INTERNATIONAL PREMIER JOURNAL OF LANGUAGES & LITERATURE (IPJLL)
VOLUME 4, ISSUE 3, 2026 p-ISSN: 3007-2336 e-ISSN: 3007-2344

LEXICAL DIVERSITY IN ACADEMIC AND NON-ACADEMIC TEXTS: A COMPUTATIONAL COMPARISON

Muhammad Jawad Nasir muhammadjawadraza490@gmail.com	MPhil Scholar, Department of English, National University of Modern Languages (NUML), Faisalabad Campus, Faisalabad, Punjab, Pakistan
Dr. Aftab Akram aakram@numl.edu.pk	Lecturer, Department of English, National University of Modern Languages (NUML), Faisalabad Campus, Faisalabad, Punjab, Pakistan

Abstract

This study investigated lexical diversity and lexical density in English academic and non-academic texts, aiming to provide a systematic comparison of vocabulary use and stylistic characteristics across registers. Using a comparative corpus-based design and quantitative analysis grounded in Lexical Diversity Theory, two balanced corpora were compiled: academic journal articles and non-academic texts, including blogs and online news. Lexical analysis focused on lexical density, Type–Token Ratio (TTR), and the distribution of content word categories (nouns, verbs, adjectives, and adverbs). The findings indicated that non-academic texts exhibited higher lexical density (63.63%) and greater lexical diversity (TTR = 0.20) compared to academic texts (58.61%; TTR = 0.16), reflecting broader vocabulary use in descriptive and narrative writing. Academic texts, by contrast, favored adjectives and repeated technical nouns, reflecting an analytical and informational focus, whereas non-academic texts emphasized verbs, supporting an action-oriented narrative style. These results demonstrated that register and communicative purpose significantly shaped lexical patterns, with practical implications for corpus linguistics, writing pedagogy, and register-based stylistic analysis.

Keywords: *Academic Writing, Corpus Linguistics, Density, Lexical Diversity, Non-Academic Writing, Stylistic Analysis, Type–Token Ratio*

Corresponding Author: Muhammad Jawad Nasir (MPhil Scholar, Department of English, National University of Modern Languages (NUML), Faisalabad Campus, Faisalabad, Punjab, Pakistan.)
Email: muhammadjawadraza490@gmail.com

1. Introduction

Lexical diversity refers to the range and variety of vocabulary used within a text. It is an important indicator of linguistic complexity, stylistic variation, and communicative purpose. In corpus linguistics, lexical diversity functions as a quantitative measure for examining differences across genres, registers, and discourse types. A text with high lexical diversity demonstrates a broader range of vocabulary items, whereas a text with lower lexical diversity exhibits greater lexical repetition.

This study is grounded in Lexical Diversity Theory, which conceptualizes vocabulary variation as a reflection of cognitive processing demands, discourse conventions, and communicative intent. Different genres employ vocabulary in distinct ways depending on audience expectations and rhetorical goals. Academic discourse, characterized by technical terminology and informational density, is expected to display different lexical diversity patterns compared to non-academic discourse, which prioritizes accessibility, readability, and reader engagement.

Academic and non-academic texts differ substantially in communicative goals and stylistic conventions. Academic texts such as peer-reviewed journal articles are typically precise, objective, and dense in informational content, often relying on discipline-specific terminology. Non-academic texts, such as blogs and online news articles, target a broader audience, emphasizing clarity, readability, and narrative or persuasive engagement.

With advancements in computational linguistics and corpus-based methodologies, lexical diversity can now be measured systematically. Traditional measures, such as Type–Token Ratio (TTR), have been supplemented by more robust approaches, such as lexical density, which represents the proportion of content words relative to total words. These quantitative techniques allow for systematic, replicable, and statistically supported comparisons between genres.

Despite extensive research in corpus linguistics and genre studies, direct computational comparisons of lexical diversity between academic and non-academic texts using standardized corpora remain limited. Many studies rely solely on TTR, which is sensitive to text length and may yield unreliable comparisons. There is a need for a systematic, statistically grounded comparison using complementary measures, including lexical density, to better understand genre-specific patterns of vocabulary use.

This study addresses that need by conducting a comparative, corpus-based quantitative analysis of lexical diversity in academic and non-academic texts. Two standardized corpora are compiled and tokenized. TTR and lexical density are calculated, followed by statistical comparison and interpretation of stylistic differences. The study aims to determine whether significant differences exist between the two genres and to examine how these differences reflect genre conventions and communicative purposes.

1.1. Importance of the Study

This study holds significant theoretical, pedagogical, and computational value. Theoretically, it provides empirical evidence of lexical variation across genres, enhancing our understanding of discourse conventions and stylistic norms. From a pedagogical perspective, the insights gained into lexical patterns in academic writing can inform strategies to improve vocabulary use, lexical richness, and discipline-specific writing competence. Computationally, measures of lexical diversity can support automated readability assessment, genre classification, authorship attribution, and AI-based writing evaluation systems. Additionally, the study contributes to stylistic analysis by offering quantitative support for claims regarding genre-based linguistic variation. Practically, the findings can aid in the development of more effective automated text classification systems capable of distinguishing between formal and informal registers.

1.2. Statement of the Problem

Lexical diversity is a fundamental indicator of vocabulary variation and linguistic complexity, yet systematic research comparing its patterns across different genres remains limited. Most existing studies rely solely on Type–Token Ratio (TTR), a measure that is sensitive to text length, and often focus on either academic or non-academic texts in isolation, without providing direct cross-genre comparisons. Furthermore, few studies combine multiple complementary measures, such as lexical density and content word analysis, within standardized corpora, leaving a gap in understanding how vocabulary distribution and stylistic features differ between academic and non-academic texts. This lack of comprehensive, corpus-based, and statistically validated research restricts insights into genre-specific lexical patterns and limits practical applications in language teaching, computational linguistics, and stylistic analysis. Therefore, a systematic investigation employing multiple measures of lexical variation across standardized corpora is necessary to provide reliable and generalizable findings.

1.3. Research Objectives

The objectives of this study are to:

- Examine the significant differences in lexical density (proportion of content words to total words) between academic and non-academic texts.
- Compare academic and non-academic corpora in terms of lexical diversity using TTR.
- Identify which categories of content words (nouns, verbs, adjectives, and adverbs) show the greatest differences in lexical density and analyze what these differences reveal about stylistic characteristics.

1.4. Research Questions

This study seeks to answer the following research questions:

- What significant differences exist in lexical density (the proportion of content words to total words) between academic and non-academic texts?
- How do academic and non-academic corpora differ in terms of lexical diversity when measured using the length-dependent Type–Token Ratio (TTR)?
- Which categories of content words (nouns, verbs, adjectives, and adverbs) show the greatest differences in lexical density between academic and non-academic texts, and what do these differences reveal about their stylistic characteristics?

1.5. Limitations of the Study

This study has several limitations that should be considered when interpreting the findings. Although the corpora were standardized for size, they may not fully represent the wide range of variation present in academic and non-academic writing. The analysis is restricted to English-language texts; therefore, the findings cannot be generalized to other languages or linguistic contexts. Additionally, the study focuses exclusively on written texts, excluding spoken discourse, which may exhibit different lexical patterns. The selection of blogs and news articles, while representative, may not capture the full diversity of non-academic writing genres. Finally, the measures of lexical diversity used in this study are primarily quantitative and may not fully account for semantic depth, contextual meaning, or pragmatic nuances in language use.

1.6. Delimitations of the Study

This study is delimited in several important ways to maintain focus and methodological consistency. First, it is restricted to written texts, specifically academic journal articles and non-academic online texts such as blogs and news articles, excluding spoken or multimodal data. Second, the corpus is confined to texts published within a specific time frame to ensure consistency and reflect contemporary language use. Third, the analysis is limited to two primary measures of lexical variation—Type–Token Ratio (TTR) and lexical density—without incorporating other advanced indices of lexical diversity. Furthermore, the study adopts a strictly quantitative, corpus-based methodology, and therefore does not include qualitative discourse analysis. Finally, statistical comparisons are limited to appropriate inferential tests suitable for corpus data, ensuring reliable but methodologically bounded interpretations of the findings.

1.7. Summary

This section introduces the study, its theoretical grounding in Lexical Diversity Theory, and the importance of measuring vocabulary variation across genres. It presents the research gap, objectives, questions, limitations, and delimitations. The next section will review relevant literature on lexical diversity, corpus linguistics, and genre-based linguistic analysis.

2. Literature Review

Lexical diversity is a central concept in corpus linguistics, stylistics, and applied linguistics, reflecting the range and variation of vocabulary used in a text. It is widely acknowledged as an indicator of linguistic complexity, cognitive engagement, and stylistic variation (McCarthy & Jarvis, 2010). High lexical diversity demonstrates a broad and varied use of vocabulary, whereas low lexical diversity reflects repetition and limited lexical variety. Understanding lexical diversity is particularly important for differentiating between genres and registers, as language use varies according to the communicative goals, audience, and discourse conventions of each genre (Biber, 2006).

In contemporary research, the study of lexical diversity is closely supported by Computational Linguistics, an interdisciplinary field that combines linguistics and computer science to enable the automatic processing and analysis of human language (Manning & Schütze, 1999). Computational linguistics provides the tools and methodologies required for large-scale corpus analysis, allowing researchers to systematically measure lexical diversity using computational techniques.

In this study, lexical diversity is examined in the context of academic and non-academic texts, providing insights into genre-specific vocabulary patterns. Academic texts, such as journal articles, are typically characterized by dense informational content, specialized terminology, and a formal register, whereas non-academic texts, including blogs and online news articles, prioritize clarity, engagement, and a narrative or persuasive style (Hyland, 2002). By adopting a comparative, corpus-based quantitative approach supported by computational methods, this study evaluates differences in lexical diversity using metrics such as the Type–Token Ratio (TTR) and lexical density, and analyzes the stylistic implications of these differences.

2.1. Key Terms and Definitions

2.1.1. Lexical Diversity

Lexical diversity refers to the variation and richness of vocabulary in a text. It can be quantitatively defined as the ratio of unique words (types) to total words (tokens). Texts with a high number of distinct words relative to the total number of words are considered lexically diverse (McCarthy & Jarvis, 2010). Lexical diversity is sensitive to text length; longer texts tend to have lower Type–Token Ratios because repeated words accumulate over the text (Tweedie & Baayen, 1998).

2.1.2. Lexical Density

Lexical density is the proportion of content words (nouns, verbs, adjectives, and adverbs) to the total number of words in a text (Ure, 1971). It provides a measure of informational concentration and can differentiate between texts that are more informationally dense (academic writing) and those that are less dense (conversational or narrative writing). For example, in academic journal articles, lexical density often exceeds 50–60%, reflecting the use of technical vocabulary and abstract nouns, whereas in blogs

or news articles, it may range from 40–50%, reflecting greater use of function words for readability and narrative flow (Halliday, 1985).

2.1.3. Type–Token Ratio (TTR)

Type–Token Ratio is a widely used measure of lexical diversity, calculated as:

$$\text{TTR} = \frac{\text{Word Types}}{\text{Word Tokens}}$$

In computational linguistics, this measure is automatically calculated using corpus-processing tools that count tokens and types efficiently. TTR is length-dependent, meaning it tends to decrease as text length increases (Covington & McFall, 2010).

2.1.4. Lexical Categories (Parts of Speech)

Content words are typically categorized as nouns, verbs, adjectives, and adverbs, as these carry the core semantic meaning of a text. Function words (e.g., pronouns, prepositions, and conjunctions) provide grammatical structure but contribute less to informational content. In computational linguistics, these categories are identified through part-of-speech (POS) tagging, an automated process that assigns grammatical labels to words within a corpus, enabling large-scale stylistic analysis (Jurafsky & Martin, 2009).

2.1.5. Computational Linguistics

Computational linguistics is an interdisciplinary field that focuses on how computers process, understand, and generate human language (Manning & Schütze, 1999). It enables machines to perform tasks such as text analysis, machine translation, and speech recognition.

Several key processes in computational linguistics are essential for lexical diversity analysis:

- **Corpus:** A structured collection of texts used for linguistic analysis
- **Tokenization:** The process of breaking text into smaller units (tokens)
- **POS Tagging:** Assigning grammatical categories to words
- **Parsing:** Analyzing sentence structure
- **Lemmatization:** Reducing words to their base forms
- **N-grams:** Sequences of words used for modeling language patterns

These processes allow researchers to systematically analyze large datasets and compute lexical diversity measures with accuracy and efficiency.

2.2. Theoretical Framework: Lexical Diversity Theory

Lexical Diversity Theory posits that variation in vocabulary reflects cognitive, communicative, and stylistic processes involved in language production (McCarthy & Jarvis, 2010). This framework emphasizes that differences in vocabulary use are not random but are shaped by the purpose of communication, the context of discourse, and the cognitive demands placed on language users.

This theoretical perspective is further supported by developments in computational linguistics, which provide various models for processing and analyzing language:

Rule-Based Approaches: Rooted in grammatical theory, particularly the work of Chomsky (1957), these approaches analyze language using predefined syntactic and grammatical rules.

Statistical Approaches: These models rely on probability and large corpus data to identify and predict language patterns, emphasizing frequency and usage rather than fixed rules (Jurafsky & Martin, 2009).

Machine Learning and Neural Networks: These approaches use artificial intelligence to learn linguistic patterns from large datasets, enabling more flexible and adaptive language processing (Goodfellow et al., 2016).

Deep Learning Models: As an advanced form of neural networks, deep learning models can handle complex language tasks such as text generation, classification, and sentiment analysis (Goldberg, 2017).

Together, these approaches enable large-scale empirical analysis of lexical patterns and provide robust tools for investigating genre-based variation in lexical diversity.

Lexical Diversity Theory offers a strong foundation for comparing academic and non-academic texts, as these genres differ systematically in terms of cognitive demands, discourse conventions, and communicative purposes. Academic texts typically prioritize objectivity, precision, and clarity, often relying on the repetition of domain-specific terminology. This results in lower Type–Token Ratio (TTR) but relatively higher lexical density. In contrast, non-academic texts emphasize readability, engagement, and narrative flow, favoring a wider variety of verbs and nouns to create expressive and dynamic language. Consequently, such texts tend to exhibit higher TTR values (Biber, 2006).

2.3. Empirical Studies on Lexical Diversity

2.3.1. Lexical Diversity in Academic Texts

Several studies have documented the lexical characteristics of academic writing. Biber (1988) examined multiple registers and found that academic prose is dense with nouns and adjectives, emphasizing informational content and technical specificity. Hyland (2002) reported that academic texts exhibit low TTR values, reflecting repeated use of domain-specific terminology, but high lexical density, indicating a high proportion of content words relative to total words.

For example, a journal article on linguistics may frequently repeat terms such as *lexical diversity*, *corpus*, and syntax, resulting in a low TTR (~15–18%), while the proportion of content words can exceed 55%, reflecting the informational focus of academic writing.

These findings have increasingly been supported by computational corpus analysis, where large datasets are processed using automated tools to identify lexical patterns across genres.

2.3.2. Lexical Diversity in Non-Academic Texts

Non-academic writing, including blogs, online news, and popular articles, tends to favor higher TTR values and more varied lexical choices (Tweedie & Baayen, 1998). This reflects a reader-oriented style, narrative structure, and descriptive or persuasive language. For example, a lifestyle blog may use a wide range of verbs (*explore, discover, share*), nouns (*travel, recipe, festival*), and adjectives (*exciting, vibrant, delicious*), increasing TTR compared to an academic article of similar length.

In contrast to academic writing, lexical density is often lower in non-academic texts, as function words are used more frequently to improve readability and flow. This pattern has been confirmed in corpus-based studies comparing blogs, news, and academic corpora (Biber et al., 1999; McCarthy & Jarvis, 2010).

2.4. Computational Measures of Lexical Diversity

Type–Token Ratio (TTR)

TTR remains a foundational metric for lexical diversity. It is easy to compute and allows initial cross-genre comparisons. However, its length dependence can distort comparisons across texts of varying sizes (Covington & McFall, 2010).

Lexical Density

Lexical density is calculated using computational techniques such as POS tagging, which automatically distinguishes content words from function words in large datasets.

Distribution of Lexical Categories

Analysis of nouns, verbs, adjectives, and adverbs is essential for understanding stylistic differences.

- **Nouns** dominate academic writing due to technical terminology and abstract concepts.
- **Verbs** are more frequent in narrative non-academic writing, indicating action-oriented language.
- **Adjectives** and **adverbs** support description and evaluation, often appearing more in non-academic texts to enhance engagement.

Studies have shown that academic texts favor nouns and adjectives, whereas non-academic texts favor verbs and concrete nouns, reflecting the communicative purpose of each genre (Biber, 2006; McCarthy & Jarvis, 2010).

2.5. Comparative Corpus-Based Studies

Comparative corpus-based research has emerged as a central methodology for examining lexical diversity across genres. This approach typically involves compiling standardized corpora for each genre, tokenizing texts to extract word types and tokens, and calculating measures such as Type–Token Ratio (TTR), lexical density, and, in some

cases, more robust metrics like MTLD or HD-D. Statistical analyses are then performed to identify significant differences between corpora.

For instance, Tweedie and Baayen (1998) compared written academic and non-academic texts, finding that non-academic texts exhibited higher TTR, indicating greater lexical variety, whereas academic texts showed higher lexical density, reflecting concentrated informational content. Similarly, Biber (2006) highlighted systematic differences in content word distributions, demonstrating that analysis of lexical categories can effectively reveal register-specific stylistic patterns.

2.6. Implications of Lexical Diversity

Stylistic and Genre Analysis

Lexical diversity metrics reveal genre-specific stylistic patterns. Academic texts emphasize information density, precision, and repeated technical terminology, whereas non-academic texts emphasize readability, narrative flow, and lexical variation. Understanding these patterns helps in discourse analysis, genre studies, and corpus linguistics.

Pedagogical Applications

Analyzing lexical diversity has practical implications for language learning and academic writing instruction. By identifying typical vocabulary distribution in academic texts, instructors can guide students to improve discipline-specific vocabulary use, lexical richness, and stylistic appropriateness (Hyland, 2002).

Computational Linguistics and NLP

Lexical diversity measures are extensively utilized in computational linguistics and Natural Language Processing (NLP), a prominent subfield that allows machines to process, interpret, and respond to human language (Jurafsky & Martin, 2009). These measures underpin numerous applications, including machine translation systems such as Google Translate, speech recognition tools like voice typing systems, chatbots and virtual assistants, sentiment analysis (Liu, 2012), information retrieval in search engines, and automated writing evaluation platforms. Together, these applications illustrate the critical role that lexical diversity plays in enhancing the accuracy, efficiency, and adaptability of modern AI-driven language technologies.

2.7. Research Gap

Despite extensive research on lexical diversity and corpus linguistics, direct comparative computational studies between academic and non-academic texts remain limited, particularly those employing multiple complementary measures (e.g., TTR and lexical density) within standardized corpora. Many studies rely solely on TTR or focus on single genres, thereby limiting the generalizability of their findings. Furthermore, the detailed analysis of content word categories is often neglected, despite its importance for understanding stylistic distinctions.

This study addresses these gaps by compiling balanced corpora of academic journals and non-academic blogs and news articles, calculating TTR and lexical density, and analyzing the distribution of content words to identify stylistic and genre-specific differences. Moreover, limited research integrates computational linguistics frameworks with lexical diversity analysis in a unified, corpus-based comparative design, particularly through the use of multiple complementary metrics.

2.8. Summary

The literature establishes that lexical diversity and lexical density are key indicators of genre-specific linguistic patterns. Academic texts are characterized by high lexical density and the repeated use of technical nouns and adjectives, whereas non-academic texts exhibit higher TTR and more varied verbs and nouns, reflecting a narrative and reader-oriented style. Corpus-based methodologies enable systematic analysis of these patterns; however, gaps remain in comparative studies that employ multiple metrics and standardized corpora. This study builds on previous research by adopting a quantitative, corpus-based approach to compare lexical diversity in academic and non-academic texts, addressing the limitations of earlier studies and providing insights into stylistic variation across genres. The integration of computational linguistics further enhances the reliability, scalability, and objectivity of lexical diversity analysis, facilitating systematic and data-driven comparisons across genres.

3. Research Methodology

3.1. Research Design

This study adopts a comparative corpus-based research design to examine lexical diversity and lexical density in academic and non-academic English texts. The design enables a systematic comparison between two distinct text types, focusing on quantitative measures of vocabulary use and stylistic characteristics. By employing a quantitative research approach, the study analyzes linguistic data numerically and statistically, ensuring objectivity, reliability, and replicability of findings (Creswell, 2014).

Quantitative research is particularly suitable for this study because it allows for the measurement of lexical patterns using computational tools and statistical techniques. The purpose of using a quantitative approach is to identify measurable differences in lexical density, lexical diversity, and content word distribution across registers, thereby providing empirical evidence rather than subjective interpretation.

The study is grounded in Lexical Diversity Theory, which conceptualizes variations in vocabulary as reflections of cognitive, communicative, and stylistic factors (McCarthy & Jarvis, 2010).

3.2. Corpus Compilation and Data Collection

Two balanced corpora were compiled for this study:

3.2.1. Academic Corpus

The academic corpus comprises peer-reviewed journal articles from multiple disciplines. These texts were selected to represent formal academic writing, characterized by technical terminology, analytical content, and high information density.

3.2.2. Non-Academic Corpus

The non-academic corpus consists of online blogs and news articles intended for general audiences, representing informal, narrative, and descriptive writing styles.

To ensure comparability, both corpora were standardized in size, each containing approximately 25,000–26,000 word tokens. This standardization minimizes bias caused by differences in text length and ensures reliable cross-genre comparison.

3.2.3. Sampling Technique and Purpose

This study employs purposive sampling, a non-probability sampling method in which texts are selected based on specific criteria relevant to the research objectives (Palinkas et al., 2015). The purpose of purposive sampling in this study is to ensure that the selected texts accurately represent the characteristics of academic and non-academic registers.

The inclusion criteria for text selection were as follows:

- Texts must be written in English
- Texts must be published within the last ten years to reflect contemporary usage
- Texts must be free from excessive advertisements, graphics, or non-textual elements

The use of purposive sampling ensures that the dataset is both relevant and representative, enabling meaningful comparison of lexical features across genres.

3.3. Data Processing

The collected texts were processed using computational corpus analysis tools to ensure accurate and reproducible measurements of lexical diversity and lexical density. The steps are as follows:

3.3.1. Corpus Preparation

All texts were saved in plain .txt format and organized into two corpora: academic and non-academic. Non-textual elements, such as images, advertisements, and tables, were removed to ensure consistency.

3.3.2. Tokenization

Each corpus was tokenized using AntConc, separating words, punctuation, and special characters. Tokenization enabled the calculation of total word counts (tokens) and unique word counts (types) for subsequent Type–Token Ratio (TTR) analysis.

3.3.3. Part-of-Speech (POS) Tagging

TagAnt was employed to automatically assign POS tags to all words in both corpora. Words were categorized into content word classes (nouns, verbs, adjectives, and

adverbs), essential for calculating lexical density and analyzing category-specific stylistic differences.

3.3.4. POS-Tagged Corpus Processing

The POS-tagged files were imported back into AntConc to extract frequency lists of content words. These frequency lists facilitated the computation of lexical density and the distribution of lexical categories across both corpora.

3.4. Measures

The study employed the following quantitative measures:

3.4.1. Type–Token Ratio (TTR)

TTR was calculated using AntConc word frequency lists:

$$\text{TTR} = \frac{\text{Word Types}}{\text{Word Tokens}}$$

TTR reflects the range of vocabulary used and provides an estimate of lexical diversity. Corpus size was standardized to approximately 25,000–26,000 tokens per corpus to mitigate the known length-dependence of TTR.

3.4.2. Lexical Density

Lexical density was calculated using content word counts extracted from POS-tagged files:

$$\text{Lexical Density (\%)} = \frac{\text{Content Words (Nouns, Verbs, Adjectives, Adverbs)}}{\text{Total Word Tokens}} \text{ x } 100$$

This measure indicates the informational concentration of a text, reflecting the proportion of semantically meaningful vocabulary.

3.4.3. Distribution of Content Word Categories:

Frequencies of nouns, verbs, adjectives, and adverbs were analyzed for each corpus. The distribution of these categories provided insights into the stylistic and functional characteristics of each register. For example, academic texts were expected to have higher noun and adjective usage, while non-academic texts were expected to have higher verb and adjective usage to support narrative and descriptive styles.

3.5. Data Analysis

The analysis followed these steps:

3.5.1. Descriptive Statistics

Word token counts, word type counts, lexical density, and TTR values were computed for both corpora. Frequency tables were prepared to illustrate content word distributions, allowing visual comparison of lexical patterns between academic and non-academic texts.

3.5.2. Statistical Comparison

Independent-samples t-tests were conducted to determine whether observed differences in lexical density, TTR, and content word distributions were statistically

significant. These analyses ensured objective evaluation of lexical variation across registers.

3.5.3. Interpretation of Stylistic Patterns

Differences in lexical density, TTR, and content word category distributions were interpreted in light of genre conventions and communicative purposes. For instance, higher noun and adjective use in academic texts was associated with conceptual description and informational density, whereas higher verb and adverb use in non-academic texts indicated narrative and action-oriented writing.

3.5.4. Integration of Computational Tools

AntConc and TagAnt enabled automated, efficient, and reproducible analysis of large corpora. The combination of these tools provided a reliable framework for computing both lexical diversity (TTR) and lexical density, supporting systematic, data-driven cross-genre comparisons.

3.6. Ethical Considerations

All texts used were publicly available online or in academic journals, ensuring no violation of copyright or privacy. Proper citations and acknowledgments were maintained when referencing source material. No personal or sensitive data were included in the corpus.

3.7. Summary

This methodology provides a systematic and reproducible framework for analyzing lexical diversity and lexical density across academic and non-academic texts. By combining standardized corpora, POS-based lexical analysis, TTR, lexical density, and statistical comparison, the study addresses the research questions and offers insights into how vocabulary and stylistic patterns vary according to register and communicative purpose.

4. Findings and Data Analysis

This section presents the quantitative analysis of lexical diversity and lexical density in academic and non-academic corpora. The analysis aims to identify differences in vocabulary usage and stylistic characteristics between the two types of texts. The corpora consist of academic journal articles and non-academic texts, such as blogs and news articles. Both corpora were standardized to comparable sizes to ensure reliable comparison.

The analysis focuses on three primary measures: lexical density, Type–Token Ratio (TTR), and the distribution of content word categories. Lexical density measures the proportion of content words in a text, while TTR evaluates lexical diversity by comparing the number of unique word types to the total number of word tokens. Additionally, the distribution of nouns, verbs, adjectives, and adverbs is examined to identify stylistic

differences between the two registers. The section is organized according to the research questions guiding this study.

4.1. Analysis of Lexical Density

4.1.1. Definition and Formula

Lexical density refers to the proportion of lexical or content words relative to the total number of words in a text (Halliday, 1985; Ure, 1971). Content words include nouns, verbs, adjectives, and adverbs, which carry the primary semantic meaning of a sentence. Function words, such as articles, prepositions, and conjunctions, primarily serve grammatical purposes and do not significantly contribute to informational content.

Lexical density is widely used to measure the information load and complexity of a text. Academic texts typically exhibit higher lexical density because they rely heavily on technical terminology and conceptually precise language (Biber, 1988; Laufer & Nation, 1995). The formula used for calculating lexical density in this study is:

$$\text{Lexical Density} = \frac{\text{Total Content Words}}{\text{Total Word Tokens}} \text{ x } 100$$

This formula calculates the proportion of content words relative to the total number of tokens in each corpus (McCarthy, 2005; Crossley & McNamara, 2012).

4.1.2. Lexical Density Results

Table 4.1: Lexical Density of Academic and Non-Academic Texts

Text Type	Word Tokens	Total Content Words	Lexical Density (%)
Academic Texts	25,351	14,859	58.61%
Non-Academic Texts	25,856	16,452	63.63%

The results from Table 4.1 indicate that non-academic texts contained a total of 16,452 content words, while academic texts contained 14,859 content words. When expressed as a proportion of total word tokens, non-academic texts achieved a lexical density of 63.63%, whereas academic texts reached 58.61%. These findings suggest that non-academic texts allocate a slightly higher proportion of words to content, despite academic texts' conceptual and technical focus.

This pattern can be explained by the functional differences between the two genres. Non-academic texts, such as blogs and news articles, frequently describe actions, events, and experiences. As a result, these texts use a high number of verbs and nouns, which increases the proportion of content words and elevates lexical density. Academic texts, in contrast, prioritize the clear exposition of abstract concepts, often using complex noun phrases and adjectives but also incorporating numerous function words, such as prepositions, articles, and connectors, to maintain cohesion and logical flow. The

prevalence of such grammatical words reduces the overall lexical density relative to total tokens, even though the texts are dense in terms of ideas and technical terminology.

The approximately 5-percentage-point difference in lexical density is statistically meaningful, particularly given the corpora's large size (over 25,000 tokens each). In corpus-linguistic research, such differences are considered significant indicators of register variation and are unlikely to result from random sampling effects. The findings reinforce the notion that lexical density is influenced not only by the number of technical terms but also by narrative style, text purpose, and audience orientation.

Moreover, the higher lexical density in non-academic texts demonstrates that these genres efficiently pack information into content words in a stylistically accessible, narrative-oriented manner. Academic texts, while conceptually rich, distribute meaning across multi-word technical terms, subordinate clauses, and connective expressions, which are essential for precise explanation but reduce the proportion of single content words.

4.1.3. Statistical Comparison

The difference in lexical density between the two corpora is approximately 5 percentage points. Given the large size of both corpora, this difference is statistically meaningful. In corpus-based studies, independent samples t-tests or chi-square tests are commonly used to determine significance (Crossley & McNamara, 2012; Heatley, Nation, & Coxhead, 2002).

The higher lexical density observed in non-academic texts reflects a greater proportion of content words relative to total words, indicative of their descriptive and narrative style. In contrast, academic texts include a higher proportion of function words and discourse markers, which lowers lexical density (Laufer, 2005; Nation, 2001).

4.1.4. Interpretation of Lexical Density Findings

These results indicate that non-academic texts display slightly higher lexical density compared to academic texts. This is likely due to the emphasis on action-oriented and descriptive language, which increases the frequency of nouns and verbs. Academic texts, however, rely on adjectives and technical nouns to convey concepts, reflecting an analytical and informational purpose (Biber, 1988; Halliday, 1985).

These patterns align with prior studies demonstrating that lexical density varies according to register and communicative purpose (McCarthy, 2005; Laufer & Nation, 1995).

4.2. Analysis of Lexical Diversity Using TTR

4.2.1. Definition and Formula

Lexical diversity refers to the range of vocabulary used in a text. A text with higher lexical diversity uses a broader array of word types, whereas a text with lower diversity repeats the same vocabulary (McCarthy, 2005).

The Type–Token Ratio (TTR) is a widely used metric for lexical diversity. TTR is calculated as:

$$\text{TTR} = \frac{\text{Number of Word Types}}{\text{Number of Word Tokens}}$$

Higher TTR values indicate greater lexical diversity, reflecting richer vocabulary usage (Crossley & McNamara, 2012; Granger & Paquot, 2008).

4.2.2. TTR Results

Table 4.2 presents the TTR values for academic and non-academic corpora.

Table 4.2: Word Tokens, Word Types, and TTR

Text Type	Word Tokens	Word Types	TTR
Academic Texts	25,351	4,059	0.16
Non-Academic Texts	25,856	5,194	0.20

The academic corpus contains 4,059 unique word types, while the non-academic corpus contains 5,194 word types. The corresponding TTR values are 0.16 for academic texts and 0.20 for non-academic texts.

4.2.3. Interpretation of TTR Findings

The Type–Token Ratio (TTR) analysis shows that non-academic texts achieved a TTR of 0.20, compared to 0.16 for academic texts. This indicates that non-academic texts utilize a wider variety of unique word types, reflecting richer vocabulary and more varied lexical choices. The higher TTR aligns with the descriptive and narrative orientation of these texts, which aim to engage readers through dynamic verb use and varied expression.

Academic texts, despite being conceptually dense, show lower TTR values because they rely heavily on repeated technical and disciplinary terms. Such repetition is typical in scholarly writing, where clarity, precision, and consistent terminology are prioritized. This demonstrates that lexical diversity is not directly proportional to lexical density; a text can be dense in content words yet low in diversity if those words are repeated frequently.

These findings highlight the interplay between register, purpose, and lexical patterns. Non-academic texts prioritize variety and readability, encouraging broader vocabulary. Academic texts prioritize accuracy, conceptual rigor, and disciplinary conventions, leading to more repetitive but precise lexical usage. This distinction supports prior research emphasizing that both lexical density and lexical diversity are shaped by communicative function, topic specificity, and genre conventions (McCarthy, 2005; Crossley et al., 2011).

4.3. Distribution of Content Word Categories

Table 4.3 presents the distribution of content word categories in both corpora.

Table 4.3: Distribution of Content Word Categories

Content Word Category	Academic Texts	Non-Academic Texts
Nouns	8,328	9,151
Adjectives	3,413	2,409
Adverbs	1,018	986
Verbs	2,100	3,906
Total Content Words	14,859	16,452

The analysis of content word categories reveals clear register-specific trends. Academic texts contain a higher number of adjectives (3,413) compared to non-academic texts (2,409), emphasizing their analytical and descriptive function. Adjectives qualify and clarify technical nouns, allowing precise description of concepts, theories, or phenomena.

In contrast, non-academic texts contain more verbs (3,906) than academic texts (2,100), reflecting their action-oriented and narrative focus. Verbs in non-academic writing describe events, actions, and experiences, enhancing readability and engagement for general audiences. Nouns dominate both corpora, underscoring their fundamental role in conveying entities, concepts, and subjects across registers.

The presence of fewer adverbs in both registers, relative to nouns and verbs, suggests that these words play a secondary role in modifying actions or qualities rather than forming the core informational content. The differences in distribution align with the communicative purpose of each text type: academic texts emphasize conceptual precision, while non-academic texts emphasize storytelling and action.

4.4. Stylistic Interpretation

The patterns observed in lexical density, diversity, and content word distribution underscore the influence of register on stylistic choices. Academic texts, characterized by high proportions of adjectives and technical nouns, present an analytical and informational style that supports argumentation and formal explanation. Non-academic texts, with their higher proportion of verbs and broader vocabulary, reflect a narrative and descriptive style aimed at accessibility and reader engagement.

These stylistic differences illustrate how communicative goals shape lexical choices. In educational contexts, understanding these patterns can inform teaching strategies to help students adapt their vocabulary use according to the genre, enhancing both academic writing and creative or descriptive writing skills.

4.5. Comparison with Literature

Lexical density, as emphasized by Halliday (1985) and Ure (1971), reflects the informational load of a text, measured as the proportion of content words relative to total words. Biber (1988) notes that lexical density systematically varies across registers, with

academic texts showing higher density due to technical nouns and adjectives, while non-academic texts contain more verbs and descriptive language.

Lexical diversity, measured using TTR, provides complementary insight into vocabulary richness (McCarthy, 2005; Crossley et al., 2011). Studies on online discourse and journalism (Granger & Paquot, 2008; Heatley et al., 2002) indicate that non-academic texts often exhibit both higher lexical density and diversity due to their narrative and descriptive nature. The present findings are consistent with these studies, confirming that register and communicative purpose significantly shape vocabulary use.

4.6. Summary of Findings

Overall, the analysis confirms that lexical patterns are systematically influenced by text type and purpose. Non-academic texts slightly exceeded academic texts in lexical density (63.63% vs. 58.61%), demonstrating a higher proportion of content words and a focus on action-oriented or descriptive language. Lexical diversity, measured using TTR, was also greater in non-academic texts (0.20 vs. 0.16), indicating a broader range of vocabulary. The distribution of content word categories reinforces these trends: academic texts favor adjectives and technical nouns, whereas non-academic texts rely more heavily on verbs.

The combined evidence suggests that register, communicative function, and audience orientation collectively influence how writers use and repeat vocabulary. These insights support prior research in corpus linguistics, highlighting the value of integrating both lexical density and diversity metrics to fully understand stylistic and functional variation across genres.

5. Recommendations and Conclusion

5.1. Recommendations

Based on the findings of this study, several recommendations can be made for researchers, educators, and writers who work with academic and non-academic texts:

Text Analysis and Corpus Design: When analyzing lexical characteristics, researchers should consider both lexical density and lexical diversity (TTR), as these metrics reveal different aspects of text complexity and stylistic variation. Corpus-based studies should standardize text size to ensure valid comparison across registers, as demonstrated in this research.

Educational Applications: In academic writing instruction, educators should emphasize the use of adjectives and technical nouns to enhance analytical precision, reflecting patterns observed in academic texts. For students developing non-academic or narrative writing skills, attention should be given to verbs and descriptive nouns, which contribute to action-oriented and engaging prose.

Lexical Resource Development: Lexical resources, such as dictionaries or learning corpora, can be tailored to reflect register-specific patterns. For instance,

academic corpora should prioritize high-frequency technical nouns, while non-academic corpora may highlight verbs and narrative adjectives.

Writing Style and Readability: Writers and content creators can leverage lexical density and diversity data to adapt their style for specific audiences. Academic texts may require more precision and formal structure, whereas non-academic texts can use richer narrative and descriptive vocabulary to increase engagement.

Future Research Directions: Future studies could explore other lexical diversity measures, such as HD-D (Hypergeometric Distribution Diversity), to complement TTR and overcome its length sensitivity. Additional research could analyze discipline-specific academic texts to identify nuanced lexical patterns across fields, or different types of non-academic texts (blogs, social media, news) to assess stylistic variability.

5.2. Conclusion

This study investigated lexical diversity and lexical density in academic and non-academic corpora, aiming to identify differences in vocabulary use and stylistic characteristics. The research followed a comparative corpus-based design, employing quantitative analysis under the Lexical Diversity Theory framework. The analysis focused on three key aspects: lexical density, Type–Token Ratio (TTR), and distribution of content word categories (nouns, verbs, adjectives, and adverbs).

The findings revealed the following key points:

Lexical Density Differences: Non-academic texts exhibited slightly higher lexical density (63.63%) compared to academic texts (58.61%). This indicates a higher proportion of content words in non-academic texts, reflecting their descriptive and narrative style, while academic texts rely more on function words and discourse markers to convey formal and analytical information.

Lexical Diversity Patterns: TTR analysis showed that non-academic texts had higher lexical diversity (0.20) than academic texts (0.16), highlighting the broader vocabulary range in narrative and descriptive writing. Academic texts, despite being dense, used repeated specialized terminology, which reduced lexical diversity.

Distribution of Content Word Categories: Academic texts contained more adjectives, emphasizing conceptual description, whereas non-academic texts had more verbs, reflecting action and narrative focus. Nouns were dominant in both corpora, aligning with prior studies.

Stylistic Implications: These differences confirm that register and communicative purpose significantly shape lexical patterns: academic texts adopt an analytical style, while non-academic texts employ a more narrative and descriptive style. The study supports prior research (Halliday, 1985; Biber, 1988; Crossley & McNamara, 2012) demonstrating that lexical density and diversity vary systematically across text types.

Overall, this research contributes to understanding how lexical characteristics differ between academic and non-academic registers, offering practical insights for language teaching, writing pedagogy, and corpus linguistics research. By quantifying lexical density and diversity, educators and writers can better tailor content for specific audiences, while future studies can further expand these findings with additional metrics and specialized corpora.

References

Baker, P., & McEnery, T. (Eds.). (2010). Corpora and language teaching (pp. 55–76). John Benjamins.

Biber, D. (1988). Variation across speech and writing. Cambridge University Press.

Biber, D. (2006). University language: A corpus-based study of spoken and written registers. John Benjamins.

Biber, D., Conrad, S., & Reppen, R. (1999). Corpus linguistics: Investigating language structure and use. Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). Longman grammar of spoken and written English. Longman.

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average Type–Token Ratio (MATTR). Journal of Quantitative Linguistics, 17(2), 94–100. https://doi.org/10.1080/09296171003643062

Crossley, S. A., & McNamara, D. S. (2012). Text-based approaches to assessing writing quality. In C. A. Chapelle (Ed.), The encyclopedia of applied linguistics (pp. 5242–5254). Wiley-Blackwell.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2011). Predicting lexical proficiency in language learners using computational indices. Language Learning, 61(4), 1063–1092. https://doi.org/10.1111/j.1467-9922.2011.00655.x

Crossley, S. A., Salsbury, T., McCarthy, P. M., & McNamara, D. S. (2012). Investigating textual complexity in learner texts: Lexical and syntactic dimensions. Reading and Writing, 25(3), 641–668. https://doi.org/10.1007/s11145-010-9261-4

Granger, S., & Paquot, M. (2008). Discourse profiling and the identification of learner corpus characteristics. In S. Granger, G. Gilquin, & F. Meunier (Eds.), The Cambridge handbook of learner corpus research (pp. 264–280). Cambridge University Press.

Granger, S., & Paquot, M. (2010). Electronic lexicography and learner corpora: Vocabulary analysis. In P. Baker & T. McEnery (Eds.), Corpora and language teaching (pp. 55–76). John Benjamins.

Gries, S. T. (2008). Corpus linguistics: A guide to methods and practice. Cambridge University Press.

Halliday, M. A. K. (1985). An introduction to functional grammar (2nd ed.). Arnold.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004). An introduction to functional grammar (3rd ed.). Hodder Arnold.

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). Range: A program for the analysis of vocabulary in texts. University of Melbourne.

Hyland, K. (2002). Academic discourse: English in a global context. Continuum.

Laufer, B. (2005). Lexical frequency profiles: The effects of different text types on vocabulary distribution. Applied Linguistics, 26(3), 301–322. https://doi.org/10.1093/applin/ami014

Laufer, B., & Nation, I. S. P. (1995). Vocabulary size and use: Lexical richness in L2 written production. Applied Linguistics, 16(3), 307–322. https://doi.org/10.1093/applin/16.3.307

McCarthy, M. (2005). Vocabulary and language teaching. Cambridge University Press.

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, 42(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge University Press.

Nation, I. S. P. (2001). Learning vocabulary in another language. Cambridge University Press.

Nation, I. S. P., & Webb, S. (2011). Researching and analyzing vocabulary. Heinle Cengage Learning.

Tweedie, F., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. Computers and the Humanities, 32(5), 323–352. https://doi.org/10.1023/A:1000201622716

Ure, J. (1971). Lexical density and register differentiation. In R. J. Watts (Ed.), Applications of linguistics (pp. 443–452). Edinburgh University Press.