ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

A MULTIDIMENSIONAL CORPUS-BASED FRAMEWORK FOR ANALYZING LINGUISTIC COMPLEXITY AND STYLISTIC VARIATION IN HUMAN AND AI-GENERATED ACADEMIC WRITING

Kainat Zahid[1] and M. Owais Ayaz[2]*
[1] PhD Scholar, Department of English, Kohat University of Science and Technology, Kohat, Pakistan
[2] PhD Scholar, Department of English, Kohat University of Science and Technology, Kohat, Pakistan

Abstract
The development of large language models (LLMs) has drastically altered academic writing. Systems currently exist that can write scholarly prose with context, grammar, and coherence. While these systems can write at a level similar to that of human scholars, important issues remain regarding the linguistic complexity and discourse of these systems, as well as the range and type of styles they can write in. Prior studies on issues of lexical diversity, syntactic complexity, or formulaic language tend to investigate these separate factors rather than considering academic writing as a multifaceted linguistic phenomenon. This study aims to develop a multidimensional system for reviewing academic writing to compare human and AI writing. The system will utilize lexical diversity, lexical sophistication, syntactic complexity, discourse cohesion, metadiscourse, readability, information density, and stylistic variation, among other features. Two corpora, one written by humans and one written by AI, will be constructed for review. The Human Academic Writing Corpus will be constructed with research articles, and the AI Academic Writing Corpus will be constructed using texts generated by AI in a restricted writing environment. The main theory that this study is grounded in is Biber's (1988) work, although other theory and measures of lexical diversity, syntactic complexity, and metadiscourse will be integrated. This study aims to provide a framework beyond simple AI detection that considers the range of linguistic features in which human and machine academic writing converge and diverge. The framework is built to be used and researched in corpus linguistics, computational stylistics, English for Academic Purposes, academic discourse studies, and research related to generative AI.

Keywords*: Corpus Linguistics; Artificial Intelligence; Academic Writing; Linguistic Complexity; Computational Stylistics; Large Language Models; Multidimensional Analysis*

1. Introduction

1.1 Background and Problem

Discourse is structured to combine lexis, grammar, syntax, rhetoric, and other elements of discourse to achieve the functions of evaluation or description of knowledge, the construction of an argument, critique, incorporation of existing scholarship, and placement of a given proposition/claim in the domain of a given discipline. For these reasons, academic writing cannot simply be defined in terms of one linguistic feature.

The development of AI has altered conventional thinking on text generation systems. State-of-the-art AI systems or LLMs can generate coherent and grammatically accurate extended pieces of academic writing. They can produce introductions, reviews of the literature, methodology, analysis, discussion, and conclusions based on brief prompts. Thus, an important research question is whether the linguistic characteristics of AI-generated academic writing are similar to those of human writers of academic writing.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

This question moves beyond grammar. Texts generated by a human and AI may both use sophisticated vocabulary and grammatical structures but may vary in the distribution, combination, and use of linguistic features. A text may have a lot of lexis but not much syntax, while another may have a lot of discourse markers and not a lot of rhetoric.

Because linguistic variation cannot be reduced to one dimension, Biber's (1988) approach is appropriate here. Varying linguistic features may occur together and describe/evaluate a piece of text. This approach is useful to view human and AI writing in a multi-faceted way instead of a binary way.

Research Gap

Prior research has usually addressed lexical, syntactic, discourse, and readability features in isolation. Although these studies are a good starting point, they could potentially miss the relationships between linguistic dimensions. In addition, the average linguistic behavior of a group has been a greater focus than differences in the linguistic behavior of individuals.

The current model aims to overcome these limitations by incorporating multiple linguistic dimensions in a unified corpus-based model. This model deals with between-group variation, or the differences between human and machine-generated texts, as well as within-group variation, or differences among the texts in a given corpus.

Research Objectives

Objective 1: To analyze and compare the lexical, syntactic, discourse, and stylistic elements of human- and AI-authored essays in an academic context through a multidimensional corpus analysis method.

Objective 2: To analyze how variations in linguistic complexity and style define human-written academic texts from AI-generated academic texts.

Research Questions

RQ1: What lexical, syntactic, discourse, and stylistic features distinguish human-authored academic writing from AI-generated academic writing?

RQ2: To what extent do human- and AI-generated academic texts differ in their multidimensional profiles of linguistic complexity and stylistic variation?

Significance

The framework has value that is theoretical, methodological, and pedagogical. In the theory section, it creates a new dimension to contemporary human-machine discourses. In the methodology section, it uses multiple corpus-based techniques as opposed to a single approach. Lastly, it provides EAP (English for Academic Purposes) researchers and instructors insights on what distinguishes surface fluency from other intricate dimensions of academic linguistic communication. Lastly, it aids in the discussion of the ethical responsibility of using AI, as it focuses on looking at large language corpora as opposed to determining the authorship of single-text documents.

2. Literature Review and Theoretical Foundation

2.1 Corpus Linguistics and Academic Writing

Corpus linguistics is the systematic study of language using the collection of textual data. Researchers use frequency analysis, keyword analysis, concordancing, and collocation to name a few. They use the tools to find recurring patterns in their data. Because of its predictable nature, academic writing is an excellent field for application of corpus analysis. Writing provides a variety of resources for argumentation, evaluation, citation, stance, and audience.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

When comparing human and AI writing, there is a unique advantage of corpus-based studies. The studies are able to use the same protocols and methods in analysis of each group. Instead of measuring the difference in writing through subjective analysis, the difference can be measured quantitatively.

2.2 Multidimensional Analysis

The theory this study is based upon is Biber's (1988) multidimensional analysis. The theory describes the grouping of linguistic features and how the co-occurring groups can be used to identify different textual dimensions. Biber arguedthat although texts can differ in many ways, pairs of texts that are similar on one dimension can be dissimilar on another.

This theory can easily be applied to the comparison of human-AI writing. Texts generated by both humans and AI can be similar in the use of formal and traditional academic structures. Yet, the lexical choices, syntax, discourse considerations, and the overall interpersonal concerns can differ. Thus, the multidimensional analysis of this study will consider the integration of various features as opposed to the analysis of single features.

2.3 Lexical and Syntactic Complexity

Lexical diversity concerns the range and distribution of vocabulary. A traditional measure is the type-token ratio (TTR):

TTR = V / N (2.1)

where V represents the number of different word types and N represents the total number of running words.

TTR is sensitive to the length of a text. McCarthy and Jarvis (2010) studied more sophisticated metrics such as MTLD, vocd-D, and HD-D, which offer different ways to address the comparability of lexical diversity in texts of disparate lengths.

Lexical sophistication is the use of less common, specialized and/or scholarly appropriate vocabulary. However, lexical sophistication should not be confounded with high writing quality because vocabulary choice is also dependant on discipline, topic, and rhetorical aim.

Syntactic complexity is the extent and the way of grammatical elaboration. In academia, writing typically uses subordinate clauses, complex noun phrases, participial phrases, and structural embedding. Lu (2010) asserted that syntactic complexity is multi-faceted and therefore should be addressed using multidimensional assessment vs. a unidimensional assessment. Relevant assessment metrics should include length of sentences and T-units, clauses per sentence, dependence, subordination, and complex nominals.

2.4 Discourse and Metadiscourse

To understand the nature of lexical connections and transitions between ideas, academic discourse requires both lexical units and propositions. Cohesion can take several forms, including connectives, semantic connections, lexical and grammatical repetition, and reverence. AI-generated text might be lacking in subtle relationships and might thus have a reliance on transitions and discourse markers. Conversely, human writing might incorporate implicit and explicit relationships to a greater extent.

Another of Hyland's (2005) dimensions is metadiscourse. He differentiated between stance and engagement, where stance encompasses markers, hedging, boosting, and self-references, while engagement refers to the text and how the writer addresses and manages the reader. LLMs being able to parrot conventional forms in writing may demonstrate differences in positioning through the frequency and distribution of metadiscourse resources.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Therefore, the literature supports the integration of lexical, grammatical, discourse, and informational features.

3. Conceptual Framework

The study defines linguistic complexity as a multidimensional construct of lexical, syntactic, discourse, and semantic elements. Lexical complexity is the range of vocabulary used; syntactic complexity is the use of structures and embedment; discourse complexity is the interconnection of propositions and textual units; and semantic complexity is the packing of information.

Variation is defined as the extent to which linguistic choices differ from text to text. Two types of variation are recognized.

1. Between-group variation is the difference between human and AI-generated texts.
2. Within-group variation is the difference among the texts in the same group/corpus.

The second dimension is important because two corpora may have the same mean, but show substantial differences in internal variation.

The proposed framework contains six main dimensions:

Dimension	Representative Indicators
Lexical complexity	MTLD, HD-D, TTR, lexical sophistication
Syntactic complexity	Sentence length, T-unit length, clause density, subordination
Discourse cohesion	Connectives, reference, lexical overlap, semantic relations
Metadiscourse	Hedges, boosters, transitions, self-mentions, engagement
Information/readability	Lexical density, sentence length, readability
Stylistic variation	Standard deviation, variance, coefficient of variation

The overall analytical sequence is:

Corpus → Linguistic Features → Linguistic Dimensions → Statistical Comparison → Multidimensional Profile → Interpretation

This model proposes that individual linguistic features must be contextualized in terms of the overall architecture of features.

4. Methodology

4.1 Research Design and Corpus Construction

This study uses a quantitative comparative corpus-based research design. Two corpora will be developed.

Human Academic Writing Corpus (HAWC): peer reviewed academic research articles written in English.

AI Academic Writing Corpus (AIAC): academic texts written by an LLM under controlled writing conditions.

This study is a corpus based comparative study rather than an AI detection study.

The human corpus will be defined using specific criteria for genre, discipline, language, topic and rhetorical purpose. The AI corpus will be constructed using prompts designed to capture the same communicative purpose and disciplinary features of the human corpus. The prompts will articulate

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

the academic topic and genre, a target length within a given disciplinary context, a given rhetorical purpose, and a requirement for formal academic English.

In order to facilitate a valid comparison the corpora will be matched as closely as possible for genre, discipline, topic, purpose, text length, number of texts and size of the overall corpus. Normalized frequencies will be used to compensate for differences in corpus size.

4.2 Data Preparation and Analytical Tools

Machine-readable format texts will undergo standardized preprocessing. Reference lists and author affiliations, along with any accompanying unrelated tables, captions and metadata, will be eliminated. Tokenization, sentence segmentation, part-of-speech tagging, lemmatization, syntactic parsing and manual quality checks will be conducted as part of the preprocessing.

To accomplish the foregoing analysis, AntConc or a similar corpus analysis tool will be employed to conduct concordancing, keyword analysis, frequency, collocation and lexical bundling analysis. Syntactic complexity will be assessed using the existing computational assessments for measuring syntactic complexity inspired by Lu (2010), and lexical diversity will be assessed using MTLD and HD-D, following the work of McCarthy and Jarvis (2010). It is expected that statistical analysis will be accomplished in either R or Python.

4.3 Lexical and Syntactic Analysis

Lexical analysis considers the use of vocabulary in terms of variety, sophistication, use frequency, and degree of formulaicity. While the use of TTR is limited because of length, this analysis will also utilize MTLD and HD-D to mitigate the impact of length on the results.

Normalized frequency will be calculated as:

$$NF = (f / N) \times 1{,}000 \quad (4.1)$$

where f represents the frequency of a linguistic feature and N represents the total number of words. Keyword analysis will seek to identify linguistic items occurring with greater or lower frequency in one corpus compared to the other.

Syntactic analysis will consider the average length of a sentence, T-unit length, average clauses per sentence, the presence of dependent clauses, subordination, and complex nominal groups. The average sentence length will be calculated as:

$$MSL = W / S \quad (4.2)$$

where W represents the total number of words and S represents the total number of sentences.

Clause density will be determined as:

$$CD = C / S \quad (4.3)$$

where C represents the total number of clauses and S represents the total number of sentences.

These metrics will be interpreted collectively because greater length of a sentence in no way supports the presence of greater sophistication of syntax.

4.4 Discourse, Metadiscourse, and Information Density

Cohesion will be evaluated in terms of reflection, discourse connectors, lexical repetition and overlap, cohesive-semantic ties and homonymy. Contextual concordance analysis and frequency counts will be employed to describe the functions of discourse cohesive devices.

Hyland's (2005) model describes the structure and components of metadiscourse to be stance and engagement resources. Stance includes lexical hedges, boosters, sentiments and self-references, while engagement resources include references to the readers and commands. In order to ensure representative cross-corpus comparisons, I will adjust the frequencies.

Readability will be used as an indirect measure of writing quality. I will also look at information density by measuring lexical density:

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

$LD = (L / W) \times 100$ (4.4)

where L represents the lexical words and W represents the total words.

4.5 Statistical and Multidimensional Analysis

Descriptive statistics include mean, median, standard deviation, variance and of course the coefficient of variation that in this case will be calculated as follows:

$CV = \sigma / \mu$ (4.5)

where σ and most likely you know what that represents, and μ not surprisingly will be the mean.

Independent-samples t-tests will be implemented. Should parametric assumptions be violated, Mann-Whitney U tests will be implemented. For both, effect sizes will be presented to give an indication of the practical outcome of the differences tested.

The main many dimensional analysis will seek to discern relationships amongst linguistic variables. Those relationships will be considered significant and relevant enough and then subjected to factor analysis or principal component analysis (whichever is more practical). The resulting factor scores analysis will be undertaken to determine the differences between the human and AI corpora. This will be done following Biber's (1988) assumption that linguistic variation results from the interrelation of different factors.

Reliability will be enhanced via standardized preprocessing and corpus selection coupled with automated measurement. Construct validity will be enhanced by the application of multiple measures of each of the major constructs. Corpus validity will be a function of careful matching of the two corpora.

4.6 Ethical Considerations

The study will comply with copyright and licensing for published human authored materials. AI-generated texts will be ID'd, and their generation explained, to improve reproducibility. The study won't make definitive statements about authorship for individual texts. The focus is on identifying corpus level linguistic patterns.

5.Analytical Framework and Expected Interpretation

The study was developed to generate empirical findings as opposed to making assumptions. Analysis will include descriptive analysis, feature-based comparison, multidimensional analysis, intragroup stylistic analysis and integrated analysis.

Lexical analysis will assess the diversity, sophistication, and formulaicity of vocabularies for both human and AI generated texts. Syntactic analysis will examine the structure and subordination of sentences as well as the density and variance of constituent structures. Discourse analysis will assess cohesion and the organization of information. Evaluation of metadiscourse will examine stance and the orientation and engagement of the reader (Hyland, 2005).

Emphasis will be on the extent and nature of dispersion of styles. Higher standard deviation for AI writing may indicate a greater level of homogeneity compared to human writing that is more variable. The final analysis will synthesize all lexical, syntactic, discourse, metadiscourse, informational and stylistic elements, as opposed to relying on one of these alone.

6. Discussion and Implications

This framework considers how human and AI-generated academic writing are becoming more similar. LLMs can easily produce academic language, so surface fluency and correctness of grammar will not be sufficient to distinguish robust linguistic profiles.

Biber's (1988) approach is ideal as it considers multiple linguistic elements as opposed to considering one aspect of language in isolation. Therefore, this framework adopts a holistic consideration of lexical, syntactic, discourse and interpersonal elements.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Lexical sophistication, syntactic complexity, cohesion, and metadiscourse could be the differentiating factors between human and AI-generated texts. However, these features should be interpreted within the context of the writing as highly sophisticated language or great depth of cohesion does not necessarily imply good academic writing. Similarly, one long sentence does not represent great syntactic complexity.

This framework will make contributions to African Applied Linguistics and may help to identify more accessible linguistic patterns of AI-generated writing. This framework may also enable the separation of surface fluency from deep rhetorical writing. More research is needed to explore the collaborative writing between humans and AIs, understanding that the line between writing generated by AIs and writing that is purely human is becoming more blurred.

7. Limitations and Future Research

The framework has multiple limitations. AI-generated texts can differ based on the model, version, Prompts, and settings. Human academic writing can differ based on several factors including discipline, culture, country, university, and author. Automated linguistic analysis can include errors for tagging, parsing, and semantic processing.

Future work should include longitudinal analyses of academic language. This could include a variety of LLMs. The collaborative writing process with AI and humans is another important avenue of research. New studies could also include interdisciplinary work and machine learning that prioritizes interpretability. These studies may determine if multiple linguistic features can consistently differentiate between human and AI text.

8. Conclusion

The introduction of generative AI shifts how we examine discourse in academic settings. Although they can often produce high-quality academic prose, similarities at a surface level do not help us understand how these textual fragments are related.

The main contribution of this paper is the development of a new multidimensional corpus methodology composed of lexical complexity and diversity, syntactic complexity, cohesion, metadiscourse, information density, and style. This methodology is largely based on Biber (1988) and developed research on lexical complexity and diversity, syntactic complexity, and metadiscourse in an academic context (Hyland, 2005; Lu, 2010; McCarthy Jarvis, 2010).

The main objective is to move away from analyzing features in isolation. By examining the co-occurrence of linguistic features and multidimensional stylistic profiles, we should study both human and AI-generated academic writing. Additionally, when studying AI-generated texts, it is also important to consider how standardized AI writing is compared to human writing that may include more variation.

This paper aims to analyze how the differences in language both human and AI systems utilize to construct academic language differs, rather than focusing heavily on whether a particular sample is or is not machine-generated writing. The author believes the framework will make useful additions to the fields of corpus linguistics, computational stylistics, EAP, and academic writing research.

References

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

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. Continuum.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474–496.

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*, 381–392.

Molodtsov, D. (1999). Soft set theory—First results. *Computers & Mathematics with Applications, 37*(4–5), 19–31.