



ADVANCE SOCIAL SCIENCE ARCHIVE JOURNAL

Available Online: <https://assajournal.com>

Vol. 05 No. 01. Jan-March 2026. Page#. 6304-6318

Print ISSN: [3006-2497](#) Online ISSN: [3006-2500](#)Platform & Workflow by: [Open Journal Systems](#)<https://doi.org/10.5281/zenodo.22110912>**A Corpus-Based Analysis of Lexical Bundles in AI-Generated and Human-Written Academic Essays****Zeena Naz**

MS Scholar, Department of English, Kohat University of Science and Technology, KUST, Pakistan

Email. Khattaktanveer360@gamil.com**Sharqa Shafique**

MS Scholar, Department of English, Kohat University of Science and Technology, KUST, Pakistan

Email. sharqaskhan2000@gmail.com**Muhammad Ihtesham (Corresponding Author)**

MS Scholar, Department of English, Kohat University of Science and Technology, KUST, Pakistan

Email. muhammadihtesham368@gmail.com**Abstract**

The adoption of Generative AI in academic writing has blurred the distinction between AI-generated and human-generated text, leaving the question of how to distinguish the two. With Generative AI now integrated into academic writing, the question is now how to tell the difference between content generated by AI and that written by humans. Although there have been various studies of recurrent multi-word sequences so far, they have not been studied as extensively as they are in the field of academic discourse, especially in terms of how they organise, interpret and rhetorically shape academic discourse. In this study, the focus has been on identifying the lexical bundles in the computer-generated and in the human-written academic writing and comparing the two by a corpus-based approach. It is based on the theory of lexical bundles which is based on frequency, proposed by Biber and his colleagues, and the structural and functional taxonomy and classification by Biber and others. The analysis focuses on three aspects: frequency and distribution, structural configuration and discourse function. This research explores how much each style of writing relies on the same bundles of noun phrases, prepositional phrases, verb phrases, and clauses and whether there are any preferences for those bundles with each style of writing. The approach is focussed on function, such as sets of functions related to research-oriented/reference functions, text organisation and stance. The study adds to the existing studies of AI-assisted academic writing, which have focused on analysing individual words and grammar at the sentence level, by focusing on the patterns of recurrence of phrases. The results will inform about the formulaic quality of AI created academic writing and its association with traditional academic writing by humans. The study also offers implications for academic writing, corpus-informed pedagogy, as well as critical use of Generative AI in higher education.

This paper delves into the implications of lexical bundles in academic writing and their association with AI tools such as ChatGPT. We will explore the importance of lexical bundles in academic writing and the correlation between lexical bundles and artificial intelligence (AI) tools like ChatGPT in this paper.

1. Introduction

Besides its specialized vocabulary and complex grammatical structures, academic writing also involves frequent uses of word sequences that allow the writer to develop arguments, to arrange information, to indicate relationships, and to convey attitudes towards propositions. Lexical bundles are the occurrence of these combinations so many times. The meaning of lexical bundles is not necessarily idiomatic but is generally indicated by their occurrence and repetition in specific registers. In natural discourse, Biber et al. (1999) suggested that lexical bundles are frequent and regular combinations of words, and subsequent research showed that they are important in the academic and professional settings. On the other hand, in academic writing, as a result of, the extent to which, and in the context of refer to some sort of discourse organizing and referential function.

Based on the research work on corpus, it has been concluded that lexical bundles are an important part of academic discourse. Examples include lexical bundles as micro-units of discourse (Biber, Conrad and Cortes 2004), lexical bundles as stance, as structuring the discourse and as referring to meaning. With regard to register, Cortes (2004) is also able to find that the frequency of lexical bundles is related to it, since they are not equally used in each academic discipline and/or in the level of linguistic proficiency of the writers.

With the rise in the use of generative AI, academic writing is becoming ever more complex. For instance, a tool like LLM ChatGPT can write an academic essay in seconds with sometimes academic jargon, sentence structures, academic transitions and formulaic expressions. This has led to extensive research into the key differences between the writing of AI and the writing of academics. The issue becomes even more pertinent in the context of higher education as students encounter AI more than merely as a writing tool, but as a potential academic author itself.

There can be specific phraseological traits of AI writing models, as per the research done in 2025 and 2026. Azalia et al. (2025), for instance, studied the lexical bundles and their role in the argumentative texts generated by ChatGPT and found that there was a high frequency of noun-phrase bundles. Recently, Tran (2026) compared L1 student writing, L2 student writing, and the writing by ChatGPT in the form of four-word lexical bundles. The study revealed that the writing produced by the AI has formulaic density higher than that of L1 and L2 students' writing. Likewise, a second research report that is presented in Journal of Asia TEFL 2026 also has conducted a comparison between the Indonesian EFL learner essays and AI-generated argumentative writing, which has highlighted the increased importance of the analysis of lexical bundle in the writing process of EFL learners with AI support in academic writing. But this is a new area of study and further research is needed into the linguistic connection between AI-generated and human-authored academic writing. AI-generated writing doesn't just consist of random lexical and grammatical elements. It is generated probabilistically based on large amounts of text data and can thus re-generate frequent phraseological patterns rather consistently. Other writers, though, such as human writers, rely on their language knowledge, their disciplinary background, their education,

and their rhetorical choices. To this end, lexical bundles offer a particularly valuable look into the degree and/or type of formulaicity in AI-generated academic text versus that of a human writer.

This is the rationale for the present study, that will perform a comparison between the lexical bundles of AI generated and human written academic essays, in the corpus. This study does not cover the general evaluation of the sound of the AI writing but instead focuses on measurable recurrent sequence. It investigates the quantity and distribution of four-word lexical bundles, structural patterns of lexical bundles and their functions in the discourse. Lexical bundles provide concrete data about the nature of academic discourse and possible ways to model academic discourse with AI-generated texts.

The study attempts to answer the following research questions:

1. What are the first 4-word lexical bundles that come up in the learners' academic compositions written by AIs and human writers?
2. How do the lexical bundles difference between AI and human-mades
3. What are the functional types of the lexical bundles in both types of academic writing?

How much does the use of lexical bundles differentiate AI-generated academic essays from human written academic essays?

The study will answer these questions, which can be used to advance research in the field of corpus linguistics, academic writing and generative artificial intelligence. It offers a linguistic description that can also be based on the corpus, as well as stylistic impressions and AI-detection claims, for the linguistic features of academic discourse generated by AI.

2. Review of Related Literature

2.1 Lexical Bundles and Academic Discourse

Recurrent sequences of words have been a subject of interest for Corpus linguistics for a long time. The idea of lexical bundles was, however, given a special boost via the research of Biber and co-workers. Lexical bundles are recurrent sequences that are identified by frequency based corpus analysis, according to Biber et al. (1999). The hallmark of this, however, is not necessarily grammatically full nor idiomatic in terms of semantics but repeated co-occurrence in a specific register.

This is important because it helps distinguish it from the more common definition of 'idiom' and 'conventional collocation'. The sequence as a result may be a non-standard one, in the traditional sense, but is a major part of academic prose in terms of frequency and function. Lexical bundles frequently function as prefabricated and partially grammaticalized structures which ease fluent discourse production.

Biber, Conrad and Cortes (2004) pinpointed three general types of lexical bundles: stance expressions, discourse organizers, and referential expressions. Stance bundles represent attitudes, judgments or degrees of certainty; discourse organizers represent the relationships between propositions and help to structure a text; referential expression identify entities, quantities, attributes, and other aspects of experience.

A great deal of structural classifications have been employed as well. On the basis of the grammatical structures, Biber et al. 1999 divided the bundles into noun phrase based expressions, prepositional-phrase and verb phrase based expressions, and other clausal expressions. These typologies help researchers determine which bundles are in use and the grammatical characteristics that are used to develop the "recurring discourse" of academic writing.

This method proved to be helpful in Cortes' (2004) study of the difference between the lexical bundles in the published and student's writings in two subjects, history and biology. She discovered that there is discipline variation and writer variation sensitivity to lexical bundles. This suggests the potential use of lexical bundle analysis for studying developmental processes in academic writing and conventions in academic discourse.

Lexical Bundles in Student Academic Writing (2.2)

The analysis of student writing has repeatedly revealed that the use of lexical bundles is not as extensive as that of good academic writers. Student writers can use a relatively limited number of repetitive phrases and experienced writers can use a larger number. The discrepancies may be attributed to the academic discourse, disciplinary socialization or differences in language proficiency and rhetorical experience.

Lexical bundles thus play an important role in pedagogy. Students' fluency and control of academic discourse may be enhanced as a result of recurrent academic patterns. The Corpus can provide learners with authentic expressions, and also demonstrate how expressions are used in context.

But with the advent of generative AI, this pedagogical landscape is now becoming muddled. Students can now learn extremely traditional academic language without learning its linguistic knowledge of its usage. One of the key issues is if the lexical bundles that can be generated by AI-based academic writing can be more systematic than student writing?

2.3 AI-Generated Academic Writing

Generative AI has rapidly become a significant component of contemporary academic communication. LLMs can produce essays, summaries, explanations, literature reviews, and more for academic writing. Studies have therefore started to investigate whether artificial intelligence writings have a systematic difference from handcrafted writings.

For instance, Lu (2025) studied the relationship between lexical bundles and the accuracy of grammatical structures and between grammatical structures and EFL learners' writing-related outcomes in EFL writing through AI-generated corpus learning. The study shows that lexical bundles are playing an ever-greater role not just for the description of academic discourse but also as a means of facilitating AI-based language learning.

Azalia et al. (2025) conducted research into lexical bundle structure and function in texts from ChatGPT essays. According to their analysis, they are especially prevalent in noun-phrase based bundles and they also pointed out the importance of lexical bundles in indicating position and in developing academic ideas.

The most relevant recent comparison is provided by Tran (2026), who made a comparison between four-word lexical bundles in L1, L2 and argumentative texts generated by ChatGPT. The study used a large collection of student and AI-generated essays and discovered that the AI-generated words had a higher formulaic density.

These studies prove the relevance of lexical bundles for the topic of AI-generated writing, but also sets demands for more comparative studies using different corpora, different writing contexts, student groups and contexts, prompts, and analysis methods. The results of an AI system do not necessarily apply to all AI-generated academic texts because there are no guarantees that the results will be the same for different generations or versions of the AI system.

2.4 Research Gap

Three important points have been set forth in the literature. The first is lexical bundles are important in academic discourse. Secondly, they can vary in frequency, structure and function between registers, disciplines and writers. Third, new studies have shown that AI written texts can have characteristic formulaic patterns.

It remains a fairly new discipline, however. While previous work on lexical-bundle literature was primarily concerned with the human academic writing, more recent research on AI has begun to identify patterns that can be compared. Previous comparisons of L1 vs L2 vs ChatGPT have been performed, such as on a large scale in a 2026 study. The present study does not therefore suggest that there is no lexical-bundle comparison between AI and human existing so far. Rather, its contribution is to push this line of inquiry further by making a focused comparison of comparable academic essays and through simultaneous attention to frequency, structure and function.

The study is also a response to the need of careful comparability between the two corpora. Lexical-bundle frequencies can be affected by the differences in topics, length of the essay, proficiency, prompts, and genre. So it makes sense that such a comparison should be made with similar texts, as opposed to random samples of AI-generated and human-written texts.

CHAPTER 3

METHODOLOGY

3.1 Research Design

The research type applied in this research is corpus-based quantitative comparative research design to analyze the occurrence of lexical bundles of four-word in the academic essay written by students who use AI and students who do not use AI. Two similar sub-corpora were created: Human-Written Academic Essay Corpus (HWAEC) and AI-Generated Academic Essay Corpus (AIAEC). The corpora were matched as far as possible with respect to genre of the texts, topic, approximate length and demands of writing tasks.

The results of the research indicated that 4-word lexical bundles can be taken as relevant unit in studying the recurring phraseological patterns in academic text.

3.2 Corpus Construction

HWAEC consists of the essays written by the university students, while AIAEC consists of the essays created by AI using the similar or same writing prompts provided by the university students. The essays in both corpora are 20 in number, and have about 20,000 words in each corpus.

Only complete academic essays were included if they were complete. Texts that contain significant quotations, references, tables, headings and copied texts were excluded or cleaned in the same manner across both corpora.

Table 3.1

The organisation of the Corpora

corpus 1000 essays ~ 12,000 words

HWAEC	20	20,000 words
-------	----	--------------

AIAEC	20	20,000 words
-------	----	--------------

Total	40	40,000 words
-------	----	--------------

3.3 AI-Generated Corpus

The standardized prompts were used to create the AI generated essays, which were based on the writing tasks used for the human essays. When it was possible the same topics were used in both corpora. All information about the AI model, prompts, and generation was documented for methodological transparency and reproducibility.

Lexical Bundles are marked in 3.4.

The AntConc software was used to locate lexical bundles with the help of 4 words. Only "neighbouring" (i.e., 4-word sequences) were examined. The use of a frequency and dispersion criterion led to the exclusion of repeated sequences in only one essay as not representative lexical bundles. The first one was the frequency of the word in every one million words of the corpus, and the relevant number was adjusted to fit the size of the corpus.

3.5 Structural Classification

The bundles were identified and then classified according to the grammatical properties, primarily following the classification system as given by Biber et al. (1999). The largest structural groups were:

noun phrase/noun-phrase fragment;

prepositional phrase;

verb phrase;

dependent clause; and

Other clauses/other miscellaneous structures.

3.6 Functional Classification

Lexical bundles were also classified according to the discourse functions. The following were put into three general classifications:

Research-oriented/reference bundles, which identify, describe, quantify or characterize entities and ideas;

Discourse oriented bundles, that bind together discourse and create connections between propositions; and

Participant-oriented bundles, which are stance, evaluation, attitude, certainty, or writer-reader interaction.

3.7 Data Analysis

The analysis included four stages. There were four stages of analysis. Four stages were used in the analysis. Four-word bundles were first extracted qualifying for them. Second, their frequencies were adjusted on the basis of a corpus size. Third, all the bundles were structurally and functionally classified. Finally, the two corpora were compared concerning the frequencies, structural distribution, functional distribution, and shared and corpus specific bundles.

Normalized frequency was determined by:

Normalized Frequency = Raw Frequency / Corpus Size x 1,000,000

CHAPTER 4

RESULTS AND ANALYSIS

4.1 Introduction

This chapter presents the results of the comparative analysis of four-word lexical bundles in the **Human-Written Academic Essay Corpus (HWAEC)** and the **AI-Generated Academic Essay Corpus (AIAEC)**. The analysis examines the overall frequency of lexical bundles, their most frequent forms, structural distribution, and functional distribution. The results are based on the corpus design and analytical procedures described in Chapter 3.

4.2 Overall Distribution of Lexical Bundles

The first stage of analysis examined the distribution of four-word lexical bundles in the two corpora. Table 4.1 presents the overall results.

Table 4.1

Overall Distribution of Four-Word Lexical Bundles

Measure	HWAEC	AIAEC
Total words	20,184	20,436
Number of essays	20	20
Four-word bundle types	118	143
Total bundle occurrences	476	621
Normalized frequency per million words	23,581.45	30,386.77
Shared bundle types	61	61
Corpus-preferred bundle types	57	82

The results show that the AI-generated corpus contains more four-word lexical bundles than the human-written corpus. HWAEC contains 118 bundle types with 476 occurrences, whereas AIAEC contains 143 bundle types with 621 occurrences. After normalization, the AI-generated corpus records **30,386.77 occurrences per million words**, compared with **23,581.45 occurrences per million words** in the human-written corpus.

This indicates that AI-generated essays demonstrate a greater concentration of recurrent four-word sequences. The two corpora nevertheless share 61 lexical-bundle types, showing considerable overlap in their use of conventional academic phraseology.

4.3 Most Frequent Lexical Bundles

Table 4.2 presents the ten most frequent four-word lexical bundles identified in each corpus.

Table 4.2

Most Frequent Lexical Bundles

Rank	HWAEC Bundle	Frequency	AIAEC Bundle	Frequency
1	in the context of	18	in the context of	26
2	on the other hand	16	it is important to	24
3	as a result of	15	as a result of	22
4	the role of the	14	in terms of the	21

Rank	HWAEC Bundle	Frequency	AIAEC Bundle	Frequency
5	the use of the	13	the role of the	20
6	in terms of the	12	the use of the	19
7	the extent to which	11	the extent to which	18
8	one of the most	10	one of the most	17
9	it is important to	9	on the other hand	15
10	in order to understand	8	in order to understand	14

The table shows substantial similarity between the two corpora. Several bundles, including *in the context of*, *as a result of*, *the role of the*, and *the use of the*, occur among the most frequent bundles in both corpora.

However, the frequencies are generally higher in the AI-generated corpus. For example, *in the context of* occurs 18 times in HWAEC compared with 26 times in AIAEC. Similarly, *it is important to* occurs nine times in HWAEC but 24 times in AIAEC.

The results suggest that AI-generated essays rely more heavily on highly conventional academic expressions. Human-written essays also make extensive use of conventional bundles, but their frequencies are comparatively lower.

4.4 Structural Distribution of Lexical Bundles

The extracted lexical bundles were classified according to their grammatical structures. The results are shown in Table 4.3.

Table 4.3

Structural Distribution of Lexical Bundles

Structural Category	HWAEC N	HWAEC %	AIAEC N	AIAEC %
Noun phrase/NP fragment	42	35.6%	58	40.6%
Prepositional phrase	31	26.3%	38	26.6%
Verb phrase	20	16.9%	24	16.8%
Dependent clause	15	12.7%	15	10.5%
Other structures	10	8.5%	8	5.5%
Total	118	100%	143	100%

Noun-phrase and noun-phrase fragment bundles constitute the largest structural category in both corpora. They account for 35.6% of the bundles in HWAEC and 40.6% in AIAEC. Thus, AI-generated essays show a stronger preference for noun-phrase structures.

Prepositional phrase bundles are the second most frequent category, representing 26.3% of HWAEC and 26.6% of AIAEC. The relatively similar percentages indicate that both groups use prepositional bundles at comparable levels.

Verb-phrase bundles also show almost identical proportions, accounting for 16.9% of human-written bundles and 16.8% of AI-generated bundles.

Dependent-clause bundles are somewhat more prominent in HWAEC, while other structural categories occur slightly more frequently in the human corpus than in the AI corpus.

Overall, the structural findings indicate that the two corpora share broadly similar grammatical patterns, but AI-generated essays demonstrate a stronger concentration of noun-phrase bundles.

4.5 Functional Distribution of Lexical Bundles

The lexical bundles were further classified according to their discourse functions. Table 4.4 presents the results.

Table 4.4

Functional Distribution of Lexical Bundles

Functional Category	HWAEC N	HWAEC %	AIAEC N	AIAEC %
Research-oriented/reference	49	41.5%	61	42.7%
Text-oriented	47	39.8%	61	42.7%
Participant-oriented	22	18.7%	21	14.6%
Total	118	100%	143	100%

Research-oriented/reference bundles constitute the largest functional category in HWAEC, accounting for 41.5% of the identified bundles. They are also highly frequent in AIAEC, where they account for 42.7%.

Text-oriented bundles represent 39.8% of the human-written bundles and 42.7% of the AI-generated bundles. The slightly higher proportion in AIAEC indicates a greater use of bundles associated with organizing and connecting ideas.

Participant-oriented bundles account for 18.7% of HWAEC but only 14.6% of AIAEC. This suggests that human writers use proportionally more expressions associated with stance, evaluation, and writer-reader interaction.

The functional results therefore demonstrate that AI-generated essays rely somewhat more heavily on research-oriented and text-organizing expressions, while human-written essays show a comparatively greater presence of participant-oriented bundles.

4.6 Shared and Corpus-Specific Bundles

The comparison also identified lexical bundles shared by both corpora and those occurring primarily in one corpus.

Table 4.5

Shared and Corpus-Specific Bundles

Category	Number of Bundle Types
Shared bundles	61
HWAEC-preferred bundles	57
AIAEC-preferred bundles	82
Total distinct bundles	200

The two corpora share 61 lexical-bundle types, demonstrating considerable phraseological similarity. However, 57 bundles are primarily associated with the human-written corpus, while 82 are primarily associated with the AI-generated corpus.

The larger number of AI-preferred bundles corresponds with the overall higher number of bundle types found in AIAEC. This suggests that AI-generated essays employ a wider range of recurrent conventional sequences within the parameters of the present corpus.

The shared bundles also demonstrate that AI-generated academic writing does not employ a completely separate phraseological system. Rather, it draws substantially on the same conventional academic expressions found in human-written essays.

CHAPTER 5

DISCUSSION

5.1 Introduction

It discusses the crucial results of the study as related to the research objectives and the earlier research on the lexical bundles and academic writing. The study examined the four-word lexical bundles from the Human-Written Academic Essay Corpus (HWAEC) and the AI-Generated Academic Essay Corpus (AIAEC). The overall frequency of lexical bundles as well as the most frequent forms of lexical bundles, structural features of lexical bundles, functional categories, similarities and differences between the two corpora are discussed.

The results show that there is a higher number of four-word sequences that appear more frequently in AI-generated academic essays than in human-written essays. There are also a large number of common lexical bundles between the two corpora, indicating that there are strong relationships between the existing academic phrase structure in the learner's academic text production and AI's own text production. Further, any variations in the distribution of bundles also suggested that the organization of academic discourse was not the same for AI-generated essays and human-written essays.

5.2 Overall Frequency of Lexical Bundles

The first main conclusion is that there are more four-word lexical bundles in AI-generated essays than in the human-written essays. There were 621 occurrences of bundles in the AI-generated corpus, while there were 476 occurrences of bundles in the human-written corpus. After normalization, the bundle frequency per million words of the AI corpus was higher.

The results indicate that AI-generated academic texts are more inclined towards formulaic and repetitive expressions. This type of behavior can be attributed to the language generation mechanism of generative AI. The system behind AI is trained on a large collection of texts, which it uses to generate a sequence that is most likely to follow the structure of the language. It is likely that conventional academic expressions will be used in generated texts often.

The finding is special in two aspects: first, lexical bundles are not necessarily very frequent words and, second, the occurrence of lexical bundles in discourse is not always apparent. They are patterns that recur and help organize and build discourse. In Biber et al. (1999) the authors have demonstrated that multi-word sequences is one of the important features of academic discourse. This is a relatively high rate of bundles in AIAEC, therefore, AI-generated essays heavily depend on the academic jargon.

However, more frequent is not necessarily more suitable and more academically appropriate AI writing. Academic writing also requires originality, fluency of argumentation, critical engagement, and context appropriate uses of language; while formulaic writing may be helpful in developing fluency and organization in the text.

5.3 similarity in the Most Frequent Bundles

The most frequent lexical bundles in the two corpora are quite similar, though there is variation in frequency. The words in the context of, in the result of, the role of the, and the use of the, occur very often in both corpora.

It appears that there is some commonality between human and AI use of common academic terms. Academic writing has repetitions of phraseological structures, especially those structures that introduce ideas, create connections, show cause/effect and provide structure for arguments.

This is not surprising as the same bundles are found in both corpora. AI-generated essays are written in the style of academic writing and thus follow many of the patterns of human academic writing.

What bundle types are being used is not really the important factor, it's how much they're used. For instance, in the case of occurs, it is found 18 times in HWAEC and 26 times in AIAEC. Likewise, it is a lot more common in the AI-generated corpus.

This means that the AI-generated written content is more likely to replicate more traditional academic writing, whereas the human writers might be more inclined to employ a variety of phrases.

The 5.4 Structural patterns of lexical bundles looks into the structural patterns of lexical bundles.

From the structural analysis, it can be seen that in both corpora, the most common structural category is noun-phrase and noun-phrase fragment bundles. The proportion is greater in their case in the AI-generated writing, however.

This result is similar to that of the nominal structures' significance in academic discourse. Noun phrases are often used to package complex ideas, represent abstract concepts and to condense information in academic writing. The role of the and the use of the allows writers to build informational units.

The higher ratio of noun phrase bundles in the AI generated texts suggests a greater inclination toward the conventional Academic Information Packaging of texts. AI created essays often use a structure similar to that of formal and informational writing which is often found in academic writing.

The results for prepositional bundles and verb-phrase bundles are fairly similar in the two corpora. This indicates that there is a certain degree of stability in some grammatical structures between academic writing produced by humans and by AI.

In the human corpus, however, there are fairly higher numbers of dependent-clause and miscellaneous structures. This may indicate more variability in the way that human writers form propositions and link ideas together. Human writers might have more flexibility in their structure decisions and not go back to overly conventional patterns.

5.5 Functional Patterns of Lexical Bundles.

The functional analysis confirms the differences between the two kinds of academic texts. The principle categories in both corpora are research-oriented and reference bundles, and text-oriented bundles. This is in line with the communicative demands of academic writing.

Writer can identify and characterize concepts and establish references using the research-oriented bundles. In academic writing, topics are introduced, phenomena are described, and relationships between concepts are discussed. This category is therefore quite common in both corpora, and represents one of the basic features of academic discourse.

The proportion of research-oriented bundles in the AI-generated corpus is slightly higher than that in the human corpus. This could suggest that traditional language for starting and presenting academic ideas is heavily used in AI-generated essays.

Text oriented packages are also very common, with an emphasis on the AI generated corpus. These linkages are used to create logical relationships among propositions and to direct the reader through an argument, and so are known as these bundles. Dense occurrences of text-oriented bundles in AI-generated writing can help to create a sense of organization and coherence in the produced text.

On the other hand and as a result are explicit signals of relations between ideas. The following are some common phrases used to structure conventional essays that are often made use of in AI-generated essays. The following are some common phrases for structuring conventional essays that are often used in AI-generated essays.

Participants can design their own bundles and write about themselves.

One of the more noticeable differences is found in participant-oriented bundles. The human written corpus has a larger proportion of these bundles than the AI written corpus.

Participant oriented bundles are related with attitude, certainty, evaluation, stance and interaction between the writer and the reader. The higher representation of these in human writing may suggest more writer perspectives being involved in the creation of academic arguments.

Academic writing is a product of the writer's own experiences, judgments, confidence and rhetorical tastes. As such, they may employ stance related expressions in different ways depending on the communicative goals of each speaker.

AI-generated writing, on the other hand, might have a preference for more formal and abstract academic language. While stance expressions can be produced by AI systems, the relatively small percentage in the present corpus points to a greater preference for informational/organisational functions than for interpersonal positioning.

This is important because academic writing is not so much about information as it is about ideas. In addition, effective academic discourse involves stating claims, assessing evidence, stating degrees of certainty, and establishing relationships with the reader. This distinction between the two corpora thus reflects an important aspect of human – AI differences.

Shared and Corpus-Specific Bundles.

The study results revealed that there are 61 shared lexical bundles in human written and AI generated academic writing. This discovery lends some weight to the idea that AI-generated writing does not represent yet another language of its own making but is part of the broader academic linguistic framework.

The two corpora, on the other hand, show different phraseological profiles, as can be proved by the presence of corpus-preferred bundles. AI also has a higher frequency of bundle types, as well as a greater number of preferred bundle types, consistent with the greater overall bundle frequency.

These corpus specific patterns are especially helpful to differentiate between the two types of writing. The similarities between human and AI writing could be even more striking, if only shared bundles were analyzed. Analysis of the corpus shows that the two types of writing differ in terms of the amount and frequency of repetition of phrases.

The results accordingly warranted a balanced conclusion: AI-generated academic writing is similar to and different from human-written academic writing. It mimics many existing academic patterns, but exhibits increased use of some recurring sequences.

For Academic Writing, implications will be considered.

The results bring a number of implications for academic writing. Firstly, lexical bundles are commonly found in AI-generated compositions, so students and teachers must be aware of the occurrence of formulaic expressions in addition to correct grammar.

If used correctly, lexical bundles can enhance the cohesion and fluency of the text. But, if too many of the conventional expressions are used, the writing may become repetitive or predictable. To learn lexical bundles, students should therefore learn to understand what they are doing rather than memorizing or reproducing.

Second, the findings have implications for the teaching of academic writing. Lexical bundles can be explicitly introduced as helpful resources to present arguments, make relationships, present concepts, or set a stance. Students can become aware of the functions of such expressions in real academic writing with the support of a corpus-based approach.

Thirdly, the results are applicable to the increasing application of generative AI in academic writing. AI-generated essays often sound very academic due to the fact that they are full of academic terms. It is therefore important for teachers and evaluators to not rate the quality of writing based exclusively on formal academic terminology.

Lexical bundles may be used too often in a text without it having originality, critical thinking or sufficiently developed arguments. The lexical-bundle analysis can detect some crucial linguistic characteristics, but cannot, alone, be used to evaluate the overall quality of academic writing.

5.9: Implications for AI-assisted Academic Writing.5.9: Implications for Academic Writing with AI.

The results also have implications for the responsible use of AI in academic writing. The relatively high rate of lexical bundles in traditional texts suggests that AI can generate texts that share characteristics of traditional academic language.

This may be helpful for students who need help structuring ideas or for those who are not familiar with some of the more common expressions used in school. But, the student should not solely rely on AI-generated formulations. Human writers still have to verify the accuracy, relevance, originality, and appropriateness of the language generated by AI systems.

The results also indicate that AI written text could be improved through human editing. Revision can eliminate over repetition, add more expressions that are context specific, and enhance writer stance as needed. So, AI can be considered a language generation tool instead of a human academic judgment tool.

5.10 Conclusion

This research was an exploratory quantitative corpus comparison study on four word lexical bundles in human and AI written academic texts. The results show significant similarities and significant differences between the two kinds of academic writing.

More lexical bundles are found and more frequently in the AI-generated corpus than in the human-written corpus. This indicates that AI text tends to be more formulaic in academic writing. However,

many of the most common bundles are common to both corpora, showing that much of the writing produced by AI models is based on mainstream forms of academic English.

From the structural analysis, it is clear that noun-phrase and noun-phrase fragment bundles dominate both the corpora with a greater proportion in the AI-generated writing. The functional analysis shows that research-oriented bundles and text-oriented bundles represent the majority of the bundles in both corpora, and that participant-oriented bundles are more frequently occurring in the human written essays.

Overall, the study demonstrates that lexical bundles are a useful tool in the identification of difference between corpora composed of artificial and human academic discourse. The results indicate that AI-created writing was more formulaic and conventional than human-written writing, while human writing had more interpersonal and phraseological variation.

The study thus adds to the growing literature on how to approach the study of academic discourse generated by AI in both vocabulary and grammatical accuracy, but also in terms of recurring multi-word phrases.

REFERENCES

- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Longman.
- Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1), 4–21.
- Hyland, K. (2008). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, 18(1), 41–62.
- Hyland, K. (2009). Academic discourse: English in a global context. Continuum.
- Biber, D., Conrad, S., & Cortes, V. (2004). If you look at ...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), 371–405. <https://doi.org/10.1093/applin/25.3.371>
- Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, 26(3), 263–286.
- Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, 23(4), 397–423. <https://doi.org/10.1016/j.esp.2003.12.001>
- Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, 31(2), 81–92. <https://doi.org/10.1016/j.esp.2011.08.004>
- Simpson-Vlach, R., & Ellis, N. C. (2010). An Academic Formulas List: New methods in phraseology research. *Applied Linguistics*, 31(4), 487–512. <https://doi.org/10.1093/applin/amp058>
- Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press.
- Chen, Y.-H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, 14(2), 30–49.
- Bychkovska, T., & Lee, J. J. (2017). At the same time: Lexical bundles in L1 and L2 university student argumentative writing. *Journal of English for Academic Purposes*, 30, 38–52.

Anthony, L. (2005). AntConc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom. In Proceedings of the 2005 IEEE International Professional Communication Conference (pp. 729–737). IEEE.

A corpus-driven comparative analysis of AI in academic discourse: Investigating ChatGPT-generated academic texts in social sciences. (2024). *Lingua*, 312, 103838.

<https://doi.org/10.1016/j.lingua.2024.103838>