman, L. (2015). “Mixed methods research”. https://ro.uow.edu.au/smhpapers/2656

Hariri, W. (2023). “Unlocking the potential of ChatGPT: A comprehensive exploration of its applications, advantages, limitations, and future directions in natural language processing”. https://doi.org/10.48550/arXiv.2304.02017

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Alexander Trautsch, A. (2023). “A large-scale comparison of human-written versus ChatGPT-generated essays”, *Nature Portfolio*, vol. 13, no.18617, pp. 1-11. https://doi.org/10.1038/s41598-023-45644-9

Howie, A. (2022). “Analysis of the correlation between the lexical profile and Coh-Metrix 3.0 text easability and readability indices of the Korean CSAT from 1994–2022”. https://digitalcommons.cwu.edu/etd/1784

Ifelebuegu, A. O., Kulume, P., & Cherukut, P. (2023). “Chatbots and AI in Education (AIEd) tools: The good, the bad, and the ugly”, *Journal of Applied Learning & Teaching*, vol. 6, no. 2, pp. 1-15. https://doi.org/10.37074/jalt.2023.6.2.29

Ikram, A., & Castle, B. (2020). “Automated essay scoring (AES); a semantic analysis inspired machine learning approach: An automated essay scoring system using semantic analysis and machine learning is presented in this research”, in *Proceedings of the 12th International Conference on Education Technology and Computers*. Pp. 147-151. https://doi.org/10.1145/3436756.3437036

Jafarie, M., & Tabrizi, H. (2022). “The utility of Coh-Metrix application for the selection of core texts: Trialling of texts for students of computer sciences”, *Journal of New Advances in English Language Teaching and Applied Linguistics*, vol. 4. No. 2, pp. 1029-1050. https://doi.org/10.22034/jeltal.2022.4.2.11

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). "ChatGPT for good? On opportunities and challenges of large language models for education", *Learning and Individual Differences*, vol. 103, no. 102274. https://doi.org/10.1016/j.lindif.2023.102274

Khaldi, K. (2017). "Quantitative, qualitative or mixed research: which research paradigm to use?" *Journal of Educational and Social Research*, vol. 7, no. 2, pp. 15-24. https://doi.org/10.5901/jesr.2017.v7n2p15.

Khare, Y. (2023). "Prompt engineering: Rising lucrative career path AI chatbots age. https://www.analyticsvidhya.com/blog/2023/04/prompt-engineering-rising-lucrative-career-path-ai-chatbots-age/

Kim, J. (2022). "The use of cohesive devices in Korean EFL writing across different proficiency levels", *Korean Journal of English Language and Literature*, vol. 22, pp. 1078- https://doi.org/10.15738/kjell.22..202210.1078

Kitamura, F. C. (2023). "ChatGPT is shaping the future of medical writing but still requires human judgment", *Radiology*, vol. 307, no. 2, 230171. https://doi.org/10.1148/radiol.230171

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). "ChatGPT for language teaching and learning", *RELC Journal*, vol. 54, no. 2, pp. 1-14. https://doi.org/10.1177/00336882231162868

Latifi, S., & Gierl, M. (2021). "Automated scoring of junior and senior high essays using Coh-Metrix features: Implications for large-scale language testing", *Language Testing*, vol. 38, no. 1, pp. 62-85. https://doi.org/10.1177/0265532220929918

Leal, S. E., Duran, M. S., Scarton, C. E., Hartmann, N. S., & Aluísio, S. M. (2021). "NILC-Metrix: Assessing the complexity of written and spoken language in Brazilian Portuguese". https://doi.org/10.48550/arXiv.2201.03445

Lee, J. (2021). "A comparative analysis of NAEA English reading passages for middle and high school students using Coh-Metrix", *Modern English Language* (*현대영어교육*), vol. 22, no. 3, pp.12-23. https://doi.org/10.18095/meeso.2021.22.3.12

Lund, B. D., & Wang, T. (2023). Chatting about ChatGPT: how may AI and GPT impact academia and libraries?" *Library Hi Tech News*, vol. 40, no. 3, pp. 26-29. https://doi.org/10.1108/LHTN-01-2023-0009

Mahadini, M. K., Setyaningsih, E., & Sarosa, T. (2021). "Using conventional rubric and Coh-Metrix to assess EFL students' essays", *International Journal of Language Education*, vol. 5, no. 4, pp. 260-270. https://doi.org/10.26858/ijole.v5i4.19105

Mattas, P. S. (2023). "ChatGPT: A study of AI language processing and its implications", *International Journal of Research Publication and Reviews*, vol. 04, no. 02, pp. 435–440. https://doi.org/10.55248/gengpi.2023.4218

McCarthy, P. M., Kaddoura, N. W., Al-Harthy, A., Thomas, A. M., Duran, N. D., & Ahmed, K. (2022). "Corpus analysis on students' counter and support arguments in argumentative writing", *PEGEM Journal of Education and Instruction*, vol. 12, no. 1, pp. 256-271. https://doi.org/10.47750/pegegog.12.01.27

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). "Coh-Metrix: capturing linguistic features of cohesion", *Discourse Processes*, vol. 47, no. 4, pp. 292-330. https://doi.org/10.1080/01638530902959943

Mohammed, A. A., Al-ghazali, A. & Alqohfa, K. A. (2023). "Exploring ChatGPT uses in higher studies: A case study of Arab postgraduates in India", *Journal of English Studies in Arabia Felix*, vol. 2, no. 2, pp. 9-17. https://doi.org/10.56540/jesaf.v2i2.55

Nasseri, M., & Thompson, P. (2021). "Lexical density and diversity in dissertation abstracts: Revisiting English L1 vs. L2 text differences", *Assessing Writing*, vol. 47, pp. 100-511. https://doi.org/10.1016/j.asw.2020.100511

Nkhobo, T., & Chaka, C. (2023). "Syntactic pattern density, connectives, text easability, and text readability indices in students' written essays: A Coh-Metrix analysis", *Research Papers in Language Teaching and Learning*, vol. 13, no. 1, pp. 121-136. http://ijlter.org/index.php/ijlter

Rahi, S. (2017). "Research design and methods: A systematic review of research paradigms, sampling issues and instruments development", *International Journal of Economics & Management Sciences*, vol. 6, no. 2, pp. 1-5. https://doi.org/10.4172/2162-6359.1000403

Reilly, J. M., & Schneider, B. (2019). "Predicting the quality of collaborative problem solving through linguistic analysis of discourse", in *EDM 2019 - Proceedings of the 12th International Conference on Educational Data Mining*), pp. 149-157. https://eric.ed.gov/?id=ED599226

Reiter, B. (2017). "Theory and methodology of exploratory social science research", *International Journal of Science and Research*, vol. 5, no. 4, pp. 129-150.

Rudolph, J., Tan, S., & Tan, S. (2023). "ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?" *Journal of Applied Learning and Teaching,* vol. 6, no. 1, pp. 342-363. https://doi.org/10.37074/jalt.2023.6.1.9

Seals, S. M., & Shalin, V. L. (2023). "Long-form analogies generated by ChatGPT lack human-like psycholinguistic properties". https://doi.org/10.48550/arXiv.2306.04537

Shpit, E. I. (2022). "The use of Coh-Metrix by individual Russian novice writers for developing self-assessment and self-correction skills", *International Journal of English for Specific Purposes*, vol. 3, no. 1, pp. 6-33.

Straume, I., & Anson, C. (2022). "Amazement and trepidation: Implications of AI-based natural language production for the teaching of writing", *Journal of Academic Writing*, vol. 12, no. 1, pp. 1-9. https://doi.org/10.18552/joaw.v12i1.820

Su, Y., Lin, Y., & Lai, C. (2023). "Collaborating with ChatGPT in argumentative writing classrooms", *Assessing Writing*, vol. 57, 100752.

Swedberg, R. (2020). "Exploratory research", in Elman, C., Geering, J., & Mahoney, J. (eds.), *The production of knowledge: Enhancing progress in social science*. Cambridge: Cambridge University Press, pp. 17-41.

Tabassum, R., Mahmood, M. A., Mahmood, R., & Haider, S. (2022). "Readability assessment in the writings of Pakistani graduate level learners with reference to Coh-Metrix", *Hayatian Journal of Linguistics and Literature*, vol. 6, no. 1, pp. 107-115. http://111.68.104.137/index.php/HJLL/article/view/66

Tam, A. (20230. “What are zero-shot prompting and few-shot prompting”. https://machinelearningmastery.com/what-are-zero-shot-prompting-and-few-shot-prompting/

Wang, J., Engelhard Jr., G., & Combs, T. (2023). “Exploring difficult-to-score essays with a hyperbolic cosine accuracy model and Coh-Metrix indices” *The Journal of Experimental Education*, vol. 91, no. 1, pp. 125-144. https://doi.org/10.1080/00220973.2021.1993774

Yildiz, M., & Yeşilyurt, S. (2021). “Effects of task complexity on text easibility and coherence of EFL learners’ narrative writing”, *Adıyaman University Journal of Educational Sciences*, vol. 11, no. 1, pp. 36-47. https://doi.org/10.17984/adyuebd.839236

Zhou, T., Cao, S., Zhou, S., Zhang, Y., & He, A. (2023). “Chinese Intermediate English Learners outdid ChatGPT in deep cohesion: Evidence from English narrative writing”. https://doi.org/10.48550/arXiv.2303.11812