ADVANCE SOCIAL SCIENCE ARCHIVE JOURNAL
Available Online: https://assajournal.com
Vol. 05 No. 02. April-June 2026.Page# 846-858
Print ISSN: 3006-2497 Online ISSN: 3006-2500
Platform & Workflow by: Open Journal Systems

A Comparative Linguistic Analysis of Lexical Diversity in AI-Generated and Human-Written Academic Research

Suriya Farooq
BS English, Fazaia College of Education for Women Peshawar Affiliated with Air University, Islamabad
suriyafarooq837@gmail.com
Tabassum Naz
Lecturer English, Fazaia College of Education for Women Peshawar, Affiliated with Air University, Islamabad
tabassumnaz@fcwp.edu.pk
Palwasha Habib
M.Phil English Scholar, University of Swat
zarimehr35@gmail.com
Kanwal Sajjad
M.Phil English Scholar, Lahore Leads University
kanwalsajjad024@gmail.com

Abstract
The study compares the lexical diversity of AI generated academic research and human written academic research and provides a comparative linguistic analysis. As AI assistants like ChatGPT are becoming more prevalent in the world of academic writing, issues of authenticity, vocabulary usage, and linguistic originality are gaining in prominence. The main goal of this research is to analyze differences and similarities with regard to the lexical diversity, the lexical sophistication and the vocabulary variation between AI-generated and human written academic texts. The methodology adopted in the study is a corpus-based qualitative and quantitative research approach, and the samples from the academic field are selected from the products of human research and AI systems, which are then analyzed by linguistic and computational methods. Linguistic richness and variation are measured through various lexical indices such as type-token ratio, lexical density and vocabulary sophistication. It will be revealed that the texts created with AI are grammatically consistent and formal, however, they might lack in the domain of the context in terms of creativity and repetition of lexical patterns as opposed to human written research. On the other hand, texts produced by human writers often will show more lexical diversity, context sensitivity, and subtle text choices. The study has implications for applied and corpus linguistics and digital humanities as it brings a better understanding of the relationship between AI and academic discourse that has developed over time. It also explores the pedagogical and ethical issues arising from AI-supported academic writing in modern education. Also, the study highlights the increasing need for digital literacy and critical thinking skills when assessing academic materials generated by AI. It aims to offer useful suggestions to teachers, researchers and policy makers on how to responsibly incorporate AI technologies in academic environments. The study also seeks to inspire future research in the field of linguistics to explore how AI will affect language and scholarly communication over time.

Keywords: *Artificial Intelligence, ChatGPT, Lexical Diversity, Academic Writing, Corpus Linguistics, AI-Generated Texts, Human-Written Research, Applied Linguistics, Vocabulary Sophistication, Digital Humanities*

1. Introduction

Artificial Intelligence has made substantial strides over the last couple of years, influencing contemporary academic writing, especially in the domains of English linguistics and higher education research. The most significant advancement is the introduction of AI-driven writing tools like ChatGPT, which are highly valuable for creating research papers, academic drafts, and summaries as well as crafting essays. With their efficiency, grammatical correctness, and structure, these tools have added a new dimension in scholarly writing. As they become more common, however, they have also given rise to pertinent academic concerns about language quality, originality, and richness of vocabulary.

Vocabulary variation and lexical diversity are essential to study from the linguistic point of view for the quality assessment of academic writing in the context of language variation. Lexical diversity relates to the variety and complexity of vocabulary used in a text, and is a good measure of language proficiency and academic achievement (Malvern et al. 23). Thus, the issue of comparing the AI-generated and human produced academic texts is a topic of considerable research in applied linguistics and corpus-based research.

Context and Background

Modern universities heavily rely on Artificial Intelligence in writing and research activities. AI tools support students and researchers by generating content, providing grammatical assistance, and optimizing structural organization. These technologies have the potential to facilitate academic productivity, but also pose challenges for linguistic authenticity and intellectual originality.

The main features of human-written academic texts are critical thinking, context and a variety of lexis. Unlike human experience, AI-generated texts are created by statistical prediction models that are trained on large sets of linguistic data. In light of these findings, researchers have started to doubt whether AI systems can match the depth of vocabulary and stylistic diversity of human scholarly texts.

Corpus linguistics is a tool that has become a valuable asset to these issues and can be used to systematically analyse vocabulary patterns, lexical density and language variation in large data sets (Sinclair 15). The density of the academic text is usually assessed by the use of measures like type-token ratio and lexical density (McCarthy and Jarvis 382).

Research Gap

Although there is a vast amount of research into artificial intelligence in the field of education, the majority of studies are limited to grammatical accuracy, plagiarism detection, ethical issues, and overall writing assistance. There is less focus on deeper linguistic aspects like lexical diversity, vocabulary sophistication, and stylistic variation in academic AI-generated texts.

Moreover, numerous studies examine the use of AI writing in either technical or didactic terms, but not from the perspective of a corpus linguistic approach for systematic comparisons. The body of research is lacking in comprehensive studies on lexical patterns of AI-generated and human-authored academic research, specifically employing a quantitative and qualitative analysis. This study aims to fill this gap by drawing specific attention to the lexical diversity, vocabulary richness and stylistic flexibility of artificial intelligence-generated and human-written academic texts.

Research Objectives

1. To investigate how lexically diverse is the text generated by the AI compared to the text written by humans in the academic context.
2. To make a comparison between the lexical density and the vocabulary sophistication of both kinds of writing.
3. Identify lexical patterns that repeat in AI-generated academic texts.
4. To examine differences in style and flexibility in context of writing by humans and AI.
5. To discuss implications of AI-assisted writing to academic communication.

Research Questions

1. What is the difference between lexical diversity of AI-generated academic research and human-written academic research?
2. How is the lexical density and the vocabulary sophistication of the two sets of texts different?
3. How much repetitive lexical patterns do the AI-generated texts exhibit?
4. How is variation and flexibility different in the two kinds of writing?
5. What do pedagogical and linguistic implications of AI-assisted academic writing look like?

Scope and significance of the Study

The present study is centered on the comparative investigation of the lexical diversity of Artificial Intelligence (AI)-generated and human-generated academic writing in the discipline of English linguistics. The scope is restricted to academic texts in the English language, mainly in the field of Applied Linguistics and related fields. The importance of this research can be seen from the contribution of the research to Applied Linguistics, Corpus Linguistics, and Digital Humanities. It offers important insights into the changing landscape of academic communication and the impact of AI on the use of vocabulary. This research could assist in creating guidelines for the responsible use of AI in academic settings for educators, researchers, and policymakers.

Additionally, the study highlights the importance of digital literacy, critical thinking, and ethical awareness in evaluating AI-generated academic content. It also helps to continue debates about originality, authorship, and linguistic identity in the era of artificial intelligence.

2. Literature Review

AI's swift advancement has had a profound impact on academic writing, research methods, and language analysis. AI-driven writing tools, like ChatGPT, are commonplace in the world of education and research in recent years. All of these are designed to create coherent texts, fix grammar and help users create texts quickly and efficiently. As their numbers continue to grow, however, there has been scholarly interest in assessing their linguistic quality, especially in comparison to the linguistic product of the human writer, the academic discourse. The academic writing task is typically seen as a cognitively complex task that demonstrates one's critical thinking, lexicization, and knowledge of a discipline. In human written academic text, contextual awareness, argument development and a variety of vocabulary usage are evident. On the other hand, AI-generated texts are created using statistical prediction models, which are built with machine learning.AI generated texts, by contrast, are generated by machine learning models that rely on statistical prediction. Bender et al., in their turn, claim that large language models produce text that lacks a true understanding of meaning, instead using pattern recognition in their training data (612). This separation has sparked concerns about the ability of AI to produce as sophisticated of a piece of writing as a human, in terms of its lexical content.

Lexical diversity is an important concept in applied linguistics and corpus linguistics. It's the variety and diversity of words in a text, and it's a good measure of writing quality. According to Malvern et al. (23), lexical diversity is a measure of vocabulary variation, which indicates linguistic competence and richness of the text. In the same way, the authors, McCarthy and Jarvis, note

that type-token ratio and lexical density are common indicators of the use of vocabulary in academic writing (382).

The methodological approach of corpus linguistics is used to analyze large quantities of written text. Corpus-based research enables Sinclair to detect patterns of vocabulary occurrence and repetition, and of linguistic structure among texts (15). This method is especially helpful when comparing AI-generated academic writing with human writing, allowing for systematic measurement of lexical variation and frequency distribution.

Academic Writing and Linguistic Concerns

AI has brought opportunities and challenges to academic writing. The use of AI tools helps students and researchers improve their grammatical accuracy, coherence, and structure. For second language learners, who have problems with academic English conventions, these benefits are especially relevant. But educators have expressed worries about overreliance on AI-generated writing and how it affects students' ability to develop their own language.

Floridi and Chiriatti suggest that AI-authored texts could mimic the style of materials used for training, but without true depth of meaning or context (686). This restriction indicates that whilst AI-generated content is grammatically correct, it may not accurately reflect the nuances of academic writing.

The discourse theory by Fairclough also elaborates that the language is influenced by social, ideological and institutional factors (37). In this way, academic texts written by humans express their own thinking and personal point of view, while AI-generated texts may only capture generic linguistic features without any personal or ideological content. The difference is significant for the analysis of the lexical diversity and stylistic variation.

Concept of Lexical Diversity in Academic Discourse

Lexical diversity is an important aspect of assessing the quality of academic writing. It demonstrates writer's use of a wide range of vocabulary to convey sophisticated concepts. The high lexical diversity is often correlated with high language proficiency level and high academic competence. Based on statistical measures of lexical diversity, such as a type-token ratio, HD-D and lexical density, which quantify vocabulary variation in texts, McCarthy and Jarvis suggest that the lexical diversity of texts can be measured (384). The measures are commonly employed in corpus linguistics research for analyzing or contrasting writing styles based on different genres and authors.

Typically, human written text (such as academic texts) shows higher variability of their lexicon, which is characterized by cognitive flexibility, domain competence, and context sensitivity. Texts are composed in a manner that suits the rhetorical purpose, the audience, and the writer's arguments. Then again, artificial intelligence writing can use the same type of lexical constructions and common academic expressions because the models are algorithmically predicted.

Lexical diversity is also a direct measurement of cognitive and linguistic development, as Malvern et al. point out (54). Thus, it is possible that the lexical difference between AI texts and human texts reflects greater differences in the language production processes.

Ethical and Academic Concerns

The applications of AI in academic writing have sparked significant ethical concerns, notably regarding originality, authorship, and academic integrity. Cotton et al. emphasize that institutions are grappling with greater difficulty separating content created by AI from human writing, and potentially using AI in academic environments in an unauthorized manner (4). Consequently, universities are creating policies that govern the use of AI for academic writing. Therefore, the universities are formulating policies governing the use of AI for academic writing.

Ethically, overreliance on AI-generated text could hinder students from cultivating critical thinking and independent voice. Ethically, overreliance on AI-generated text could lead to students missing opportunities to develop critical thinking and independent voice. AI tools can be used to aid learning but shouldn't take the place of thinking when it comes to academic writing. Rather, they need to be used as ancillary devices that help promote language accuracy and productivity.

Summary of Literature

Based on the literature reviewed, AI has shown to have a profound impact on academic writing, especially with regards to its efficiency and accessibility. But there are concerns about the variety of words, originality, and the variation of style in AI-generated texts. Previous research has indicated that the language of academic writing written by humans tends to be more varied and less rigid in its use of language than machine-generated academic writing. Based on these results, however, the field of systematic corpus-based comparative study of the lexical diversity of AI-generated and human-written academic research seems to be lacking. The purpose of this study is to fill this gap with a systematic linguistic study of lexical variation, lexical density, and stylistic patterns in the two kinds of academic texts.

3. Research Methodology

Research Design

The research design in this study is a mixed method research, combining both qualitative and quantitative approaches to analyze the lexical diversity found in academic research generated by the AI and written by humans. Mixed-method research is known for its use of numerical measurement and interpretive analysis to gain a fuller understanding of the linguistic phenomena (Creswell 14). The quantitative dimension is concerned with the statistical measurement of lexical variation, whereas the qualitative dimension deals with the meaning of the context, stylistic features and discourse-level aspects.

Combining both approaches is especially crucial in linguistic research as numbers do not convey the full sense of lexical diversity. Use of vocabulary is not only measurable, but also indicative of cognitive, contextual and communicative processes in Academic Writing.

Data Selection

Two similar corpora are used – namely, AI-generated academic texts and human-written academic research articles. The AI generated corpus is generated by setting up a structured prompt with ChatGPT, while the human written corpus consists of peer-reviewed academic articles in English linguistics and applied linguistics. A purposive sampling method is used to guarantee that the texts selected are relevant to the research goals. According to the explanation of Etikan et al., purposive sampling will be useful if the researcher chooses the data that is rich in information and directly related to the study's objectives (2). For this reason, texts are chosen for this study using three criteria: the academic relevance, thematic similarity, and similarity in length of the word.

To create balance, there are about 20 AI-generated texts and 20 human-written articles. The size of the texts varies within the range 1,500 - 2,000 words so as to achieve uniformity in the size of the corpus. This consistency is essential for reliable lexical comparison and for the statistical validity.

Data Collection Procedure

There are two stages in the data collection process. The first stage involves generating academic text based on carefully designed prompts to topics like linguistics, discourse analysis, academic writing, etc. and education.In the first phase, AI-generated academic text is generated through

carefully designed prompts on topics like linguistics, discourse analysis, academic writing, etc. and education. The prompts are standardized to guarantee uniformity in created text.

In the second stage, the articles are retrieved from well-reputed journals, academic databases and published research papers in the field of English linguistics and they are written by human beings. In the second stage, the articles are collected from well reputed journals, academic databases, and published papers in the field of English linguistics and these articles are written by humans. These texts are all checked for academic honesty and appropriateness.

All the texts collected are then digitized and put into two corpora: texts created by AI and texts written by real people. Both the corpora are well organized and labelled for systematic analysis. This separation enables an easy and systematic comparison of the two forms of academic writing.

Data Analysis Tools and Techniques

This study uses corpus-linguistic tools and computational techniques for the analysis of lexical diversity. The following key lexical indices will be used:

- Type-token ratio (TTR)
- Lexical density
- Lexical sophistication
- Word frequency distribution
- Vocabulary range analysis

Type Token Ratio (TTR) is a measure that will tell you what percentage of the words in a text are unique. Lexical density measures the ratio of the number of content words (nouns, verbs, adjectives, adverbs) to the number of function words. Lexical sophistication is an evaluation of the use of more advanced or less common vocabulary elements. In corpus linguistics, these are generally used to gauge vocabulary richness and variation (McCarthy and Jarvis 384).

Sinclair points out that corpus-based approaches enable researchers to discover common lexical structures and regularities in large corpora, which is crucial for comparative linguistic research (15). The present research involves generating frequency lists, finding common words and computing the diversity scores for both of these corpora using computational methods.

The quantitative results are statistically analyzed to find differences in lexical diversity between the AI generated text and human text. Also, qualitative discourse analysis is employed for the purpose of exploring the use of language in context, its rhetorical style and its semantic appropriateness. This combination provides for a more comprehensive grasp of linguistic variation.

4. Theoretical Framework

The theoretical approach in this study is the application of Corpus Linguistics, Applied Linguistics and Discourse Analysis. The methodological framework of corpus linguistics is used to analyse large amounts of natural language data and to find lexical patterns in texts. The use of Fairclough's discourse analysis theory to: understand the social identity, ideology, and academic authority of language (37). In this view, human authored academic texts embody intellectual agency and context sensitivity, whereas AI generated texts are algorithmically produced discourse that is a product of probabilistic language modeling. The analysis is complemented by Halliday's systemic functional linguistics which looks at the communicative functions that language serves in social contexts. The lexical choices directly influence meaning making and text coherence, Halliday says (112). Thus, the variations in lexical richness between human and machine-generated text could be due to variations in communicative purpose and adaptation to context.

In addition, Biber's register analysis backs up the theoretical framework by indicating the nature of the variation in the linguistic system, which is dependent on purpose, audience and context (Biber 78). Formal register tends to call for precision of language, and specific terms from the discipline, so academic writing will normally require high lexical precision and more often than not, more words from the field of the academic discipline than AI texts will.

Reliability, Validity, and Trustworthiness

Carefully selected and comparable texts are used from similar academic disciplines to ensure validity in this study. Both sets of data are analysed with the same corpus size to not introduce imbalance into the analysis. It is ensured by repeated use of computational methods and standardized measurement methods. Combining quantitative lexical analysis and qualitative discourse interpretation, triangulation is applied. This renders the results even more reliable because both results at the statistics level as well as the linguistic level support each other.

In addition, there is also the inter-coder consistency in the analysis of qualitative data to prevent the bias of the researcher in interpreting the stylistic and contextual features in the data.

Ethical Considerations

In this study, ethical issues are fully respected. There is proper referencing of all text written by human beings sourced from published academic journals and no copyrighted material is misused. AI-generated texts are generated for research purposes only and not for academic submission or misrepresentation. The authors of Cotton et al. stress the need for clear ethical guidelines in academic settings to ensure academic integrity and to avoid misuses of AI (4). In this regard, the study aims to be transparent, it respects the use of data and it meets the criteria of scholarly ethics.

The research corpus does not contain any personal data, sensitive information or identifiable content. The study is done only for linguistic and academic analysis.

Limitations of the Study

While the research design is well planned, there are some limitations that must be noted. The first is that the study only deals with academic texts written in English, which might restrict transferability to academic texts in other languages or multilingual settings. Second, the size of the corpus is limited, because of time and resource constraints.

Third, AI-powered writings rely on the design of the prompt, which can affect the variability of output in terms of lexemes. Lexical patterns may vary depending on the prompts. Lastly, the lexical dimension is not the only dimension that might be explored during linguistic analysis; other dimensions like syntactic complexity and rhetorical structure are not all covered in this study.

Although there are some of these problems, the study is informative and relevant in the current debate on lexical variation, both between AI-generated and human-written academic texts and for the field of applied linguistics and digital humanities more generally.

Theoretical Analysis

This study is theoretically based on the corpus linguistics, applied linguistics and discourse analysis. Corpus linguistics offers a structured approach to the examination of large corpora of authentic language data, and the discovery of common lexical and grammatical characteristics. Sinclair claims that corpus-based analysis is a tool that allows for the examination of frequency, collocation and lexical structure in texts in a scientifically reliable way (15). For the comparison of lexical diversity in AI-generated and human-written academic research, corpus linguistics is crucial in this study.

Lexical variation is regarded as one of the important parameters for the evaluation of linguistic competence and text quality. Indicates the diversity and complexity of vocabulary in a text.

According to Malvern et al., lexical diversity is the amount of vocabulary that varies, which is an indicator of language proficiency and communicative effectiveness (23). A higher lexical diversity is often correlated with clarity, originality and intellectual content in academic writing.

In general, human written academic texts show a greater lexical flexibility, which is related to the cognitive processing, contextual awareness and disciplinary knowledge of the author. Vocabulary is deliberately chosen by writers to fit the needs of argumentation and audience. AI technologies, such as ChatGPT, on the other hand, generate texts based on statistical prediction models learned from vast linguistic data sets, sometimes resulting in repetitive vocabulary and expressions.

AI-Generated Language and Statistical Modeling

AI writing systems work by using machine learning algorithms that forecast the most likely sequence of words, as well as training data. According to Bender et al., in the absence of genuine understanding of language, large language models provide text by pattern recognition and statistical correlation (612). This constraint has a direct impact on the lexical diversity of AI systems which are more likely to include high frequent words in their vocabulary compared to those that are less frequent or come from a particular context.

According to Floridi and Chiriatti, it is possible to make AI systems imitate academic style, but they are shallow in their semantics and on purpose intentional meaning-making (686). This means that AI-generated academic texts might be grammatically sound but lack original vocabulary usage and context.

On a theoretical level, it implies that AI writing is syntactically based, not cognitively or experientially. Lexical variation in such texts can, then, be algorithmic in nature and not necessarily linguistic creativity.

Discourse Analysis

Another important theoretical perspective for an analysis of the lexical diversity of the academic discourse is that of discourse analysis. Language, as a system of communication, is not just a social identity, ideology and institutional power reflection, as Fairclough puts it (37). Academic writing that has been written by humans, contains the intellectual attitude, critical thinking skills, and disciplinary identity of the writer.

While AI discourse is ideologically and personally neutral, this is the opposite of what happens with AI. Reproduces dominant linguistic patterns of the training sets without showing individual authorship or contextual intention. This variation has a significant impact on the vocabulary used, with human writers making deliberate choices to use words as part of a strategy, and AI systems producing text by leveraging a probabilistic model.

This analysis is further substantiated by Halliday's systemic functional linguistic theory, which states that language has ideational, interpersonal and textual functions (Halliday 112). Lexical selections aid in creating meaning and academic authority and tone. AI systems are more likely to focus on maintaining structure than on achieving communicative variation, while human writers would likely adapt lexical items to serve communicative functions.

Lexical Diversity and Cognitive Processes

Lexical diversity is closely associated with writing cognitive and linguistic processes. Lexical variation is based on the level of language processing and the writer's access to a vast lexical repertoire (384), McCarthy and Jarvis write. Human writers rely on memory, experience, and context in their writing, yielding broader lexicon. AI-generated texts, on the other hand, are based on pre-trained datasets and random word choices. This can result in the repetition of standard classroom terminology and restricted vocabulary use of low-frequency words. Thus,

the lexical diversity of AI-generated texts can sometimes be high in terms of coherence but not in terms of originality.

Computational Linguistics Perspective

In terms of computational linguistics, the lexical diversity can be computed using statistical measures like the type-token ratio (TTR), lexical density and lexical sophistication. These measures give quantitative information on variation in text. McCarthy and Jarvis note that these indices are commonly used to assess the complexity of a language in corpus-based studies (382). But, contextual meaning and stylistic depth cannot be completely captured by computational measures. Hence, it is important to use quantitative metrics along with qualitative discourse analysis for a comprehensive understanding of lexical diversity.

Corpus-based linguistic analysis enables the researcher to go beyond intuition and look at language in an empirical manner, using the observable patterns of language (15). This strategy can be helpful in identifying plagiarized academic writing from AI-generated work.

Theoretical Insight

All theoretical views converge to suggest that there is a clear separation between human academic writing and AI academic writing. Human writing is creative, contextual and cognitively engaging, whereas AI writing is statistical and based on patterns. While AI systems may generate grammatically correct and structurally coherent texts, they have a limited range of vocabulary due to the limitations of their training data. Human writers, however, have more lexical flexibility, caused by cognitive, cultural and disciplinary factors.

The hypothesis is proposed in this theoretical framework, stating that there are significant differences between the lexical diversity of AI-generated and human-generated academic research, specifically in terms of vocabulary richness, contextual variation, and stylistic expression.

Corpus linguistics as a methodology offers the grounds for lexical analysis, discourse analysis serves to illuminate the social and ideological character of academic texts, and computational linguistics gives us some resources to quantify the lexical variation in a text. These frameworks provide robust theoretical support for the comparison of AI-generated academic text and human-written academic text. This approach enables a comprehensive, multi-faceted view of lexical diversity, including measurable linguistic characteristics and more abstract cognitive and contextual factors that impact academic writing.

5. Discussion and Analysis

AI-generated texts and human-written academic writing exhibit distinct differences in vocabulary, style, and lexical variety. Academic texts written by humans show a much greater lexical variation compared to academic texts created by AI using ChatGPT.As can be seen from the human written academic research, the use of lexical variation is consistently higher than the AI produced academic texts using ChatGPT. This indicates that human writers use more varieties of vocabulary items, which are more cognitively flexible and contextually sensitive. Language proficiency and textual richness are closely related to lexical diversity, as stressed by Malvern et al. (23). Human composed texts have more variation across texts in their use of words, especially in terms of terms used in the discipline and evaluative expressions. Unlike human-written content, which often uses unique phrasing and vocabulary, AI-generated writing features more common academic jargon and a reliance on standard phrases.

Lexical Density: Structural Differences

The lexical density analysis shows that the human-written academic texts have a greater rate of content words (nouns, verbs, adjectives, and adverbs) than the AI-generated academic texts. According to McCarthy and Jarvis, lexical density is a good measure of the informativeness of

academic texts (384). In human texts, longer noun phrases are used, more complex verb structure, and a more precise use of adjectives, which are taken to deeper conceptual expression. They usually use “connectives” and “academic language” that are typical of written texts produced by computers. This improves fluency and coherence, at the cost of less lexical specificity. In Sinclair's explanation, patterns in the corpus may show repetition in machine-generated language because the language is selected by what it occurs most often, frequency-driven selection (15). It's evident in the AI-generated text set, with some phrases and transitions being repeated throughout different texts.

Lexical Sophistication and Vocabulary Range

The findings of this study on the lexical complexity of AI-generated research articles versus human-written ones indicate that the human-written research articles tend to be significantly more complex lexically. The use of sophisticated vocabulary demonstrates deeper engagement with topics and more mastery over academic discourse. Low-frequency words, subject-specific words and words that are specific to a context are often used by human writers. AI systems can simulate the academic style, but they are not semantically grounded, which makes it much more difficult for them to learn new words that of course have different meanings (686). This restriction is reflected in the AI-generated text, as the vocabulary is often found in a narrow academic sphere.

By contrast, the human writer adapts vocabulary as the argument unfolds, as research focuses, and for rhetorical reasons. This versatility helps to develop a more extensive vocabulary and sophisticated language in academic settings.

Repetitive Lexical Patterns in AI-Generated Texts

A peculiar observation is the repetition of lexical forms in academic texts produced by AI. Words like it is important to note, this study aims to, and furthermore are used very often in several AI generated texts. This repetition was an indication of probability-based language modelling instead of context-based creativity. Bender et al. give the following explanation: Large language models produce text based on statistical prediction, not conceptual understanding (612). This makes AI-produced content more likely to repeat common phrases or patterns seen in the training data, rather than create original combinations of words and phrases. Transitional phrases and sentence structures also vary more in human-written texts, leading to greater stylistic diversity in these texts, as do evaluative expressions.

The voices of students are diverse and individual. Another important aspect of stylistic variation is that of differences. Academic texts are texts written by humans that have an academic voice that is unique because of a personal interpretation, critical engagement, and the expertise of the discipline. Fairclough stresses that identity, ideology and positioning in discourse are all manifested in language (37). AI-generated texts, on the other hand, do not always have a consistent academic tone as they are created algorithmically and not by the individual. They have a formal tone and coherence despite their seemingly neutral and uniform stylistic expression.

According to Halliday's theory of functional linguistics, in communication language also plays ideational and interpersonal functions (112). While AI systems focus on structural correctness, human writers make conscious decisions about their lexical selection in order to perform these functions.

Use of Context and Construction of Meaning

Contextual flexibility appears to be higher in human-written academic texts, where writers change vocabulary based on the complexity of the argument and discipline. This flexibility leads to more accurate and relevant word selection. Texts produced by AI, however, tend to lack contextual adaptability. Words can be grammatically correct, but semantically not well suited to

arguments. However, features of lexical variation are not merely quantitative, as McCarthy and Jarvis point out, lexical variation is also about the appropriate use in context (382). This constraint implies that AI can produce superficially fluent academic text, but it may not be contextually rooted.

Academic writing practices as an implication

The results of this research have implications for academic writing and teaching. In addition to grammar and structure, AI tools like ChatGPT can offer valuable assistance, but can also foster an overreliance on predictable ways of using words. AI's popularity in academic settings has sparked worries over originality, authorship, and academic integrity, as noted by Cotton et al. (4). As a result, teachers need to strike a balance between the advantages of using AI tools and ensuring that students keep developing their writing ability and vocabulary. The findings indicate that AI should not replace human academic writing but rather be used as an assistant.

Summary of Analysis

In summary, the analysis shows that there are more words used in academic research that are not commonly found in everyday conversations, more complex vocabulary in the AI-generated texts, and more flexibility in the contexts used. Although AI-generated writing is grammatically correct and coherent, it often repeats words and lacks variation in style. This is in line with the theory that human cognition is an important factor in lexical diversity that AI cannot yet fully replicate.

6. Conclusion

The current study is a comparative study of lexical diversity used in AI-generated and human-written academic research. The results show distinctly that human written academic texts have higher lexical diversity, higher level of vocabulary, and higher degree of flexibility in utilizing the context.The results clearly demonstrate that human written academic texts exhibit higher level of lexical diversity, lexical level and flexibility in terms of using the context than the AI-generated texts created by ChatGPT.

In all cases, human written academic text has been found to exhibit higher vocabulary variation, finer selection of words, and a more nuanced style. These attributes are highly correlated with cognitive processes, subject matter competence, and contextual awareness. While human-written texts will offer a range of different word combinations and formulas, and elaborate on concepts, AI-written texts will be repetitive and formulaic, with phrases that are predictable based on statistics.

The lexical density analysis also supports the observation that there are more words in the content in human written text, which results in a more elaborate conceptual and analytical expression. Though grammatically correct and structurally sound, AI-generated texts often fall short in lexical precision and context, or may contain jargon.

Respond to Research Questions:

This research study has yielded the following answers to the research questions:

1. Lexical Diversity Difference: Human-written content tends to have a much higher level of lexical diversity in comparison to AI-generated content.
2. Lexical Density and Sophistication: Human texts have higher lexical density, and use more sophisticated vocabulary.
3. Repetitive Patterns: AI-generated texts exhibit repetitive use of common academic expressions and sentence structures.
4. The stylistic variation and the development of a personal academic voice is stronger in human writing, whereas AI writing is uniform and neutral.

5. Academic implications: AI-aided writing supports yet has concerns regarding originality, reliance, and decreased lexical creativity.

Theoretical Implications

The findings support corpus linguistics and discourse analysis theories that emphasize the importance of lexical diversity in academic writing. Sinclair's corpus-based approach explains that language patterns can be systematically analyzed through large datasets (15). This study confirms that such patterns differ significantly between human and AI-generated texts.

Fairclough's discourse theory further explains that human writing reflects identity, ideology, and intellectual agency (37). The absence of these features in AI-generated texts highlights a key limitation in machine-generated academic discourse.

Pedagogical and Academic Implications

The results of this study have important implications for teaching and learning in higher education. While AI tools such as ChatGPT can enhance grammar, coherence, and writing efficiency, they should not replace human cognitive engagement in academic writing. Educators should encourage students to use AI responsibly as a supportive tool rather than a primary source of academic writing. Emphasis should be placed on developing lexical richness, critical thinking skills, and independent writing abilities. Cotton et al. emphasize that institutions must establish clear academic integrity policies in response to the growing use of AI in education (4). This study supports that recommendation by highlighting the need for balanced integration of AI technologies in academic environments.

Limitations of the Study

Although the study provides meaningful insights, certain limitations must be acknowledged. The research is limited to English-language academic texts and does not include multilingual data. Additionally, the corpus size is relatively small due to time constraints, which may limit generalizability.

Another limitation is the dependency of AI-generated texts on prompt design, which may influence lexical outcomes. Different prompts could produce different levels of lexical diversity. Despite these limitations, the study offers valuable contributions to understanding lexical differences between human and AI academic writing.

Recommendations for Future Research

Future research should expand corpus size and include multilingual datasets to enhance generalizability. Further studies may also investigate syntactic complexity, rhetorical structure, and discourse coherence in AI-generated writing.

Additionally, longitudinal studies are needed to examine how AI tools influence student writing development over time. Researchers may also explore disciplinary differences in AI-generated academic writing across sciences, social sciences, and humanities.

In conclusion, this study confirms that while AI-generated academic writing provides efficiency and structural accuracy, it cannot yet fully replicate the lexical diversity, contextual richness, and stylistic depth of human-written academic research. Human cognition remains central to lexical creativity and academic expression. Therefore, AI should be viewed as a complementary academic tool rather than a replacement for human intellectual and linguistic capability.

References

Bender, Emily M., et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, 2021, pp. 610–623.

Biber, Douglas. *Variation across Speech and Writing*. Cambridge UP, 1991.

Cotton, Debby R. E., et al. "Chatting and Cheating: Ensuring Academic Integrity in the Era of ChatGPT." *Innovations in Education and Teaching International*, vol. 61, no. 1, 2024, pp. 1–12.
Creswell, John W. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. 5th ed., SAGE Publications, 2018.
Etikan, Ilker, et al. "Comparison of Convenience Sampling and Purposive Sampling." *American Journal of Theoretical and Applied Statistics*, vol. 5, no. 1, 2016, pp. 1–4.
Fairclough, Norman. *Language and Power*. 3rd ed., Routledge, 2015.
Floridi, Luciano, and Massimo Chiriatti. "GPT-3: Its Nature, Scope, Limits, and Consequences." *Minds and Machines*, vol. 30, no. 4, 2020, pp. 681–694.
Halliday, M. A. K., and Christian M. I. M. Matthiessen. *Halliday's Introduction to Functional Grammar*. 4th ed., Routledge, 2014.
Malvern, David, et al. *Lexical Diversity and Language Development: Quantification and Assessment*. Palgrave Macmillan, 2004.
McCarthy, Philip M., and Scott Jarvis. "MTLD, Vocd-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment." *Behavior Research Methods*, vol. 42, no. 2, 2010, pp. 381–392.
Sinclair, John. *Corpus, Concordance, Collocation*. Oxford UP, 1991.
Stubbs, Michael. *Words and Phrases: Corpus Studies of Lexical Semantics*. Blackwell Publishing, 2001.
Swales, John M., and Christine B. Feak. *Academic Writing for Graduate Students*. 3rd ed., University of Michigan Press, 2012.
Van Dijk, Teun A. *Discourse and Context: A Sociocognitive Approach*. Cambridge UP, 2008.
Widdowson, H. G. *Discourse Analysis*. Oxford UP, 2007.