AL-AHQAQ **Recognized by HEC in Y-Category**
https://ahqaq.com/index.php/Journal/issue/view/9 **ONLINE - ISSN- 3006-595X**
Volume.4, Issue.1 (2026) **PRINT - ISSN- 3006-5941**

The Stylistic Fingerprint of AI: Mapping Lexical Bundles and N-Grams in Automated Academic Writing

Noshaba Fraz
Air University Kamra, Pakistan, Email. noshaba.fraz@gmail.com

Abstract

Large Language Models (LLMs) have transformed students' approach to high-stakes English language assessment, including the widely used International English Language Testing System (IELTS). Although LLMs produce grammatically correct sentences, they are based on probabilistic models that focus on likely rather than unusual linguistic patterns rather than the idiosyncratic variations associated with natural human fluency. The present study uses a corpus-driven analysis to explore the formulaic stylistic pattern(s) that emerge from the IELTS Academic Writing Task 2 essays generated by LLMs (or "AI-ese"). This study builds and analyzes a specially designed corpus of 20 model essays and finds that the majority of the essays share a unique, algorithmically extracted stylistic fingerprint with a great deal of repetitive 3-word and 4-word lexical bundles. The findings reveal that the model always uses the pre-packaged rhetorical units, for instance, as is widely accepted and the modern era, to build up the illusion of academic objectivity. The aim of this study is to show that these formulaic sequences are structurally correct but the epistemic stance markers and authorial presence they generate are not sufficiently complex or sophisticated in accordance with the evaluation principles of IELTS. The findings indicate that relying on AI-generated models without critical evaluation for test preparation might be detrimental to the cultivation of true communicative ability and to the test takers' performance.

Keywords:

IELTS, Large Language Models, Corpus Linguistics, Lexical Bundles, AI-ese, Academic Writing

Introduction

Large Language Models (LLMs) have revolutionized the academic landscape by fundamentally changing how students engage with academic writing. With the advent of high-stakes English proficiency exams like the International English Language Testing System (IELTS), students are increasingly using AI tools to create what they call model essays for studying and learning the skills required for this test. These models can create grammatically correct texts, but they are based on probabilistic structures that produce texts that are more likely to be linguistically "highly probable" than the natural variation seen in human speech. This technological move raises a significant pedagogical question: can AI replicate the form of an academic essay, but at the same time reinforce a highly repetitive, robotic lexicogrammar, often referred to as "AI-ese," which could be detrimental to the acquisition of authentic communicative competence for students aiming for it?

The essence of this stylistic homogeneity is the use of a large number of formulaic language units, namely lexical bundles, which are recurring combinations of 3 or 4 words used as the units of academic discourse. As the previous studies have shown, human writers tend to use these bundles to convey their stance and complex argumentation, whereas AI models seem to be more limited in their choice of words and stances (Jiang & Hyland, 2024). The consistent style produces a sense of academic objectivity, but is not always accompanied by nuances such as epistemic stance markers as demanded by the IELTS assessment criteria for ‘Lexical Resource' and ‘Coherence and Cohesion'. Even with this newfound understanding, there is still a great deal of research needed to more precisely quantify the specific, repetitive N-grams the LLM tends to fall back on when completing the 275-word requirement in an IELTS Task 2 essay.

Therefore, this study uses a corpus-driven analysis approach to find and classify the most common formulaic patterns produced by the LLMs when answering IELTS Task 2.As a result, this study uses the corpus-driven analysis approach to search and classify the most frequent formulaic patterns that LLMs generate to answer IELTS task 2. This study aims to answer two main questions: what are the most common 3-word and 4-word lexical bundles in the AI-generated IELTS responses, and how do they work rhetorically to create a synthetic academic voice? This study seeks to offer empirical proof for educators and test takers about the stylistic constraints of AI model-generated texts. Finally, the results will highlight the need for a more critical stance to writing using AI-based tools, encouraging the use of more sophisticated understandings of the role of formulaic language in shaping the perceived quality of a response and the actual improvement of a learner's writing.

Literature Review

Study of multi-word sequences is central to applied linguistics because the knowledge of such structures is crucial for academic fluency. The basic design of the extraction and analysis of these structures, known as "lexical bundles", was developed by finding the statistical frequencies in large corpora (Biber et al., 2004). This basic research showed that the discourse-building blocks that define the organization of information and stance in the academic context are different to those found in other academic registers, such as university teaching and textbooks (Biber et al., 2004). On this note, Hyland (2008) pointed out that the notion of lexical bundles is not uniformly applied within the academy but rather that they are used for different rhetorical purposes in different disciplines. These basic corpus research projects managed to chart the authentic linguistic properties of the human writing process of academic discourse, but they failed to envision the recent and ongoing algorithmically produced text that is now present in the language learning landscape leaving a gap in today's assessment of model texts.

Recently, the use of Large Language Models (LLMs) has led to a reevaluation of the problem of corpus-assisted instruction (CAI) in English Language Teaching (ELT). The ability of models such as ChatGPT to generate large amounts of grammatically correct text poses challenges to traditional studies of discourse as argued by Curry et al. (2024), as it can make it less clear what knowledge is produced by humans and what is produced by the algorithmic pattern recognition. Likewise, Crosthwaite and Baisa (2023) warn of replacing traditional data-driven learning with AI. They mention that while AI systems can give learners real-time, context-appropriate feedback, the generated texts might not capture the communicative subtleties and nuances of real human texts. This requires a deeper understanding of the stylistic features of AI-generated texts, since students are now likely to read more of these when they meet them in their learning environment than real human texts.

Studies over the last few years have started examining this particular lexicogrammatical trace of generative AI. When comparing with the essays written by ChatGPT, Jiang and Hyland (2024) reported that the former has a significantly lower overall frequency of lexical bundles, but a significantly higher type-token ratio for the few lexical bundles that it produces, suggesting that it constructs an essay with a more rigid and formulaic approach than a human writer. In particular, the AI seems to be relying too heavily on abstract, noun-based bundles and on transition phrases, often failing to recognize the subtle epistemic stance markers that human students would have been able to identify to engage their readers in persuasion (Jiang & Hyland, 2024). The results are consistent with

previous studies in scientific writing, where reviewers complained that abstracts produced by ChatGPT lacked original writing voice and were overly mechanical and formulaic in style (Gao et al., 2023). Although the previous study by Jiang and Hyland (2024) and the study by Gao et al. (2023) successfully demonstrate the mechanisms by which AI works, the range of applicability of each study is narrow, namely general A-level argumentation in the former, and medical abstracts in the latter.

This is a very important and clear gap in the field of language testing to measure English proficiency, specifically the IELTS test. As a test requiring very specific lexicogrammar, the IELTS exam has been shown to benefit from corpus-based tools to help candidates understand and produce such academic texts (Qin, 2023). It is particularly important, however, to understand what kind of stylistic biases are being absorbed by the candidates, who more and more are using AI tools to create ‘model' IELTS Task 2 type essays to study. Though we have an understanding of the formulaic nature of AI in other settings, we currently don't have any empirical data that quantifies the exact and specific words and phrases (3-word and 4-word lexical bundles) that LLMs fall back on when using them in the essay prompt for the IELTS test. It is important to be able to recognize this synthetic stylistic sign, in order to avoid the assimilation of an unnatural, robotic academic style, which may lead to loss of points.

Methodology and Theoretical Framework

This study adopts a corpus-driven approach to exploring the formulaic stylistic strategies, or “AI-ese”, that are characteristic of academic writing created by Large Language Models (LLMs). The corpus includes 20 model essays produced using Gemini Flashlight 3.1, which have been created with a systematic approach and cater to common IELTS Academic Writing Task 2 prompts. A strict system prompt was used throughout: “Write an essay for Academic Task 2 answering the following prompt of 275 words. Take an academic, objective, formal voice. Do not have a title or introduction header.” To ensure consistency of data and reduce variance in produced style, a uniform prompt was given. The total number of words is about 3099 and it was compiled into a single .txt file in UTF-8 encoding. High fidelity data extraction from this uniform processing is obtained via AntConc (Version 4.4.0). The analytical framework focuses on the identification of recurring lexical bundles (lexical sequences of 3-4 words, which have a high statistical frequency, Biber et al., 2004). The analysis sets a minimum frequency of 5, which eliminates idiosyncratic noise in the data and emphasizes the systematic and repeated "rhetorical fingerprint" of the model. This quantitative analysis is complemented by a qualitative analysis of the rhetorical functions of the bundles, classified based on how they are involved in stance and discourse organization (Hyland, 2008).

The paradigm of this research is Corpus-Assisted Discourse Studies (CADS) and Formulaic Language Theory. The number and distribution of lexical bundles, the "building blocks" of specialized discourse as Biber and Conrad (2009) argue, are largely responsible for the register variation of a text type, such as the specific set of requirements in the IELTS Academic Task 2. In the context of IELTS, where every second counts, the candidate's ability to use natural and varied lexical resources is assessed. But the structure of LLMs, learned through probabilistic text generation, is inherently biased towards sequences that are very likely to occur (Curry et al., 2024). This architectural bias leads to the creation of formulaic, predictable and standardized language, which is different from the idiosyncratic variance seen in the high-scoring human essays (Jiang & Hyland, 2024).

In the present study, the functional classification of lexical bundles developed by Hyland (2008) is used to classify sequences into three main rhetorical categories: stance markers (e.g., it is undeniable that), discourse organizers (e.g., in conclusion it is) and referential expressions. This study will seek to illustrate this "synthetic fingerprint" by observing the corpus and show that AI-generated essays are characterized by a disproportionate use of stance markers, which creates the illusion of academic objectivity but does not actually result in a true diversity of lexical items. Although grammatically correct, the use of a limited range of N-grams is a stylistic hallmark of robotic fluency and does not necessarily correspond to the presence of the author that is required by the assessment criteria for IELTS (Jiang & Hyland, 2024). Based on this, this study assumes that there is a measurable stylistic overuse of these particular bundles of transitional and stance-taking devices that may be problematic for the learners who use AI as a main source of writing learning.

Findings and Discussions

The quantitative analysis of the 20-essay corpus shows that the texts have a very distinctive language profile, marked by extreme stylistic homogeneity and predictable formulaic structures. The corpus is very small, with 5,500 tokens, which results in a very low ratio of types to tokens, meaning that the model will have a very small lexicon with high probability in order to handle a variety of topics in the argumentative text. The most common content words found in Word results file are listed in table 1. These frequencies reflect a thematic interest in the dilemmas of society, and the major lexical units of the AI's argumentative strategy are the words "argue," "pivotal" and "economic.

Table 1: Top Content Words

Rank	Word	Frequency
14	argue	24
15	essay	20
21	modern	15
33	pivotal	12
35	economic	11
40	global	10
42	integral	9

The 3-word and 4-word lexical bundles in N-Gram results are in accord with the occurrence of a 'rigid' 'AI-ese' fingerprint. The model frequently uses anticipatory "it" sentences and frequently used stance markers as shown in Table 2 to set the academic tone in the model.As indicated in Table 2, the model always uses anticipatory "it" sentences and also the use of frequently used stance markers to set the academic tone in the model. It is a common expression and its variants are used frequently and act as main discourse organisers.

Table 2: Top Identified Lexical Bundles

N-gram	Frequency
is widely acknowledged	14

it is widely	14
widely acknowledged that	14
in my opinion	13
in conclusion the	9
others argue that	9
the modern era	9

Moreover, collocational analysis gives information on the structural dependence of the AI. The data shows that the model can consistently match rigid noun phrases with certain adjectives, is a pivotal, and the modern era. This indicates the AI is not attempting to create these phrases from a large vocabulary set, but is rather accessing them as packaged phrases. Likewise, there is a strong statistical connection between widely and it is; the terms free and university are strongly connected as fixed parts of the approach to the topic of education as a model. Lastly, there is the argumentative structure that is consistently composed by the binary opposition of some and others which is used as a recurring formula for presenting divergent ideas.

Discussion

The results of this corpus-based analysis can be considered empirical evidence of a synthetic academic voice that is extremely uniform and has been algorithmically produced. It is well known and I think it is the very reason why the Large Language Model (LLM) is capable of composing argumentative texts when there are many bundles that can be found in the N-Gram results that are marked as 'in common'. The repetitive use of these sequences creates a monotonous style with little variation in the lexicon or in the way the ideas are presented which does not match the idiosyncratic variation in the lexicon and the flexibility in presentation of ideas which are expected of higher scoring human IELTS candidates. This predictability is not just a stylistic trait of the model, but a

certain constraint of the model to produce a real tone, a vital criterion of the rubrics for IELTS Writing Task 2, band 8.

Moreover, the information given by collocate results show that there is a preference for the top-probability sequences as opposed to original, context-dependent thinking. The uniform pairing of the 'modern' and the 'era', and the "excessive use of is" are key to mark a range of topics which are diverse and complex socio-economic, suggesting that the model favours structural probability, how likely it is for one word to follow another, over substantive lexical diversity and semantic depth. The dependence on limited range of N-grams can be seen as an "AI-ese" fingerprint, similar to the phenomenon that was noted by Jiang and Hyland (2024) in relation to the limited range of "N-grams" and "AI-ese" that ChatGPT uses when composing an argumentative text. If a model always assumes these "pre-packaged" units of discourse, it reduces the writing to a lack of nuance, complexity and originality of ideas that are characteristic of expert human performance.

Pedagogically the findings are far reaching and problematic. The markers of the IELTS examiners are trained to recognise memorised responses, formulaic language and "over-used" words, since these are indicative of the lack of linguistic control and adaptability. Learners who use AI responses as their models in writing, therefore, might end up internalizing a robotic, superficial writing style that is not suitable for the IELTS criteria for "Lexical Resource" and "Coherence and Cohesion". Students may fall into the trap of copying the style of these AI essays, which could lead to an artificial impression of writing by the examiner and impact their band marks.

In conclusion, the results highlight the critical need for more critical use of AI to support learning. The passive approach to learning the models of AI as warned by Crosthwaite and Baisa (2023), promotes the use of the model rather than developing communicative competence. Next steps include an active, formula-deconstructive approach to "AI-ese" using a pedagogical lens informed by the corpus, whereby learners are motivated to recognise and critique the highly predictable formula of the machine. Further studies are needed to identify the “threshold of predictability” when a text becomes algorithmically uniform and is negatively evaluated by human evaluators and by the standardized evaluation measures. While LLMs are being deployed as "models" for high-stakes academic writing, these deconstructive strategies will not be put in place until then.

Conclusion

The results of this corpus-driven study confirm that Large Language Models (LLMs) produce IELTS Task 2 responses with an algorithmic fingerprint which is clearly present throughout the responses. From the systematic analysis of recurrent lexical units found in the file N-Gram results, it is clear that the model is based on a limited, probabilistic structure that favors predictable structures over lexical and rhetorical diversity. The empirical data shows that these essays are grammatically correct, but their use of 'pre-packaged' language patterns or repetitive language patterns leads to an overtly robotic register. Therefore, the study suggests that AI-ese produced by these models is quite lacking in terms of epistemic stance markers and authorial stance, which are needed for high-level academic achievement. The outputs produced by the model are never the "gold standard" to which learners and educators aspire when preparing for the IELTS; this is because the model is systematically unable to produce new, context-specific arguments, which means that the voice one gets from it is performative but still rather shallow, with the risk of being rejected by standardized assessment criteria. The bundles are made by repeating the same word, phrase, and/or sentence multiple times, and these same words, phrases, and sentences are frequently used as discourse markers, such as in my opinion and is widely acknowledged. This repetitive use of the same word, phrase and/or sentence, coupled with the absence of genuine communicative intent, suggests an algorithmic convenience over the dynamic intellectual work of argumentation. The increasing differences between the language AI models produce and the rhetorical complexity found in academic writing are a danger to the diversity of academic writing, which could become more homogenized if these models were relied upon. In conclusion, this study suggests that the use of LLMs in IELTS preparation without a critical lens will still continue to cause problems, not solutions, for achieving high levels of language proficiency for academic and professional success, unless educators and learners become aware of the constraints of "AI-ese."

References

1. 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
2. Crosthwaite, P., & Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics*, *3*(3), 100066. https://doi.org/10.1016/j.acorp.2023.100066
3. Curry, N., Baker, P., & Brookes, G. (2024). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. *Applied Corpus Linguistics*, *4*(1), 100082. https://doi.org/10.1016/j.acorp.2024.100082
4. Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ Digital Medicine*, *6*(1), 75. https://doi.org/10.1038/s41746-023-00819-6
5. Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, *27*(1), 4–21. https://doi.org/10.1016/j.esp.2007.06.001
6. Jiang, F., & Hyland, K. (2024). Does ChatGPT argue like students? Bundles in argumentative essays. *Applied Linguistics*, *46*(3), 375–391. https://doi.org/10.1093/applin/amae052
7. Qin, L. X. (2023). An Analysis of IELTS Academic Reading Texts through A Corpus-based Approach. *Journal of Studies in the English Language*, *18*(1), 1–38. https://doi.org/10.14456/jsel.2023.1