International Journal of Education and Development using Information and Communication Technology (IJEDICT), 2023, Vol. 19, Issue 3, pp. 50-68

Comparing Measures of Syntactic and Lexical Complexity in Artificial Intelligence and L2 Human-Generated Argumentative Essays

Nomsa Zindela
University of South Africa

ABSTRACT

This study explores syntactic and lexical complexity in argumentative essays written by Artificial Intelligence (AI) and Humans (HS). First year Tswana second language (L2) learners of English in a South African University wrote the HS argumentative essays. The AI essays have been generated by ChatGPT-3.5 on the same topics. Using a multidimensional perspective to genre analysis, this study examines differences in the essays using linguistic measures of syntactic and lexical complexity. Even though not all researchers agree on the nature of linguistic differences, the extant literature links features such as syntactic and lexical complexity to writing style and quality. However, little is known about this variation in content generated by ChatGPT when compared to writing by university students studying English as a second language (L2). Findings from this study show differences in the use of complexity features between AI-generated-generated content (AIGC) and human-generated content (HGC). AIGC was found to be using more sophisticated and varied vocabulary. HGC was found to use more function words and less diverse vocabulary. The use of content words was comparable with HGC using a lot more content words than AIGC.

Keywords: *ChatGPT; AI-generated-generated content (AIGC); human-generated content (HGC); syntactic; lexical complexity; lexical density; argumentative essays.*

INTRODUCTION AND BACKGROUND

Detecting linguistic features in texts is not an easy process, yet it is an important construct for understanding differences in writing styles (Lu 2012; Wolfe-Quintero, Inagaki, & Kim 1998, McNamara, Louwerse, McCarthy, & Graesser (2010).

An important area of research that has received much attention is evaluating textual and linguistic features in university students' argumentative writing. Studies in this area have epistemic markers to measure levels of language acquisition, language proficiency, and development in learners' writing (Laufer & Nation 1995, Wolfe-Quintero et al 1998: Read 2000; Bulté & Housen 2014: Yoon & Polio 2017; Zenker & Kyle 2021). Most of this research has been conducted using learners' authentic texts that are free of social media influence and unadulterated by Artificial Intelligence. However, there is a scarcity of research that measures the same linguistic and textual features to distinguish between AI generated content and human-generated content.

Understanding variance in AIGC and HGC is the focus of this study. While a wealth of knowledge exists about how humans write, it is still unclear how AI generative models including ChatGPT write. What is clear is that the styles of AI models differ significantly from humans. AI models generate significantly higher-quality argumentative essays than humans and the linguistic diversity of Natural Language Generation (NLG) is improving faster than we can catch up. The models themselves outperform each other, for example, ChatGPT-4 outperforms ChatGPT-3 (Herbold, et al. 2023). However, little is known about the linguistic features that contribute to the difference between AI and Human-generated texts.

This study addresses this gap and makes a small contribution to an understanding of the linguistic complexity markers that make an AI output different and 'better' than that of a human. Analysing features of what is believed to be good writing is an area of interest to language practitioners who

are always searching for ways of improving learner proficiency and quality of academic writing. Saricaoglu & Atak (2022, p.57) point out that an understanding of, for example, the development of learners' academic writing skills is enhanced by analyses of syntactic features in their writing. Studies in this area have used both longitudinal and cross-sectional observations to understand issues of complexity in writing development (e.g., Bulté & Housen, 2014; Lu, 2011; Mazgutova & Kormos, 2015). These and other studies have shown that '…more complex syntactic structures within a written text can indicate more advanced-level writing skills' (Crossley, 2020; Larsen Freman, 1978 cited in Saricaoglu & Atak (2022 p. 59). The question this study addresses relates to whether the AI output, which is considered better than that of the human, contains the linguistic features that characterise good and complex writing to which every L2 learner aspires.

This study comes on the heels of the recent challenges that educators face in differentiating between students' authentic work from work generated with the help of AI bots such as the infamous/famous generative model; ChatGPT.

Large Language Models

The advent of Large Language Models (LLMs), and powered AI chatbots (e.g., OpenAI's ChatGPT, Microsoft's Bing, Google's Bard), which can produce impressive human-like and high-quality written output, is challenging educators, and researchers alike to shift the pendulum to evaluating the authenticity of written output in addition to other essay quality features. It is common knowledge that many university students are turning to ChatGPT to write essays, assignments, and online examinations. While acknowledging the efficacy of these bots in supporting students' writing, there are rising concerns among educators regarding the impact this is having on higher education in general and academic writing, in particular (Diwaker, Sharma, & Tomar 2021). Most concerning is the realisation that educators, lecturers, and college tutors are not able to distinguish GPT generated from human written content (Clark et al 2021).

Understanding linguistic features that characterise human vs AI-generated writing is, therefore, a useful exercise when one considers that many educators lack the experience and skill to make this distinction. Many educators rely on the 'naked eye' judgment, a set of predefined rubrics, and subject matter knowledge to assess essays. However, in the face of AI, the limitations of such traditional methods are becoming even more pronounced.

This study explores the linguistic features which are characteristic of the argumentative genre. The argumentative genre is of course a text used in academic writing - first developed in the first year of university – and is the skill of making an argument which cuts across almost all fields in academic writing. The study also draws from the accumulated body of knowledge regarding what constitutes high-quality writing, advanced/upper-level graduate writing, developed writing, and native speaker writing as compared to beginner writing, in the argumentative genre (see Crossley, et al 2020, 2011,2009). In the context of the study, the exploration is of AI versus human-generated essays.

Research in the rhetorical practices in an AI-dominated world is growing; findings are contradictory, and more research is needed. In a study by Liu et al. (2023) it has been shown that identifying the source of the text is very much linked to the quality level of the essay as well as the language model used to generate the argumentative essays. For example, some of the human evaluators in the Liu et al. (2023) study was able to identify low-level human essays and essays generated by the more advanced models such as text-DaVinci-003 and gpt-3 turbo, yet in a study by Clark et al. (2021), the evaluators underestimated the quality of the texts that the existing models can generate. These studies have also given insight into navigating this space. For example, Liu et al. (2023) found that repeated exposure to machine-written texts and repeated excises on evaluating AI-generated essays improved the evaluator's ability to make the distinction. In the Dou et al (2022) study, when participants were asked to detect whether an essay was authored by a human or a machine and

give a reason, they focussed on cues such as grammar and spelling, concluding that human essays had more of these and that they contained more personal experience anecdotes than the machine essays. In machine-generated essays, evaluators detected more similar examples and repetitive expressions. However, as Liu et al. (2023, p15) noted, 'even with the knowledge of the two features of machine essays …, participants are not very confident when identifying machine essays'' as the same repetitiveness and redundancies in expression were evident in human writing.

Based on the findings from the studies examined thus far, one can conclude that humans potentially can detect AIGC features, they can also identify some features of machine essays, especially after reading several AI-generated texts, but this ability is limited. Therefore, if educators are exposed to more AIGC, and are trained in using linguistic and automated computational tools, they will develop the necessary skills to assess the authenticity of students' writing, and by default improve their practice and awareness of what linguistic features to focus on to improve learners' writing proficiency.

Theoretical Framework

This study is framed around the theoretical lenses of Linguistic complexity and its subsystems; lexical and syntactic complexity as measures of linguistic performance, proficiency, and development in both first and second language contexts (Kuiken, Vedder & Gilabert 2010). The study draws on Bulté & Housen's (2014) lexical complexity theorisation which expounds that the study of metrics such as lexical diversity (e.g., type-token ratio), lexical sophistication (e.g., frequency of words beyond the 1,000 most common words), and lexical density (e.g., the ratio of lexical words per function words) are the best indicator of proficiency (Bulté & Housen, 2014, p 23). The lexical complexity theorises that the skill to communicate effectively lies in a writer's word choices. Lu (2012) states three indices for measuring text complexity; lexical density, lexical variation, and lexical sophistication. The lexical words such as *nouns, verbs, adjectives, and some adverb*s contribute to text density. Lexical density itself is indicative of mature writing, informative writing, better writing, and advanced writing (Bulté & Housen (2014). Read (2000) used the term lexical richness to refer to the same concept. To gain a comprehensive gauge of a text's lexical complexity, one must ideally explore all three indices.

For the present study which aims to identify linguistic features that potentially differentiate between AIGC and HGC, I have opted to examine syntactic and lexical complexity features. These tell us just as much about the writer's language level and quality as is necessary for an exploratory study.

Syntactic Complexity

According to Ortega (2003, p.492), syntactic complexity refers to the range of forms that surface in language production and the degree of sophistication of such forms. This makes syntactic complexity a multidimensional construct and sometimes challenging to operationalise or measure. Calculated as the ratio of clauses per T-unit (CT) and subordinate clauses per T-unit (DC/T), syntactic complexity can be used to assess writing quality (Bulté & Housen, 2014).

Several measures have been used to measure syntactic complexity, including the Mean Length of Sentences, Mean length of T- units, Mean Length of Clauses, Mean Number of clauses, and Mean number of Dependent clauses.

Measuring syntactic complexity has been made easy through corpus applications and computational tools. In this study, I use Lu's (2010, 2011) framework which addresses 14 indices at different sentence levels, the phrase, and the clause. Lu (2011) categorised the indices under five dimensions, 'length of the production unit, amount of subordination, amount of coordination, degree of phrasal sophistication, and overall sentence complexity" (Lu, 2011 pp. 43-44).

Lexical Complexity

As part of the linguistic complexity, theorisation is the notion that language production is layered in its complexity in terms of diversity/variation, density, and sophistication (Lu, 2012). Advanced writing is associated with these dimensions of how unique words are, how diverse and how frequent or infrequent, as these serve as indicators of writing quality.

Lexical density and lexical diversity are easy to confuse in that they both measure the occurrence of words and clause types in a text. Lexical diversity is measured as the proportion of different words to the total words in a text (Lu, 2012). It is usually measured using the type-token ratio (TTR). Lexical Density is measured as the proportion of lexical/content words to the total number of words - both lexical/content and grammatical/functional (Ure, 1971).

Lexical sophistication is measured as the proportion of infrequent as well as advanced words to the total number of words in the text (Read, 2000).

The lexical and syntactical measures discussed above underscore why the lexicon and syntax are an integral part of written language proficiency and complexity and that an analysis of writing whose aim is to explore writing differences and quality, should examine lexical and syntactic features. One may refer to a study by Grant & Ginther (2000) which compared written essays using type-token ratio and average word length and found that both these complexity features increase as the proficiency of the learners increase. In another study by Durrant & Brenchley (2019 p. 1950) in which they examined speech patterns in language development it was found that while other indices such as word frequency did not differ significantly with age, lexical parts of speech such as verbs and adjectives decreased with age, and the frequency of nouns increased significantly.

For the present study which examined lexical complexity measuring the lexical diversity, density and sophistication are indicators of which essays use more diverse vocabulary, more sophisticated academic vocabulary, and more grammatical structures which give more information about the topic.

LITERATURE REVIEW

ChatGPT - the Open AI chatbot

ChatGPT which stands for Generative Pre-trained Transformer, is an Artificial Intelligence (AI) chatbot created by OpenAI. Since its launch in November 2022, ChatGPT has rapidly developed and to date, it has crossed over 100 million users, 12.31% of which are from the United States. Seven countries including China, Russia, Ukraine, Iran, Venezuela, Afghanistan, and Belarus have no access to ChatGPT (Demand Sage.com, August 2023). Users of ChatGPT in South Africa are not officially listed in the Demand Sage database, as such it is not clear how widespread its use is. However, the South African education policies have not yet pronounced on the country's position on the use of the bot. The country has not banned the use of ChatGPT, nor have Higher Education institutions taken a clear policy stand on the place of ChatGPT in the South African classroom. This means its use in South Africa is unregulated.

Compared to other social media platforms such as Facebook, YouTube, and TikTok which took 4.5 years, 1.5 years, and 9 months respectively, to reach 100 million users; it has taken ChatGPT 2 months to reach the same mark. In a short space of time, OpenAI has brought out different versions of ChatGPT. The version used in this study, GPT-3, is an iteration of earlier models, featuring parameters of 175 billion compared to the 117 million and 1.5 billion of GPT-1 and 2 respectively, making it the largest language model at the time of authoring this paper. Even though GPT-4 has been released, the popular version and the one that the present study uses is GPT-3.5.

While ChatGPT is recognised as the most powerful chatbot the world has ever known, this paper argues that it cannot replace certain features of human language. Apart from the differences in the sizes of each model, mentioned above, and that the models can perform the same tasks, be it at different levels of precision and accuracy, the training data for the models has been sourced from publicly available texts on the Internet. For example, GPT-3 has been trained on data from a wide and diverse range of sources, including the Internet, books, and articles. To date, data used to train ChatGPT goes up to 2021; here-in lies the first limitations of ChatGPT.

The unprecedented growth in efficiency and usage of this Chatbot is disruptive to educational practices and information generation (Herbold et al 2023). Even top leaders and AI experts such as Elon Musk and Joshua Bengio fear the 'power' of the LLMs and have called for temporary bans on further development of more powerful models. At the same time, educators and researchers are only just beginning to conduct scientific studies to understand the capabilities of bots such as ChatGPT to match humans in academic writing. In a novel study Herbold et al (2023), have produced such research in which they compare AIGC to HGC in argumentative writing of high school students in Germany. The results of their study reveal some key factors which the present study builds upon, namely, the first AI generates high-quality argumentative essays than the essays by high school students. For example, ChatGPT uses more nominalisation and higher sentence complexity whereas humans use more models and epistemic constructions of speaker attitude than GPT. The second is that the writing styles between human and generative AI differ significantly, and the third is that the linguistic complexity of AI models is changing with linguistic diversity even though lower in humans compared to ChatGPT3, it is on a significant rise with GPT-4.

Concerns have been raised regarding the use and abuse of the AI. However, in the context of L2 writing, ChatGPT is posing a different type of threat in that learning English and mastering its use in academic writing is the major preoccupation of language experts. The biggest concern for language educators is how will they measure language proficiency and development, in the face of a freely available 'ghost' writer called ChatGPT.

The study has chosen to analyse salient syntactic and lexical features in one of the most popular text types in academic writing, the argumentative essays. With an unstated objective of detecting AI writing vs human writing, this study uses several computational analysis tools to achieve results.

Detecting AIGC vs HGC

Work in this area has been going on in different permutations. Before the release of Large Language Models, researchers were already conducting detection work to distinguish machine and human-generated texts. Detectors such as GLTR by Gehrmann, Strobelt, & Rush (2019) another by Zellers et al (2019), called Grover and one by Uchendu et al (2020) are but a few models that were trained to detect machine-generated texts. However, with the introduction of LLMs, researchers are also exploring AIGC detectors.

There is a fast-growing industry that is developing AIGC detectors such as GPTZero with built-ins to get holistic scores for how much a document is written by artificial intelligence (GPTzero 2023). Other tools of a similar nature include Originality.ai, OpenAI's AI text classifier, Unicheck; Copyleaks AI Content Detector, Writer AI Content Detector, Crossplag AI Content Detector, and Sapling AI Content. All these detectors have been developed in 2023. The recency means that these tools are either not easily available or that their viability has not been fully tested. Additionally, these detectors have been modeled and trained on first language (L1), and therefore potentially not suited for L2 contexts. As noted by Lui et al. (2023), not all detectors, if any have a 100% rate of success in detecting AIGC. There is a need for educators to familiarise themselves with statistical tools that can be used to detect the authenticity of writing.

Studies of AI vs Human Writing

Since the advent of ChatGPT, there has been a burst of studies, comments, opinions, and pronouncements on the impact that it will have and is having on society and areas such as academic writing, scientific writing, customer service, and medical operations. Some see positive gains from this innovation and some sceptics lament the negative effects including the compromise of ethical standards, academic integrity, and general authenticity of content (Lui et al 2023 p.33).

Thorp (2023) in an article cleverly entitled 'Chatgpt is fun, but not an author', argues that ChatGPT could potentially provide high accuracy on English translation and proofreading to non-native English speakers, therefore reducing the chances of their work being rejected in scientific journals. In a similar vein, van Dis et al. (2023) present what they call five priorities for research that ChatGPT offers which include 'increased efficiency and the possibility to review articles' (Liu et al (2023 p. 33).

In the context of students who are still building skills and knowledge through high school and university study, researchers are keen to explore how some of the text genres such as argumentative writing should be handled to retain authenticity and academic integrity. Claimed as the first study to conduct such a comparison is a study by four German scientists, Herbold, et al (2023), which uses a large corpus of argumentative essays by high school students that are later assessed by human experts (teachers). Drawing upon linguistic theories of fingerprinting which posit that each human has a unique way of using language to express thoughts, opinions, and ideas, these scholars selected specific linguistic features to compare (Herbold et al 2023, p. 3). In their study Herbold et al (2023) used different models of ChatGPT, GPT-3, and GPT-4 which they instructed to write an essay following the prompt 'Should students be taught to compete or to cooperate?'. The GPT-generated essays were then compared to essays by student writers. The key questions in their research were to ascertain how good GPT-3 was compared to GPT-4, how the essays compare, and what linguistic features characterise human vs AI content.

The essays in Herbold et al (2023) study, were assessed using first, a rubric with six descriptors - topic and completeness, logic and composition, expressiveness and comprehension, complexity, and vocabulary, text linking and language constructs - and then checked for the distribution of semantic complexity, sentence complexity, linguistic diversity, discourse markers, modals, nominalisation, and epistemic markers. As the researchers found, AI significantly outperformed humans in the quality of the essays (Herbold et al 2023, p. 12). The findings also noted significant differences with the AI presenting highly structured essays. Notable similarities were found in the structure of the first sentence and all the AI essays started the last paragraph with 'In conclusion'. This rigidity in structure seems to have reduced the need to use discourse markers in the essays, as such the AI essays characteristically had a low count of discourse markers compared to human essays. These results are backed up by Guo, Zhang, Wang, Nie, Ding, et al (2023) and Zhao, Zhao, Lu, Wang, Tong, Qin (2023) who also investigated the fingerprint phenomenon in ChatGPT-3 generated writings.

Significance of the Study

This study is significant in that it adds to the scholarship to explore ways of distinguishing AIGC from HGC. The exploration is conducted using argumentative essays written by first-year university students at a South African university and then compared with AI-generated essays on the same topic. The students' essays are regarded as authentic human-generated content since the data has been drawn from a corpus of almost two decades ago (2001/2) at a time when there was no Internet and social media platforms, let alone open-AI supported writing. This study is also significant in that no study to my knowledge has compared argumentative essays of first-year university students in a South African University, nor has there been a study that compares human-written essays to AI-

generated essays on the same topics. One recent study by Nkhobo & Chaka (2023) which uses essays by students in a South University, measures linguistic features by examining the lexical and syntactic nature, but it does not compare these to AI-generated content which is what the present study does.

Given the extant literature base, most of which is not older than a year, the present study fills a real gap and adds to existing writing complexity literature. Using a small exploratory corpus, the study compares features found in the AI and human-generated essays. Analysing these features provides insights into AI and the dimension it brings to writing scholarship.

Research questions or hypotheses

The following overarching questions frame this investigation:

1. How do the measures of lexical complexity compare between human and AI-generated essays?
2. How do the measures of syntactic complexity compare between human and AI-generated essays?

Research Methods

Corpus Construction

As already noted, data used in this study originated from a larger corpus of argumentative essays by Tswana learners of English in a first year South African University. The corpus, the Tswana Learner English Corpus (TLEC) was compiled in 2001/2002 and contains 501 argumentative essays. Significant about this corpus is that the essays were written prior the Internet and the proliferation of social media in South Africa and before AI and most today talked about generative models such as ChatGPT. In the TLEC each topic features several essays by different students.

For this exploratory study a sample of 40 essays is used, 20 essays by student writers and 20 by ChatGPT-3 on the same topics. The students' essays were randomly selected, and all three topics were equally represented. The students are first language speakers of Setswana, one of the eleven official South African languages. The language proficiency of these students would be low to average, and it is for that reason that ChatGPT was prompted to imagine it was writing as a student during this period.

To create ChatGPT-3 generated essays, the browser version of ChatGPT-3.5, was used to collect the 20 essays between September and November 2023. The ChatGPT-3.5 version was considered the most appropriate to use as it is the most likely version that is freely available and most likely used by students as well. Three topics were chosen from several topics in the corpus. These were the most popular topics and would therefore be representative of the nature of the corpus.

The following instruction was given to ChatGPT-3.5

> Write an essay of 400- 500 words long on the following topics:
> 1. South African soccer players should be paid more to ensure that they play in South Africa.
> 2. Poverty is the cause of the HIV/AIDS epidemic in South Africa.
> 3. The Prison system is outdated. Society should not punish its criminals but rehabilitate them.

Further, ChatGPT was instructed to write the essays as if you were a first-year university student at a South African University and not give the essays a title. The instruction given to the content generators (the human and AI) was to produce an essay of between 400 – 500 words.

It was important to modify the prompt to ensure the essays were as close to the human-generated essays as possible. Generating many essays was a challenge as at some point ChatGPT was beginning to question why it is being required to generate essays on the same topic. For this reason, I had to skip a few days before requesting more essays. For an exploratory study, 20 essays were deemed sufficient to glean insights into which lexical and syntactic features are prevalent in AIGC. The corpus used in the study is provided in Table 1 below.

Table 1: *Corpus Summary*

Corpus	No of essays	Total words/Tokens
HGC	20	9619
AIGC	20	6334
Total words	40	15953

The corpus amounted to a total of 15953 words. The average length of the essays was 300-500 words. The analysis considered the differences in length and its effect on the interpretation of results.

Research Design

This study is exploratory in nature as it aims to gain insight and understanding of a research area that has limited existing knowledge. The study investigates the lexical and syntactic complexity features in argumentative essays generated by ChatGPT-3 and undergraduate first-year university students in a South African University. It employs a quantitative approach to examine how complexity features differ between the essays. Ethical approval was granted by the ethics committee of the University from where the students studied and I as a researcher am using the same data for a bigger project.

Data Preparation and Analysis

Several tools were used to prepare and analyse the data. These were sourced from the freely available suit of linguistics analysis tools (SALAT) as provided on the Natural Language Processing site. As stated on their website, the tools run on several operating systems, and provide measures related to lexical sophistication, text cohesion, syntactic complexity, lexical diversity, grammar/mechanisms, and sentiment analysis. The tools span years of collaborative work by many linguists (For this study I used three tools. The links to the tools are provided below.

- The Tool for the Automatic Analysis of Lexical Diversity (TAALED)
- The Tool for the Automatic Analysis of Lexical Sophistication (TAALES)
- The Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC)

Syntactic complexity was assessed using the L2 Syntactic Complexity Analyser (L2SCA), a computational system for the automatic analysis of syntactic complexity developed by Lu (2010, 2011) at Pennsylvania State University. The L2ACA is a subset of a TAASSC tool. L2SCA is a

multidimensional measure of different grammatical structures at phrasal and clausal levels (non-finite clauses, noun complement clauses, pre-modifying nouns, as well as prepositional phrases as modifiers, and many more (Saricaoglu & Atak, 2022 p. 58). It is because of its multidimensionality and reliability that this tool was chosen for this study.

The data analysis was conducted in several stages. The first stage was to separately load each of the data sets onto the automated analysis tools. The output data was computed and saved in an Excel file. Each file was then interrogated to extract the desired lexical and syntactic complexity measures.

FINDINGS

The findings of the study are discussed below in the context of the research questions.

To answer the first question of the study, I calculated the lexical complexity indices of diversity, sophistication, and density.

Lexical Diversity

As noted earlier lexical diversity (LD) is one of the indicators of lexical complexity. LD has been used in several studies related to L2 writing quality (e.g., Crossley, Salsbury, McNamara, & Jarvis, 2011; Wolfe-Quintero, Inagaki, & Kim 1998). In the present study, the TAALED 1.3.1 tool was used.

Measured through the Type-Token Ratio, LD indices are susceptible to text length. The traditional LD measures such as TTR: simple TTR, TTR, Root TTR, and Log TTR were used by Hess, Sefton, & Landry (1986) in a study analysing 50 utterances of children's spoken language. Their results suggested that none of these TTR measures were stable. The two most relevant studies for our purpose is that by McCarthy & Jarvis (2007, 2010) who used a corpus with 9 texts sampled from a wide range of genres to test the stability of MTLD for text length. Their findings did not show a significant correlation between values and text length even with texts as short as 100 tokens. The other study is by Zenker F. and K. (2021) who analysed a large corpus of essays of various lengths written by native and non-native speakers of English from 10 countries. As noted by Zenker F. and K. (2021) the sampled texts in McCarthy and Jarvis's (2007, 2010) study do not include L2 texts, as does their study, however, they were confident in the stability of this measure.

For the present study, I opted for the MTLD (Measure of Textual Lexical Diversity, or LDAT, Lexical Diversity Assessment Tool) to calculate the LD of the essays. This measure is derived from the average length of continuous text units above a certain Type-Token Ratio. Figure 1 below shows the calculated Lexical diversity of the two sets of argumentative essays.

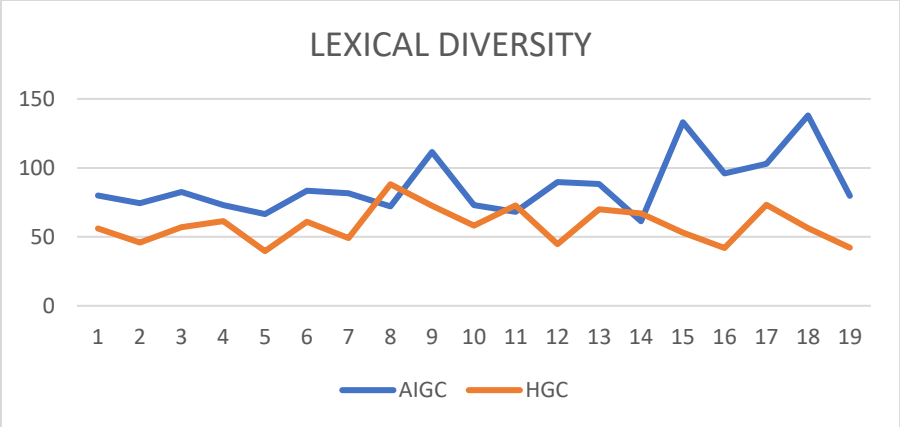

Figure 1: Lexical Diversity (MTLD) measure of the AIGC and HGC

The results show a high LD count for the AI-generated essays compared to the human essays. The result contradicts what Reviriego et al (2023) found when measuring vocabulary diversity and lexical richness in different tasks performed in TOEFL, Finance, Medicine, computing, and open-ended questions. They found that AI used few vocabularies, and the lexical richness was lower than in the tasks performed by humans. However, their study was preliminary and requires more conditions to confirm the findings, for example, there was no control on the type of tasks and the version of ChatGPT used. The results from this study may be small and not easy to generalise but are based on data that has been controlled for task type and ChatGPT version for generating content.

Lexical Sophistication

To check the level of sophistication, I used the Tool for the Automatic Analyses of Lexical Sophistication, TALAES. The measure of sophistication in this tool allows for a choice from different sub-corpora of British National Corpus (written and spoken) and other corpora. For this analysis, I chose the BNC academic written option. This is in line with what South African education follows - British English as opposed to US English. As noted in Kyle & Crossley (2015, p. 766) TAALES offers a series of indices based on several frequency logarithms, which when combined show the frequency (Freq), the register range, and the word range of the vocabulary employed by the students when compared with well-known corpora. In this tool, frequency scores are calculated by dividing the sum of the frequency scores for the tokens in a text by the number of tokens in the text that received a frequency score. The scores for all words (AW), content words (CW), and function words (FW) are calculated.

The mean scores of the lexical sophistication for the present data are shown in Figure 2. The sophistication was calculated based on the BNC frequency of content words. Content words are a good indicator of the range of vocabulary the writers employed.

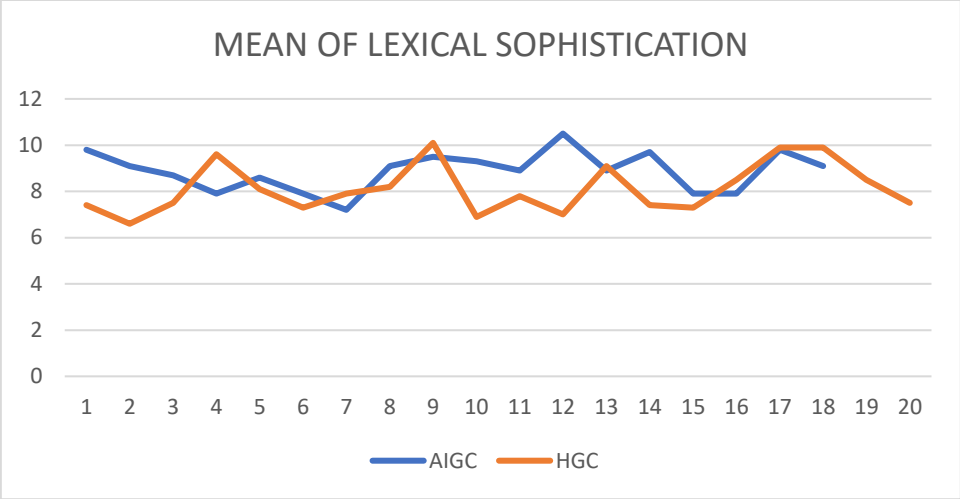

Figure 2: Mean of Lexical Sophistication

Even though the graph shows some overlaps, again the AIGC maintains a higher frequency of sophistication in terms of content words**.**

Lexical Density

The next step of the analysis was to calculate the lexical density of the texts to establish the level of the informative nature of the essays. The lexical density was measured through the type-token ratio. This meant dividing the total number of types by the total number of tokens. Tokens are all the words in a passage, while types are words that are different from each other. The lexical density when calculated using Ure's (1971) method yielded the following results shown in Table 2.

Table 2: *Lexical Density Measure*

Corpus	Total token Types	Total words/Tokens	Lexical Density
HGC	3500	9619	42%
AIGC	2935	6334	46%

The lexical Density of the AIGC is again slightly higher than the HGC, another indication that the AIGC texts are richer in terms of word choice and content. Despite that, the number of tokens is higher in the HGC, meaning there could be more sentence production, and the richness of the words is low.

To further examine and confirm the results from the complexity measures, I extracted from the TAAED output the measures of content and function words shown in Figure 3 and Figure 4 respectively.

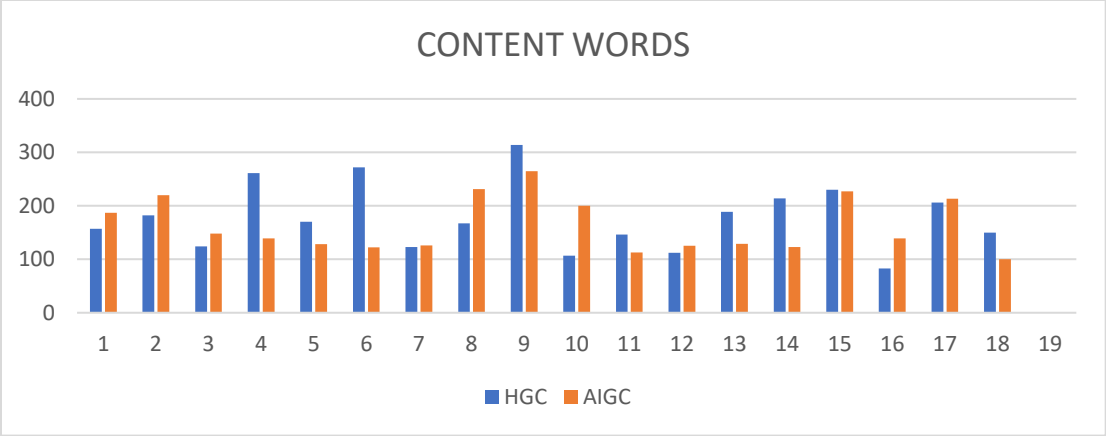

Figure 3: Measure of content word count from each data set

The data in Figure 3 clearly shows the narrow margin between counts of density in terms of content words. However, the picture is slightly different with function words as shown in Figure 4.

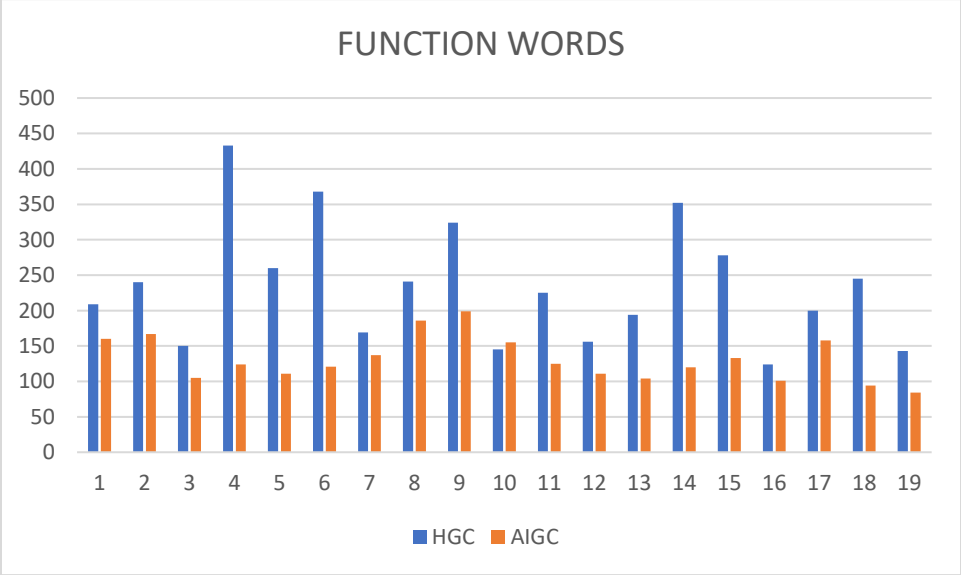

Figure 4: Measure of Function words by data set

The graph above shows a marked difference between AIGC and HGC in terms of function words with the students producing much more function words than ChatGPT. The student writer is clearly preoccupied with ensuring that the grammar and structure conform to the expectations of the teacher. In so doing they overcompensates by using more functional words.

To answer the second question, the syntactic complexity using TAASSC 1.1.9 was conducted. Data was entered and only the 14 L2SCA index was chosen. The index suit is shown in Table 2 below.

Table 2: *14 syntactic complexity measures (Lu, 2010, 2011)*

	Index Abbreviation	***Index Name***
Length of production unit	*MLS*	*The mean length of the sentence*
	MLT	*The mean length of the T-unit*
	MLC	*The mean length of the clause*
Sentence complexity subordination	*C/S*	*Clause per sentence*
	VP/T	*Verb phrase per unit*
	C/T	*Clause per T-unit*
	DC/C	*Dependent clause per clause*
	DC/T	*Dependent clause per T-unit*
	T/S	*T-unit per sentence*
	CT/T	*Complex T-unit ratio*
Coordination	*CP/T*	*Coordinate phrase per T-unit*
	CP/C	*Coordinate phrase per clause*
	CN/T	*Complex nominal per T-unit*
Particular Structures	*CN/C*	*Complex nominal per clause*

W = word number, S = Sentence, C = clause, and T= T-unit.

The analysis using the L2SCA syntactic complexity suit unfortunately did not yield the desired output. For example, while the run of the HGC (students' essays) yielded all categories, the run on the AIGC (ChatGPT-3.5 essays) yielded all categories but gave 0,00 values on all indices except in the Mean Length of sentence (MLS). The decision was to use one index. Using one index is common in previous research. Analysis was therefore confined to discussing the MLS index and the results are worthy of the task. As explained in Kyle (2016 p.10) the mean length of sentence (MLS) index is simply the number of words in a sentence. As a measure it has advantages compared to the T-unit counts that we could not find in our computation. According to Kyle (2016 pp.10-11), the MLS is less ambiguous and therefore can be counted quickly and reliably. "MLS is strongly correlated with MLTU". As noted in Kyle (2016) several studies have demonstrated positive relationships between MLS and language proficiency (see Wolfe-Quintero et al., 1998; Ortega, 2003).

The mean length of sentences was calculated using the lowest MLS and the highest from each data set. The standard deviation was also calculated to establish the distance from the mean.

Table 3*: Descriptive stats of syntactic complexity index of Mean Length of Sentences*

		Min	**Max**	**mean**	**STD.**
MLS	HGC	12.77	51.44	24.44	11.32
	AIGC	15.10	23.43	23.23	2.19

The data in Table 3 shows that the MLS for AIGC is smaller than that of AIGC which means there are differences in terms of the mean length of sentences. The HGC seems to have longer sentences than the AIGC. What this means in terms of the number of phrases and clauses may require further investigation. This finding relates to Bulte & Housen's (2014) study which measured syntactic changes in the writing of L2 learners over a course period. They found that by the end of

the semester, the L2 essays had an increased number of clauses, and sentences as well as T-units.

The standard deviation, which is the average amount of variability in the data, tells us how far each value is from the mean. For the present data, the deviation is wider for the HGC at 11.32 spread from the mean than it is at 2.19 for the AIGC. It is not surprising that the deviation is small for the AIGC, because the AIGC essays examined in this study use similar generic, and homogenous styles such as always ending an essay with 'in conclusion' and beginning new paragraphs with several conjunctive adjuncts such as 'firstly', 'secondly', 'moreover'.

DISCUSSION AND CONCLUSION

The first conclusion to be drawn from these results is the differences in the quantity of language output. The students produced relatively more words than the ChatGPT-3.5. This may be interpreted in several ways. Language proficiency is a factor in that less proficient writers use more words to express what more proficient writers would express in fewer words. This aligns with the mean length of syntactic structures (sentences) where the HGC had slightly longer sentences than the AIGC with its use of more succinct and slightly sophisticated words. In addition, the results show a higher output of function words by the human writers compared to the ChatGPT. This output is clearly affected by essay length as most ChatGPT essays were shorter despite the instruction to write 400-500 words essays.

As hypothesized for the first question, in most cases the quality of AI-generated content as evidenced in the lexical density, sophistication, and diversity is higher than that of human-generated content. When a text is at above 40% density level, then it is readable and balanced in terms of information delivery (Ure, 1971). As noted by Johannson (2008) since content words carry the bulk of the information, they are an important indicator of textual richness.

Despite the differences, the scores based on the indices examined in this study do not differ significantly. For example, the lexical density was averaged between 40% - 49%. Of interest is what Read (2000) has noted about lexical density, if it is over 50% the text is most likely produced by native speakers and not L2. While the lexical density of the AIGC is higher than the human HGC, it is of interest that it is below the native speaker benchmark of 50%. As noted earlier, while the high quality of written output by ChatGPT is a reality, the 'humanness' in the language output is a moot point and one that linguists must explore further.

Concerning lexical diversity, the findings, not surprisingly, also show that in the sample essays by AI, the range of vocabulary is wider, and the language repertoire is larger. The Human-generated essays show evidence of larger language output which contains more function words, these being words that focus on textual features of cohesion rather than on the content and the message.

The second question was to explore syntactic complexity which included the amount of output, phrase and clause complexity, and coordination. However, not all indices were explored. Based on the findings on the MLS (mean length of sentences), the mean, and standard deviation, it can be concluded that content produced by ChatGPT is mostly homogenous and predictable, whereas content produced by student writers tends to be voluminous, semantically sound but not lexical rich and diverse.

This study provides a first window to what is available for university educators in L2 context to further explore their student essays and get an understanding how these differ from AI generated one. Educators may consider for example using ChatGPT to generate essays prior an assessment on the same topic and get students to benchmark their essays using the measures of density,

sophistication and diversity against the AI essays. Such an exercise would raise the student's awareness of the capabilities of AI and what to learn and not to learn from the experience.

LIMITATIONS

The study would have benefitted from using human assessors and larger data. While the results give useful insights into what can be measured to establish differences, the results may not be easily generalisable. Most importantly what this study has shown is that each of the linguistic indices requires further in-depth research even if it means a replication of previous work to include an analysis of language as used by the new 'kids on the block' ; generative models and AI bots.

CONCLUSION

The argumentative text is ideal ground for providing insight into students' writing. This study investigated the features that distinguish AI-generated from human-written argumentative essays. A combination of features was explored, namely, lexical complexity features and syntactic complexity features of self-mention. Data used was drawn from a corpus of authentic texts written by university L2 students in 2001/2002, this being a period of pre-Internet and AI. The second set of data were essays generated by ChatGPT on the same topics.

The study found that: the text generated by humans and AI are significantly different; the AI-generated texts had higher quality output compared to the human content, the texts were lexically richer and used varied vocabulary. However, there was no significant difference in the language output about content words.

The study concludes that if educators can pay attention to linguistic features that focus on linguistic complexity, and richness, they stand a better chance of keeping pace with the fast-developing world of AI. This study makes a valuable contribution to research in academic writing and assessment which confronts what Naidu & Sevnarayan (2023) term the disruptive innovation; ChatGPT. Based on the results of this study ChatGPT still must prove whether it is a robot, a non-native user, or an L2 user of English or a new user.

REFERENCES

Bulté, B., & Housen, A. 2014. "Conceptualizing and measuring short-term changes in L2 writing complexity." *Journal of Second Language Writing*, vol. 26; pp. 42-65.

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. 2021. "All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text." *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7282–7296.

Creswell, J. W. 2009. *Research Design: Qualitative, Quantitative and Mixed Method Approaches* (3rd ed.). Thousand Oaks, CA: SAGE Publications.

Crossley, S. A., & McNamara, D. S. 2009. Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing*, vol. 18, pp.119-135.

Crossley, S. A., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *IJCEELL,* vol. 21, pp. 170-191.

Crossley, S. A., Louwerse, M., McCarthy, P. M., & McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *Modern Language Journal*, vol. 91, pp.15-30.

Demand Sage.com, August 2023.

Diwaker, C., Sharma, A., & Tomar, P. (2021). Artificial Intelligence in Higher Education and Learning: Impact of AI Technologies on Teaching, Learning, and Research in Higher Education, 11 pages.

Dou, Y., Forbes, M., Koncel-Kedziorski, R., Smith, N. A., & Choi, Y. 2022. "Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text." *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7250–7274.

Durrant & M. Brenchley (2019). Development of vocabulary sophistication across genres in English children's writing. *Reading and Writing*, vol. 32, no. 8, pp. 1927-1953.

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., and Pearson, A. T. 2022. "Comparing Scientific Abstracts Generated by ChatGPT to Original Abstracts Using an Artificial Intelligence Output Detector, Plagiarism Detector, and Blinded Human Reviewers." *bioRxiv*.

Gehrmann, S., Strobelt, H., and Rush, A. M. 2019. "GLTR: Statistical Detection and Visualization of Generated Text." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

GPT-3. 2020. "A Robot Wrote This Entire Article. Does That Scare You, Human?" *The Guardian*, September 8, 2020.

Grant, L., & Ginther, A. (2000). Using Computer-Tagged Linguistic Features to Describe L2 Writing Differences. *Journal of Second Language Writing*. Vol. 9, pp.123-145. 10.1016/S1060-3743(00)00019-9.

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. 2023. "How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection."

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. 2023. "AI, write an essay for me: A large-scale comparison of human-written versus ChatGPT-generated essays."

Hess, C.W. Sefton, K.M., &. Landry, R.G. (1986). Sample size and type-token ratios for oral language of preschool children *Journal of Speech and Hearing Research*, vol. 29, no. 1, pp. 129-134

Housen, A. 2002. "A corpus-based study of the L2-acquisition of the English verb system." *Computer learner corpora, second language acquisition, and foreign language teaching*, vol. 6, pp. 2002–77.

Hunt, K. 1965. Grammatical structures are written at grade levels. Urbana: NCTE.

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: a developmental perspective. Lund University, Dept. of Linguistics and Phonetics, Working Papers 53 (2008), pp. 61-79

Kormos J. 2011 Task Complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing*, vol. 20, no. 2, pp.148-161.

Kyle, K. & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, vol. 49, no. 4, pp. 757-786.

Kyle, K. 2016. Measuring syntactic development in L2 writing: Fine-grained indices of syntactic complexity and usage-based indices of syntactic sophistication (Doctoral Dissertation).

Laufer, B., and Nation, P. 1995. "Vocabulary size and use: Lexical richness in L2 production." *Applied Linguistics*, vol. 16, no. 3, pp. 307-322.

Larsen-Freeman, D. 1978. An ESL index of development. *TESOL Quarterly*, vol. 12, no. 4, pp. 439-48.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer,

Liu, Y., Zhang, W., Zhao, X., Zhang, Y., Zhang, Z., Yue, S., Cheng, X., & Hu, H. (2023). "ArguGPT: Evaluating, Understanding and Identifying Argumentative Essays Generated by GPT Models." *Computational Linguistics*, vol. 1, no. 1, 1048.

Lu, X. 2010. "Automatic analysis of syntactic complexity in second language writing." *International Journal of Corpus Linguistics*, vol. 15, no. 4, pp. 474-496.

Lu, X. 2011. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, vol. 45, no. 1, pp.36-62.

Lu, X. 2012. The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal,* vol. 96, no. 2, pp.190-208.

McCarthy, P. M., & Jarvis, S. 2007. vocd-D, A theoretical and empirical evaluation. Language Testing, vol. 24, no. 4, pp. 459–488.

McCarthy, P. M., & Jarvis, S. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, vol. 42, no. 2, pp. 381–392.

McNamara, D.S., Louwerse, M.M., McCarthy, P.M., & Graesser, A.C. (2010). "Coh-Metrix: Capturing linguistic features of cohesion." *Discourse Processes*, vol. 47, pp. 292-330.

Naidu, K., & Sevnarayan, K. 2023. "ChatGPT: An ever-increasing encroachment of artificial intelligence in online assessment in distance education." *Online Journal of Communication and Media Technologies*, vol. 13, no. 3, e202336. DOI: 10.30935/ojcmt/13291.

Nkhobo T., & Chaka, C. 2023. *Research Papers in Language Teaching and Learning*, vol. 13, no. 1, January 2023, pp. 121-136.

Ortega, L.. (2003). Syntactic Complexity Measures and their Relationship to L2 Proficiency: A Research Synthesis of College-level L2 Writing. *Applied Linguistics,* vol. 24. pp. 492-518. 10.1093/applin/24.4.492.

Ramoroka, B.T., 2016. "The use of interactional metadiscourse features to present a textual voice: A case study of undergraduate writing in two departments at the University of Botswana." *Reading & Writing*, vol. 8, no. 1, pp128. DOI: 10.4102/rw.v8i1.128.

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

Saricaoglu, A., & Atak, N. (2022). Syntactic complexity and lexical complexity in argumentative writing: Variation by proficiency. Novitas- ROYAL (Research on Youth and Language), vol. 16, no. 1, pp. 56–73.

Thorp, H. 2023. "ChatGPT is fun, but not an author." *Science*, vol. 379(6630), pp. 313.

Uchendu, C., Windle, R., & Blake, H. (2020). Perceived Facilitators and Barriers to Nigerian Nurses' Engagement in Health Promoting Behaviors: A Socio-Ecological Model Approach. *Int J Environ Res Public Health*. 2020 Feb 18; vol. 17, no. 4, pp.1314. doi: 10.3390/ijerph17041314. PMID: 32085607; PMCID: PMC7068510.

Ure, J. (1971) Lexical Density and Register Differentiation. *Contemporary Educational Psychology*, vol. 5, pp. 96-104.

van Dis, E.A.M., Bollen, J., Zuidema, W., van Rooij, R., Bockting, C.L.(2023). ChatGPT: five priorities for research. *Nature.* 2023 Feb;614(7947):pp. 224-226. doi: 10.1038/d41586-023-00288-7. PMID: 36737653.

Van Rooy, B. 2009. Tswana learner English corpus. (In Granger, S., Dagneaux, E., Meunier, F. & Paquot, M., (eds). International corpus of learner English handbook and CDROM. Version 2. pp. 198-204.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. 1998. Second Language Development In Writing: Measures Of Fluency, Accuracy, And Complexity. *Studies in Second Language Acquisition*, vol. 23, no. 3, pp. 423–425.

Yoon, H. J. 2017. Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality". *System*.

Youseff, A. 2019. Syntactic complexity and lexical diversity in English conference abstracts: investigating cross-disciplinary effects with native speaker baseline. *HERMS*, vol. 8, no. 2, pp. 33-70.

Zellers, R. Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending Against Neural Fake News. *In Advances in Neural Information Processing Systems*, pp. 9054–9065.

Zenker, F. and Kyle, K. 2021. Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, vol. 47, 100505.

Zhai, X. 2022. ChatGPT user experience: Implications for education. *SSRN Electron*. J.

Zhao, W., Zhao, Y., Lu, X., Wang, S., Tong, Y., & Qin, B. (2023) Is ChatGPT Equipped with Emotional Dialogue Capabilities? Preprint: https://arxiv.org/abs/2304.09582