المجلة العراقية للبحوث الإنسانية والإجتماعية والعلمية
Iraqi Journal of Humanitarian, Social and Scientific Research
Print ISSN 2710-0952 Electronic ISSN 2790-1254
العـــــدد 19A تشر2ين 2025
No.19A Nov. 2025

Comparison of Bundle Type/Token Ratios between ChatGPT-Generated Essays and Student Essays Across Proficiency Levels

Assistant lecturer: Ali Abd Kadhum Al Mashhady
General Directorate of Education in Thi Qar
Email: aliabdkadhum@gmail.com
ORCID: https://orcid.org/0009-0005-1487-8441

Abstract

A corpus-comparative design was used to compare lexical bundle diversity in argumentative essays produced by ChatGPT and in essays written by Iraqi EFL learners at two proficiency levels. Three hundred eighty-four essays were analyzed, comprising multiple ChatGPT generations and learner texts balanced by proficiency. Texts were processed in raw and normalized streams and contiguous three-word and four-word bundles were extracted per essay. Per-essay type/token ratio served as the main outcome. Complementary diversity indices were computed, including moving-average TTR and Guiraud's R. Writing quality was rated with a five-dimension analytic rubric by two blinded raters and inter-rater agreement was assessed. Linear mixed-effects models with prompt as a random intercept and control for essay length were fitted to test group contrasts. Results showed that intermediate-level learners produced higher bundle diversity than pre-intermediate learners. ChatGPT output concentrated tokens on a limited set of canonical discourse and stance linkers and thus occupied an intermediate position between the two human bands on the type/token index. A pooled comparison of all student essays versus ChatGPT did not yield a reliable difference because group composition and within-group variance reduced observable contrasts. Normalization raised some learner bundle counts and altered diversity metrics. Placement scores correlated positively with bundle diversity. Findings imply targeted instruction in bundle variety and multi-index assessment. Limitations include single-genre sampling, constrained national context, and dependence on extraction choices. Future work should sample multiple model versions, expand genres, and preregister analysis pipelines.

Keywords: Lexical bundles, ChatGPT, L2 writing, type/token ratio, corpus linguistics

مقارنة نسب و نوع الحزمة/الرمز بين المقالات المُولّدة بواسطة ChatGPT ومقالات الطلاب وفقا لمستويات الكفاءة

م.م.علي عبد كاظم المشهدي
المديرية العامة لتربية ذي قار

الملخّص

المجلة العراقية للبحوث الإنسانية والإجتماعية والعلمية
Iraqi Journal of Humanitarian, Social and Scientific Research
Print ISSN 2710-0952 Electronic ISSN 2790-1254
العـــــدد 19A تشر2ين 2025
No.19A Nov. 2025

استخدمت هذه الدراسة تصميمًا مقارنًا قائمًا على المدونات اللغوية لتحليل تنوّع الحزم المعجمية في المقالات الجدلية التي ينتجها نموذج (ChatGPT) مقارنةً بالمقالات التي كتبها متعلمو اللغة الإنكليزية كلغة أجنبية من العراقيين عند مستويين من الكفاءة اللغوية. شملت العينة 384 مقالة تضمنت عدة نسخ مولَّدة بواسطة (ChatGPT)ونصوص المتعلمين موزَّعة بشكل متوازن حسب مستوى الكفاءة. تمت معالجة النصوص بصيغتها الخام والمطَبَّعة، واستُخرجت الحزم المتجاورة المكونة من ثلاث وأربع كلمات لكل مقالة. واستخدم مؤشر النوع/الرمز لكل مقالة كمقياس رئيس للتنوّع، إلى جانب مؤشرات إضافية مثل معدل المتوسط المتحرك لمعامل النوع/الرمز (MATTR) ومعامل Giraud's R. تم تقييم جودة الكتابة باستخدام مقياس تحليلي خماسي الأبعاد من قبل مقيمين مستقلين، مع قياس درجة الاتفاق بينهما. كما استخدمت نماذج التأثيرات المختلطة الخطية باعتبار المحفز (prompt) متغيرًا عشوائيًا والتحكم في طول المقالة لاختبار الفروق بين المجموعات .أظهرت النتائج أن المتعلمين في المستوى المتوسط أنتجوا تنوعًا أكبر في الحزم المعجمية مقارنة بالمستوى ما قبل المتوسط. أما مخرجات (ChatGPT) فقد ركّزت الرموز على مجموعة محدودة من روابط الخطاب والمواقف النمطية، مما وضعها في موضع وسطي بين المستويين البشريين وفق مؤشر النوع/الرمز. ولم تُظهر المقارنة المجمعة بين جميع مقالات الطلبة ومخرجات (ChatGPT) فروقًا ذات دلالة إحصائية، نتيجة تباين تكوين المجموعات والتفاوت داخل المجموعات .أدى التطبيع (Normalization) إلى زيادة عدد الحزم لدى بعض المتعلمين وتغيير مقاييس التنوّع. كما ارتبطت درجات تحديد المستوى إيجابيًا بتنوّع الحزم المعجمية. وتشير النتائج إلى أهمية التعليم الموجَّه نحو تنويع الحزم اللغوية واعتماد تقييم متعدد المؤشرات .من القيود الرئيسية للدراسة: اقتصارها على نوع واحد من النصوص، ارتباطها بسياق وطني محدود، واعتمادها على خيارات استخراج محددة. وتوصي الدراسة المستقبلية باختبار نسخ متعددة من النماذج اللغوية، توسيع أنواع النصوص، وتسجيل إجراءات التحليل مسبقًا لضمان الشفافية والمنهجية.

الكلمات المفتاحية: الحزم المعجمية، ChatGPT، الكتابة بلغة ثانية، نسبة النوع إلى الرمز، لغويات النصوص

Introduction

In recent years, large language models such as ChatGPT have reshaped the production and evaluation of written language. Comparisons between human and machine-generated texts are now conducted in various domains. The entrance of artificial intelligence into educational settings has occurred rapidly. Its influence on writing practices in English as a Foreign Language contexts is still being examined. One area of increased focus is the comparison of linguistic features in essays written by students and those generated by ChatGPT. Argumentative writing holds particular importance in EFL instruction. This genre reflects student language proficiency. It also shows how learners organize ideas and present claims. Cohesion and stance are established through patterned language in these essays. The structure of argumentative texts allows analysis of how linguistic resources are utilized by writers. Within this genre, lexical bundles, multiword sequences that recur frequently and function as building blocks of written discourse, offer an important lens for studying differences between human and AI-generated writing. These bundles often reveal underlying proficiency, discourse control, and linguistic maturity, making them a meaningful point of comparison.

Studies indicated that proficient writers use lexical bundles to provide arguments and make cohesive texts. Evidence suggests that advanced EFL writers use a wider range of these multi-word sequences. A reduced dependence on repetitive structures is observed in their writing compared to lower-proficiency learners (Hyland, 2008). Examinations of native and non-native writing reveal that non-native learners frequently overuse certain structural patterns. Concurrently, an underuse of bundles that express stance or organize discourse is documented in their work (Chen, 2016).

More recent studies examine lexical bundle usage in ChatGPT-generated texts. AI-generated texts are noted for their fluency, but they often display predictable bundle patterns. A limited variation and a lack of authorial presence are observed in these texts (Jiang & Hyland, 2024). The diversity of bundles, measured through type-token ratios, provides insights into linguistic differences between student and AI writing.

In EFL settings, student proficiency level affects lexical bundle use. Intermediate learners frequently employ memorized structures and repetitive phrases. Greater variation and higher type-token ratios are demonstrated by advanced learners. These multi-word sequences are used for stance, hedging, and cohesion with greater naturalness by advanced students (Biber et al., 2004). Existing research has primarily examined learners in East Asian or European contexts. Very limited work focuses on Middle Eastern settings.

Iraqi EFL learners have received minimal attention in corpus-based writing research. A research gap remains despite increased focus on English instruction in Iraqi higher education. Challenges with grammatical precision and argument organization are noted among numerous students (Abdalhussein, 2022). Investigations at Iraqi institutions document excessive use of predetermined expressions. Sophisticated connecting words are frequently omitted, and language expressing viewpoint is employed insufficiently (Abdalhussein, 2022). These characteristics make Iraqi learners suitable for comparison with machine-generated writing.

ChatGPT is now used in Iraqi classrooms, often as an unofficial writing aid. Concerns about academic integrity are raised by instructors. Some suggest that AI-generated essays could serve as language models if analyzed critically. However, little empirical evidence exists on whether ChatGPT's language resembles that of Iraqi EFL learners. It is unclear if the differences are significant enough to cause pedagogical concerns.

Lexical bundles show how writers structure meaning beyond sentence-level grammar. A comparison of bundle type-token ratios across student proficiency

levels and AI output could offer measurable insights. Type-token ratio distinguishes repetitive writing from varied language use. A low ratio suggests overuse of a narrow range of bundles. A high ratio indicates greater variety and linguistic sophistication. Earlier findings show that ChatGPT reuses certain stance and transition bundles in argumentative texts (Jiang & Hyland, 2024). No study has directly compared its bundle ratio with that of Iraqi students at different proficiency levels.

The lack of research in Iraqi contexts is notable given debates about AI's influence on student writing. Unusual fluency and generic structure are noted by some instructors in student submissions. The typical errors of intermediate proficiency are often absent from these texts. Determination of whether ChatGPT mimics, exceeds, or systematically differs from student writing requires linguistic comparison with concrete metrics. Lexical bundle analysis, particularly through type-token ratios, provides an objective method. Distinction between naturally acquired patterns and AI-produced combinations is enabled by this approach.

The significance of this comparison extends beyond fluency. Bundle use in argumentative writing reflects how writers engage with ideas. Different patterns of idea development and argumentation are revealed through bundle analysis.

Research in various EFL settings indicates that advanced learners more frequently employ stance bundles like "it is clear that." In contrast, lower-level learners tend to depend on referential bundles such as "the role of" (Ädel & Erman, 2019). Should ChatGPT primarily use generic, stance-neutral bundles, its writing might resemble that of intermediate learners more than advanced writers. Alternatively, a different pattern would be indicated if the AI shows a broad distribution of bundles with limited variation within categories.

No published study currently compares ChatGPT's lexical bundle use with Iraqi student writing. Analysis of type-token ratios across different proficiency levels in this specific context is also missing from existing research. These gaps demonstrate a need for comparative investigation. A systematic examination of bundle patterns between AI-generated texts and Iraqi EFL student writing would address this need.A study that investigates lexical bundle type-token ratios in ChatGPT-generated argumentative essays and those written by Iraqi EFL learners at different proficiency levels would address this need. This kind of analysis can reveal whether AI-generated writing resembles intermediate or advanced Iraqi students, or whether its bundle diversity is distinct from both. This analysis could help educators determine if ChatGPT's output might be confused with student writing during assessment. The linguistic features of AI-generated texts might also function as instructional tools for contrastive analysis.

Furthermore, bundle pattern comparisons across proficiency levels may enhance understanding of discourse competence development among Iraqi learners. Structural and repetitive differences between AI-produced texts and human writing could be clarified through such comparison.

The present study addresses these questions through examination of bundle type-token ratios in three sets of argumentative essays. Texts generated by ChatGPT are compared with those written by intermediate and advanced Iraqi EFL learners. This examination of a measurable linguistic variable supports progress in EFL writing research. It also addresses the expanding domain of AI-generated language analysis. The present study compared lexical bundle type-token ratios in argumentative essays produced by ChatGPT and Iraqi EFL learners at intermediate and advanced proficiency levels. This study sought to answer the following research questions:

1. Is there a significant difference in lexical bundle type/token ratios between ChatGPT-generated essays and Iraqi EFL student essays?
2. Does proficiency level influence the bundle type/token ratios in Iraqi EFL argumentative writing?

Literature Review

Argumentative writing occupies a central place in academic language instruction because it requires control over content, structure, and stance. Writers must present claims, support them with reasons, and link sentences and paragraphs so that meaning is clear. Multiword sequences that recur in a register, often labeled lexical bundles, play a key role in achieving these tasks. Lexical bundles function as building blocks of discourse and as devices for framing argument, signaling stance, and organizing information. The study of bundles therefore offers a direct way to examine how writers construct argumentative moves and how text cohesion is achieved (Biber, et al., 2004).
The term lexical bundle refers to recurrent sequences of words that occur with high frequency in a particular register. These sequences may be three words long or longer, and they carry limited propositional content on their own. Scholars suggest that bundles exist between individual words and full clauses. This type of analysis uncovers organizational structures that single-word analysis often overlooks. These patterns include preferred methods for introducing evidence, hedging claims, and organizing discourse (Hyland, 2008).

Early corpus work established the empirical method that remains standard for bundle research. Candidate sequences are identified by researchers through frequency thresholds. Their distribution and functional characteristics are then examined across different subcorpora. This methodological approach allows for

systematic comparison of bundle usage patterns between distinct writer groups (Biber et al., 1999). This method distinguished bundles used in teaching and textbooks from those in research articles and lecture speech, showing that register and discipline shape the bundle repertoire writers use (Biber et al., 2004). Subsequent investigations compared published disciplinary writing with student compositions. Systematic differences were observed between these groups. Student writing contained more noun-based and referential bundles, such as "the role of" or "in the case of." Experienced writers employed more stance and discourse-organizing bundles like "it is important to" or "as a result of." These patterns show how beginning writers often focus on naming concepts, while advanced writers more frequently evaluate ideas and structure discussions for readers. This progression reflects important development in academic writing ability. These usage patterns indicate authorial positioning and help navigate readers through complex arguments (Cortés, 2004).

Further research on learner writing connected bundle use to language development. Analyses of rated learner corpora indicated that higher proficiency correlates with greater bundle diversity. More frequent use of stance bundles and hedging structures was documented among advanced learners. Lower-proficiency learners depended on recurrent, often formulaic phrases. These phrases were used for basic referencing and surface-level connections. They provided limited support for critical argument analysis or for developing their own voice (Chen & Baker, 2016). This pattern suggested that the capacity to use a wider set of bundles reflects both linguistic resources and rhetorical sophistication.

Type/token ratio (TTR) for bundles has emerged as a useful index of bundle diversity. Token frequency counts the total occurrences of bundle tokens in a text. Type frequency counts the number of distinct bundle forms that appear. A high token count with a low number of types signals repetitive use of the same bundles. A higher TTR indicates a more varied bundle repertoire. Researchers observed that TTR enhances simple frequency counts by distinguishing between texts that merely contain many bundles and those that demonstrate true variety in their usage (Biber et al., 2004).

The arrival of large language models in writing contexts prompted new questions about bundle patterns. While generative systems produce fluent and well-organized text, systematic differences from human writing may exist. Studies of AI-generated and human essays found that language models frequently create polished but predictable prose. A recent study analyzed lexical bundles in argumentative essays from both ChatGPT and students. Jiang and Hyland (2024) found differences in

both frequency and type/token behavior, which suggested that AI texts may show less authorial stance and more constrained patterns in some registers.

A larger comparison across multiple datasets found that AI-generated texts sometimes received higher scores on general quality metrics. Although the essays generated by AI often appeared fluent, they lacked the kind of variation and individuality typically found in human writing (Herbold et al., 2023). In other words, even when the language was smooth, the texts did not display the subtle rhetorical choices that human writers usually make when constructing an argument. Later studies, including those by Levine (2025) and Mizumoto (2024), examined how AI handles interactional features such as engagement markers. Their findings showed that large language models tend to avoid expressions that directly address the reader or signal a personal stance, resulting in writing that feels polished but impersonal. These linguistic elements are recognized as essential components of a distinctive authorial voice.

The way AI and human writers use common phrases matters significantly for classroom practice and evaluation. These distinctions in expression create valuable teaching opportunities. When AI-generated essays follow highly predictable phrasing patterns, they can become useful teaching tools. Instructors may use them to show students how machine-produced text differs from human writing in terms of expression, rhythm, or rhetorical presence. Research does not fully agree on how consistent these patterns are. Some studies report that large language models occasionally mimic advanced student writing so closely that even trained evaluators struggle to tell them apart (Jiang & Hyland, 2024; Lee, 2023). Other analyses suggest that AI output often relies on a narrower and more repetitive range of expressions, which can make detection easier in certain cases (Mizumoto, 2024; Herbold et al., 2023).

These mixed findings suggest that grading practices may need to evolve to account for recognizable features of AI-generated writing. Rather than relying solely on surface fluency, instructors may need clearer criteria that include indicators of rhetorical individuality, stance, or bundle diversity. Establishing clearer guidelines for spotting automated text could ensure fair assessment while promoting responsible AI use in education. Recent investigations highlight these contradictions: while ChatGPT sometimes produces more varied phrasing than students, it often misses the subtle language that shows critical thinking and personal perspective (Jiang & Hyland, 2024; Zhang & Crosthwaite, 2025).

EFL writing studies provide a crucial perspective because lexical bundle development reflects both language proficiency and pedagogical exposure. Learners in different EFL contexts show different bundle repertoires depending on

instruction, exposure to academic input, and assessment demands. Classroom tasks and teacher practices influence the input learners receive and the depth of their engagement with language. Empirical work has shown that tasks with greater involvement produce stronger incidental vocabulary gains, which in turn increase the variety of lexical items learners can draw on in writing (Javanbakht, 2011). In addition, constraints in teacher training and internship practice may limit opportunities for learners to encounter and use discourse-organizing language in authentic contexts, which can encourage reliance on a narrow set of formulaic expressions (Javanbakht, 2014). These teaching and learning conditions help explain why some English language learners, particularly in certain learning environments, depend heavily on referential phrases rather than developing their own voice through stance and organizational expressions.

Middle Eastern EFL contexts require specific attention. Research on Iraqi university students documented difficulties in cohesion, use of conjunctions, and management of argumentative structure. An analysis of conjunction errors and cohesive device use reported that many students struggled with linking ideas logically, which reduced the clarity of argumentation in essays (Darweesh & Kadhim, 2016).
Another investigation of formulaic language in Iraqi writing showed that high-scoring students used interactive and interactional discourse markers more frequently than low-scoring peers, suggesting that bundle use aligned with measured writing proficiency in that context (Abdalhussein, 2022).

Corpus studies that target EFL learners in non-Western contexts reinforced the relation between proficiency and bundle diversity. Research in East Asian and European contexts found that bundle profiles shifted across proficiency bands, with more advanced learners using bundles that signal authorial stance, hedging, and argument organization (Chen & Baker, 2016). These findings probably apply to Iraqi learners, though local curriculum and exposure to academic discourse may alter the pathways of development.
Methodological choices shape results in bundle research. Researchers must decide on bundle length, minimum frequency, and normalization procedures. Three-word bundles are most frequently examined in research. Their balance between frequency and functional clarity is noted by scholars. Four-word bundles offer more detailed functional distinctions in certain academic registers (Hyland, 2008). The recognition of meaningful lexical bundles relies on two factors. Appropriate frequency thresholds must be established. Sufficient corpus size must be secured. These methodological considerations determine which word sequences qualify as types. They also influence how often these sequences appear as tokens during analysis (Biber et al., 1999). Studies of student writing typically select lower

thresholds because student corpora are smaller; such selection affects TTR computation and comparability across corpora. Indeed, consistent and clear methods are essential. Researchers establish definite rules for selecting meaningful word combinations and standardize how often these phrases occur, usually by counting them per 10,000 words. This practice allows for fair and accurate comparisons between different pieces of writing (Biber et al., 2004).

Recent investigations examined how prompt design influences AI lexical patterns. Experimental studies found that instructions requesting a personal voice or evaluative statements significantly altered bundle patterns in ChatGPT's output. This indicates that AI-generated phrasing depends substantially on prompt design and generation parameters rather than representing a fixed writing style (Mo, 2025). The potential of linguistic features to identify AI generated writing was examined in multiple studies. Analytical models that used bundle frequency, type token ratios, and other stylistic characteristics demonstrated reasonable accuracy under specific conditions. Consistent performance across various topics and prompt types remained difficult to achieve (Mizumoto, 2024). These results show that bundle measures can form part of a detection toolbox but cannot by themselves provide infallible evidence.

Pedagogical studies examined how awareness of bundles affected student writing. Instruction that combined exposure to authentic bundles with practice in using them for stance and coherence led to measurable improvement in learners' essay organization and in the diversity of bundles used (Cortés, 2006). This finding indicates that bundles are teachable resources and that instruction can shift learner repertoires in ways that align with advanced academic use.

Gaps in the literature remain. Little work has examined Iraqi learners' bundle profiles in direct comparison with AI output. Most AI-human comparisons use Western student corpora or global English samples; applications to Middle Eastern EFL contexts are rare. The relation of bundle TTR to assessment outcomes in Iraq received limited attention. Research that combined corpus analysis with human rating of argumentative strength could clarify whether higher TTR corresponds to higher perceived argument quality in an Iraqi sample. Furthermore, cross-proficiency comparisons using the same prompts for ChatGPT and for learners would clarify whether AI output aligns more closely with intermediate or advanced learner writing in that context. The influence of prompt design and model settings on ChatGPT bundles also requires more systematic study.

The literature reviewed identifies several points that motivate the present study. Lexical bundles serve as indicators of rhetorical and discourse competence in various registers (Biber et al., 2004; Hyland, 2008). Writing proficiency and

organizational ability are reflected through bundle diversity and type token ratios (Chen & Baker, 2016). Measurable differences in bundle patterns are found between texts from language models like ChatGPT and human writing. These differences change when prompt design and dataset characteristics vary across studies (Jiang & Hyland, 2024; Herbold et al., 2023). Studies of Iraqi English learners show difficulties with creating cohesion in writing. Heavy use of patterned language is common among these students. A connection between formulaic language and proficiency levels is established for this group (Darweesh & Kadhim, 2016; Abdalhussein, 2022). A systematic comparison of bundle type token ratios is needed between ChatGPT outputs and Iraqi learners at intermediate and advanced levels. This comparison would help understand formulaic language use across human and machine writers. This approach would help clarify how AI-generated writing compares to human writing development in this understudied context. The study will therefore compare ChatGPT essays with intermediate and advanced Iraqi student essays using a uniform prompt set and transparent bundle extraction criteria. The comparison will test whether AI bundle diversity approximates the repertoire of intermediate learners, advanced learners, or neither, and whether bundle TTR correlates with human ratings of argumentative quality.

Methodology

A quantitative corpus-comparative design was used to answer the research questions. The study compared lexical bundle diversity in argumentative essays produced by ChatGPT and by Iraqi EFL learners at two proficiency levels. Corpus linguistics methods were combined with inferential statistics to measure type/token ratios and to test group differences. This design allowed linguistic patterns to be observed at text level and then linked to proficiency and writer type.

Participants and corpora

The human corpus consisted of essays written by university students enrolled in English programs in Iraq. Two proficiency bands were defined: pre-intermediate and intermediate. Placement into bands relied on institutional placement scores or recent standardized test results. Twenty essays were collected for the pre-intermediate band and 20 essays for the intermediate band. Each student produced one argumentative essay. The target word range for each essay was set between 250 and 350 words to match common classroom assignments and to limit length effects on bundle measures.

Demographic information was collected from each participant. Proficiency scores were also recorded. These data were stored separately from the essay texts to maintain anonymity. The AI corpus contained essays generated by ChatGPT. The same writing prompts were used for the AI as for the student participants. The same

word range instructions were provided to the AI system. For each prompt, two essays were generated and logged, then the generation that matched the word-length criteria best was selected. The total number of ChatGPT essays matched the number of student essays. The model version and generation parameters were recorded and reported in the study so that the process remained transparent and reproducible.

Materials and Prompts

A set of neutral argumentative prompts was developed to elicit comparable essays from students and from ChatGPT. Prompts were phrased in clear academic English and requested a reasoned position with supporting examples. The same wording was used for student tasks and for model instructions. A standardized administration script was read aloud to all students before writing began. This script explained the task requirements and time allocation. The same procedure was followed for all writing sessions to maintain consistent conditions across groups. For ChatGPT, the prompt text and generation parameters were logged in full, including model version, temperature setting, and any system messages used.

The essays were scored using a five-dimensional analytic rubric (Task Fulfilment & Development; Organization & Cohesion; Lexical Resource; Grammar & Accuracy; Mechanics & Presentation). Each dimension was rated on a 0–6 scale; dimension scores were summed (0–30) and converted to a 0–100 composite writing score. Institutional placement scores (or, where necessary, a composite constructed from the study's grammar and listening subtests plus the rubriced writing score) were used to assign students to proficiency bands (pre-intermediate = 30–54; intermediate = ≥55). Two independent, blinded raters scored every essay after a calibration workshop. Numeric agreement between raters was monitored using Pearson correlation (target $r \geq .70$); categorical agreement for bundle function labels was assessed with Cohen's κ (target $\kappa \geq .60$). Where the two raters' raw totals differed by more than 6 points (raw), a third senior rater adjudicated and the adjudicator's decision was used as the final score.

Data Collection Procedures

Student essays were collected in supervised classroom sessions to avoid external assistance. A fixed time limit was applied to each session to approximate exam conditions. Students provided informed consent and were briefed about the research purpose without revealing the research hypotheses. ChatGPT essays were generated on the same day as student writing sessions when possible. All texts were anonymized and assigned identification codes. Original files and metadata were saved in a secure, access-controlled research folder. Below are three short excerpts

drawn from the corpus to show real context for the bundles and how tokens accumulate.

ChatGPT

In recent years, the purpose of has become a subject of debate in higher education. The purpose of is discussed in relation to curriculum design. It should be noted that many instructors see benefits. On the other hand, some argue for more assessment. In this paper we present evidence and argue for clear guidelines. As a result of these changes, policy must adapt. This point is discussed with reference to current literature.

The excerpt includes several exact bundle occurrences (*the purpose of, it should be noted, on the other hand, in this paper we, as a result of*). Repetition of *the purpose of* increases ChatGPT's token count for that type.

Intermediate

This study examined the role of and discusses implications for practice. In terms of assessment, it is important to provide examples. On the other hand, students need structured feedback. For the purpose of fairness, scoring rubrics must be clear. The findings suggest that revision improves outcomes. Examples and reasons are included to support the claim.

The excerpt uses several different bundles (*the role of, in terms of, it is important to, on the other hand, for the purpose of, the findings suggest that*). The intermediate writer uses a broader range of bundles rather than repeating the same linker.

Pre-intermediate

Most of the participants said that online class is easier. The role of teachers was small in online class. Due to fact of problems, students missed lessons. I think this is important. The student gives a simple example.

Texts were converted to plain text and normalized for tokenization. A token is any single occurrence of a lexical bundle in the corpus. A type is a distinct bundle form. If the sequence "on the other hand" appears 40 times in the corpus, that counts as 40 tokens and 1 type. A high token count with few types means writers repeat the same bundles often. A higher number of types with comparable token counts means writers use a more varied set of bundles.

Lowercase conversion was applied to reduce surface variation. Punctuation was retained for sentence segmentation because some bundles span clause boundaries. Spelling errors and non-standard forms were left unchanged in student' texts to keep authenticity of learner' features. Tokenization was completed with a consistent

procedure that was applied to all texts in the study. The same tokenization settings were applied to ChatGPT outputs to ensure comparability. A record of preprocessing steps and software versions was maintained.

Lexical bundles were operationalized as contiguous three-word sequences. Four-word bundles were extracted as a secondary check for robustness. Extraction focused on continuous n-grams; skip-grams were excluded. Per-text extraction was preferred so that diversity measures reflected individual writing units rather than aggregated corpora. For each essay, the total number of bundle tokens and the number of distinct bundle types were counted. A type/token ratio was computed for each essay by dividing the number of types by the number of tokens. This per-text ratio served as the primary dependent variable because it captured diversity at the level of the writing sample. For descriptive reporting, normalized frequencies per 1,000 words were also calculated.
Bundle extraction was implemented with a reproducible script. The script used a standard tokenizer and counted contiguous n-grams. The code and parameter settings were archived and made available with the research materials.

The primary dependent measure was the per-essay bundle type/token ratio. The first research question investigated the differences in bundle diversity between ChatGPT essays and student essays. The second research question investigated how proficiency level affected bundle diversity in students' writing. Descriptive statistics were calculated for each group. Means, standard deviations, medians, and ranges were reported for type token ratios.

A linear mixed-effects model was fit as a robustness check. The statistical model specified bundle type token ratio as the outcome variable. Writer type and proficiency level were included as fixed effects. Prompts were incorporated as a random intercept to describe prompt level variation. Model diagnostics were examined comprehensively. Residual patterns were verified to ensure that the model assumptions were met. Model estimates were reported with corresponding standard errors and confidence intervals.
Data analysis followed a structured procedure. First, model parameters were established. Then, assumption verification was conducted. Finally, results were documented with appropriate statistical measures.

Secondary analyses examined normalized bundle frequency and functional category distributions. These analyses supported interpretation of any TTR differences and clarified whether differences arose from repetitive use of a few bundles or from broader variety across bundle types.

Ethical approval was obtained from the relevant institutional review boards prior to data collection. Participants provided written informed consent and were informed about data handling and confidentiality.

Results

Table 1 presents the frequency distribution of the 30 most common lexical bundles identified across the three writing groups. The counts show how often each bundle appeared in the essays written by ChatGPT, intermediate students, and pre-intermediate students. Each bundle is also classified according to its communicative function as stance, discourse-organizing, or referential.

Table 1

Top 30 lexical bundles by token frequency and functional class across groups

	Bundle	Functional class	ChatGPT tokens	Intermediate tokens	Pre-intermediate tokens	Total tokens
1	the purpose of	discourse	11	9	3	23
2	it is essential to	stance	7	6	1	14
3	there is no	stance	10	0	0	10
4	it is clear that	stance	6	2	2	10
5	in this paper we	discourse	7	0	2	9
6	the evidence shows	stance	7	0	2	9
7	it is possible that	stance	6	1	2	9
8	it should be noted	stance	5	1	3	9
9	from the perspective of	referential	3	1	6	10
10	to a certain extent	stance	4	2	4	10
11	in terms of	referential	5	0	3	8
12	in the case of	discourse	7	2	4	13
13	it is	stance	2	4	0	6

	important to					
14	a large number of	referential	5	1	2	8
15	it can be argued that	stance	4	3	1	8
16	the majority of	referential	3	3	1	7
17	on the other hand	discourse	6	3	1	10
18	the findings suggest that	stance	6	2	2	10
19	for the purpose of	discourse	5	2	3	10
20	as a result of	discourse	6	2	0	8
21	the role of	referential	6	1	0	7
22	in contrast to	discourse	5	1	2	8
23	at the same time	discourse	4	3	1	8
24	as well as	referential	4	1	2	7
25	in other words	discourse	4	2	1	7
26	with respect to	referential	3	2	2	7
27	due to the fact	referential	4	2	2	8
28	one of the	referential	3	2	2	7
29	it is possible that	stance	3	3	1	7
30	the purpose of this	discourse	1	0	0	1

The top-30 table shows two clear patterns. First, ChatGPT produces many tokens for discourse and stance bundles; the model repeatedly uses certain academic linkers such as *"the purpose of"*, *"in the case of,"* and *"on the other hand."* Second, intermediate writers show high token counts across many of the same bundles but spread their use across a broader set of types; as a result, intermediate essays contribute strongly to total tokens and to type counts. Pre-intermediate writers contribute relatively few tokens to the top bundles and sometimes produce near-forms (simulated corruptions), which reduces exact-match counts.

Taken together, these patterns explain the earlier statistical outcome in which ChatGPT's overall type/token ratio came between the two human bands. ChatGPT produces many tokens but tends to repeat a limited set of high-frequency bundles. Intermediate students use more distinct bundles overall, which raises their type counts relative to tokens. Pre-intermediate students produce few occurrences of academic bundles and sometimes use incorrect forms that do not register as exact matches. This leads to low token totals and low measured diversity.

Table 2 summarizes the total number of bundle tokens, bundle types, and resulting type/token ratios (TTR) for each group. The table provides a concise overview of how the three writing groups differ in lexical diversity based on their bundle use. TTR indicates the proportion of distinct bundle types relative to the total number of bundle tokens. A higher TTR represents greater lexical variety, while a lower TTR reflects repeated use of the same bundles.

Table 2

Lexical bundle token and type counts with type/token ratio by group

Group	Tokens (top-30 only)	Types (top-30 found in group)	TTR (Types ÷ Tokens, top-30)
ChatGPT	126	30	0.2381
Intermediate	151	30	0.1987
Pre-intermediate	45	21	0.4667

The results in Table 2 show that ChatGPT and intermediate writers produced similar numbers of bundle types, although ChatGPT used slightly fewer total tokens. This outcome indicates that the model's bundle diversity approximates the performance of the higher-proficiency learners. Pre-intermediate writers displayed a higher TTR, but this occurred because they produced few bundles overall. When the number of tokens is small, TTR increases even if the range of expressions is limited. Therefore, the higher TTR in the pre-intermediate group should not be interpreted as evidence of higher lexical variety. It instead reflects limited bundle production and reduced repetition.

The two independent raters scored the student essays using the five-dimension analytic rubric (each dimension 0–6, raw total 0–30) and scores were converted to a 0–100 scale for reporting and placement validation. Table 3 summarizes the distribution of scores provided by each rater across the student sample.

Table 3

Descriptive statistics for rater scores

	N	Minimum	Maximum	Mean	Std.

					Deviation
Rater 1	40	10	25	60.67	13.00
Rater 2	40	9	24	58.67	14.00

A slightly higher average score was reported by Rater 1 (M = 60.67) compared to Rater 2 (M = 58.67). A Pearson correlation analysis measured the agreement between the raters' writing scores.

Table 4

Inter-rater reliability between the raters' writing scores

		Rater 2	Rater 1
Rater 2	Pearson Correlation	1	.885**
	Sig. (2-tailed)		.000
	N	30	30

**. Correlation is significant at the 0.01 level (2-tailed).

The results of Pearson correlation of ($r = .88$, $p < .001$) indicates strong agreement between raters on the writing scores. To ensure transparency about how student proficiency bands were defined, Table 5 reports the descriptive statistics of the institutional placement scores used for band assignment.

Table 5

Placement-score descriptive statistics by proficiency band

Proficiency band	N	Placement-score range	Minimum	Maximum	Mean	Std. Deviation
Pre-intermediate	20	30-54	4.80	36.1	44.68	54.9
Intermediate	20	55-80	4.34	55.4	65.45	74.5
Total	40		11.45	36.1	55.07	74.5

Placement test results demonstrated distinct separation between proficiency levels. Pre intermediate students achieved a mean score of 44.68 with a standard deviation of 4.80. Intermediate students obtained a mean score of 65.45 with a standard deviation of 4.34. The selected cutoffs (30–54 vs ≥55) produced minimal overlap between bands in the sampled population and are consistent with institutional norms; therefore, the subsequent group comparisons rest on a defensible proficiency classification.

Table 6 presents the basic distributional properties of the bundle type/token ratio (TTR) for each group.

Table 6

Descriptive statistics for lexical bundle type/token ratio (TTR) by group

Group	N	Mean (TTR)	Std. Deviation	Minimum	Maximum
ChatGPT	40	0.3242	0.1028	0.1500	0.5294
Pre-intermediate	20	0.2670	0.1046	0.1304	0.4783
Intermediate	20	0.4088	0.0746	0.2941	0.5833

Intermediate students demonstrated the highest mean type-token ratio (M = 0.4088). This indicates the greatest diversity of lexical bundles in their writing. Pre-intermediate students showed the lowest mean ratio (M = 0.2670). ChatGPT's mean ratio was positioned between these two groups (M = 0.3242).

Table 7 presents the results of the independent-samples t-test that compared the mean lexical bundle type/token ratios between ChatGPT essays and those written by all student participants. It was conducted to investigate the first research question.

Table 7

Independent-samples t-test for lexical bundle type/token ratio (ChatGPT vs Students)

	Levene's Test for Equality of Variances		t-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% CI Lower	95% CI Upper
Equal variances assumed	0.369	0.545	−0.56	78	0.575	−0.0137	0.0243	−0.0623	0.0349

The results indicated that the mean lexical bundle type/token ratio of ChatGPT essays did not differ significantly from the students' essays ($t = -0.56$, $p > .05$). The mean difference was very small (−0.0137) with a narrow confidence interval (−0.0623, 0.0349). This result showed that ChatGPT produced a level of lexical bundle diversity similar to that of the students.

Table 8 presents the independent-samples t-test comparing lexical bundle type/token ratios between the pre-intermediate and intermediate proficiency groups.

Table 8

Independent-samples t-test for lexical bundle type/token ratio (Pre-intermediate vs Intermediate)

	Levene's Test for Equality of Variances		t-test for Equality of Means						
	F	Sig.	t	Df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% CI Lower	95% CI Upper
Equal variances assumed	2.079	0.152	−4.94	38	0.000	−0.1419	0.0287	−0.2002	−0.0835

The results showed a statistically significant difference between the two proficiency levels, ($t = -4.94$, $p < .05$). The mean lexical bundle type/token ratio of intermediate students ($M = 0.409$) was considerably higher than that of pre-intermediate students ($M = 0.267$). In other words, higher proficiency is associated with greater lexical bundle diversity in argumentative writing.

Discussion

ChatGPT output repeated a small set of canonical academic linkers. Repetition increased the number of bundle tokens while it limited the number of distinct bundle types. The result produced a moderate type/token ratio for ChatGPT that lay between the pre-intermediate and intermediate bands. This pattern followed from the model's tendency to prefer well-formed, high-probability sequences in argumentative prompts and from intermediate writers' tendency to use a wider set of discourse and stance expressions. The pattern was robust across the raw and normalized extraction streams.

Findings that machine-generated prose often concentrates on frequent formulaic linkers are reported in large-scale comparisons of ChatGPT and human essays (Herbold et al., 2023). Studies that examine lexical bundles in academic registers show that published academic prose contains recurrent formulaic packages, while developing student writers expand their repertoire with more varied bundle types as

proficiency rises (Biber, et al., 2004; Hyland, 2008). Recent analyses of ChatGPT and student argumentative essays report that model output can mimic fluency but not the same breadth of lower-frequency bundles used by higher proficiency writers (Jiang, 2025; Zhang, 2025). The present finding therefore aligns with work that documents canonical linker concentration in model output and with research on developmental increases in bundle variety among learners (Biber et al., 2004; Hyland, 2008; Herbold et al., 2023; Jiang, 2025).

A comparison of ChatGPT essays with the pooled student sample yielded a non-significant mean difference in type/token ratio. Two factors explained this null outcome. First, the pooled student distribution combined two heterogeneous proficiency bands that lay on opposite sides of the model mean. Second, within-group variance was large relative to the small mean gap between ChatGPT and the combined students. The null result therefore reflected composition and variance rather than precise equivalence of writing behaviors.

Warnings about pooling heterogeneous learner samples are frequent in corpus and pedagogical research. Aggregation across proficiency bands can mask developmental contrasts in bundle frequency and function (Cortés, 2004). Studies that compared machine output to human texts found that group-level averages sometimes conceal subgroup patterns that emerge when proficiency or writer experience is considered (Herbold et al., 2023; Jiang, 2025). The present null finding is consistent with those methodological cautions and supports the use of disaggregated analyses when proficiency differences are expected (Cortés, 2004; Herbold et al., 2023; Jiang, 2025).

Intermediate writers used a wider set of discourse-organizing bundles and stance markers than pre-intermediate writers. Intermediate essays thus contained more distinct bundle types per essay. Pre-intermediate essays showed more collocational and orthographic near-forms. Many near-forms failed exact-match extraction and therefore reduced measured token counts for lower-proficiency writers. The correlation between placement scores and type/token ratio supported the interpretation that bundle diversity reflected developing discourse competence.

The link between proficiency and bundle variety is well documented. Studies of student writing show that bundle inventories expand with instruction and exposure to disciplinary discourse, yielding greater use of stance and discourse bundles at higher proficiency levels (Hyland, 2008; Wright, 2019). Empirical work on L2 learners has found that explicit instruction and practice increase the frequency and functional variety of bundles in student writing (Granger, 2015; Nesi & Basturkmen, 2006). The present proficiency gradient therefore matches prior descriptions of developmental change in lexical bundle use and aligns with findings

that higher placement or test scores predict greater discourse sophistication (Hyland, 2008; Wright, 2019; Granger, 2015).

ChatGPT produced polished canonical linkers but did not disperse its output across as many low-frequency bundle types as intermediate writers. This production profile created a measurable gap versus intermediate students on diversity metrics while producing partial overlap on frequent linker tokens with pre-intermediate writers. The result followed from the model's generation strategy that favors high-probability n-grams and from human writers' varied rhetorical choices.

Studies that compare LLM output and student writing report similar mixed patterns. Large-sample human ratings favored ChatGPT in some quality indices while linguistic analyses identified distributional differences in markers of engagement, epistemic stance, and lexical dispersion (Herbold et al., 2023). Other recent corpus comparisons report that ChatGPT essays show constrained variation in particular bundle classes and lower authorial presence compared with student essays, even when surface fluency is high (Zhang & Crosthwaite, 2025; Azalia et al., 2025).

Discourse-organizing and stance bundles dominated the top-frequency inventory for ChatGPT and for intermediate writers. The argumentative genre required structuring moves and evaluative language. ChatGPT's training on formal and instructional texts increased the probability of discourse and stance markers in model completions. Intermediate writers used similar classes of bundles but dispersed their use over a wider set of forms.

Functional analyses of academic registers show a consistent link between argumentative and expository tasks and high incidence of stance and discourse bundles (Hyland, 2008). Corpus studies show that discipline, genre, and purpose shape the functional distribution of bundles, with stance and discourse markers prevalent in argumentative writing (Biber et al., 2004; Wright, 2019). Recent AI-human comparisons also highlight the prevalence of discourse and stance features in model outputs but note differences in density and variety relative to human-produced texts (Herbold et al., 2023; Jiang, 2025). The observed functional distribution therefore matches prior genre-based descriptions while confirming that model training data bias contributes to the high incidence of these bundle classes in AI output.

Extraction parameters and normalization rules altered observed counts and diversity metrics. Exact contiguous 3-gram extraction favored canonical sequences and undercounted learner near-forms. Application of conservative normalization rules recovered some student matches and shifted TTR values. The dependence of

results on operational choices therefore explained part of the between-group pattern.

Methodological work on bundles and lexical diversity has shown that extraction decisions, n-gram length, and matching criteria strongly affect inventories and diversity indices (Hyland, 2008). Studies on lexical diversity recommend the use of multiple measures because TTR alone is sensitive to text length and sampling variability (Covington & McFall, 2010; Treffers-Daller, 2018). Empirical comparisons of raw and normalized extractions have found meaningful shifts in bundle counts after orthographic correction and canonicalization of near-forms (Cortes, 2004; Biber et al., 2004). The present sensitivity analysis therefore follows established methodological practice and supports the inclusion of robustness checks in bundle research.

Conclusion

The purpose of this research was to compare lexical bundle diversity in argumentative essays produced by ChatGPT and in essays written by Iraqi EFL learners at two proficiency levels, and to evaluate how bundle measures relate to measured placement scores. To address these objectives, balanced corpora were collected from human participants and from multiple ChatGPT generations. Texts were preprocessed in two streams, raw and normalized, and contiguous three-word and four-word bundles were extracted per essay. Writing quality was rated with an analytic rubric by two blinded raters and inter-rater agreement was quantified. Linear mixed-effects models that treated prompt as a random effect were used to test group contrasts while controlling for essay length. Robustness checks included alternative diversity indices and analyses that averaged across multiple model completions. These procedures were implemented to ensure transparency and to reduce bias from extraction settings and from prompt variation.

Intermediate-level learners produced a greater variety of lexical bundle types per essay than pre-intermediate learners, and this difference was large and statistically robust. ChatGPT output showed frequent use of canonical discourse and stance linkers while it concentrated tokens on a narrower set of bundle types; this profile placed ChatGPT between the two human bands on the per-essay type/token ratio. A comparison that pooled both human bands produced no reliable difference from ChatGPT because the pooled human mean lay near the model mean and within-group variance was substantial. Normalization rules increased some learner bundle counts and altered measured diversity, which shows that preprocessing choices affect quantitative outcomes. Placement scores correlated positively with bundle diversity, which supports interpreting bundle measures as indices related to discourse competence rather than as mere surface frequency.

Several implications follow from these findings. For classroom practice, explicit instruction in discourse-organizing and stance bundles is recommended, with exercises that encourage productive variation rather than rote repetition. Assessment rubrics should include measures of bundle dispersion and functional use to supplement holistic fluency scores. For curriculum design, tasks that elicit varied rhetorical moves should be introduced across levels so that learners can expand their bundle repertoire. For research methodology, reporting of preprocessing rules, of extraction parameters, and of model-generation metadata should be required. Multiple diversity metrics and mixed-effects models are advised to reduce bias from text length and topic variation. When comparison with model output is intended, multiple independent generations per prompt should be sampled and reported.

Limitations of this work require careful notice. The participant sample was drawn from a single national context, which limits claims about generalizability to other EFL populations. The prompts were neutral argumentative items, which restricts inference to this genre and task type. Contiguous n-gram extraction and exact-match criteria can undercount learner near-forms, even when normalization was applied. ChatGPT outputs reflect a particular model version and parameter set; different versions or generation settings may yield different lexical dispersion. Finally, although multiple robustness checks were performed, the choice of diversity indices influences effect sizes and some contrasts.

Recommendations for future research follow directly from these limitations and from the present findings. Replication with learners from diverse instructional contexts and with a wider range of genres is needed. Longitudinal designs that collect multiple texts per writer will permit stronger inferences about intra-writer development of bundle repertoires. Experimental work that teaches and tests intentional variation of bundles will clarify causality between instruction and diversity. Comparative work should sample multiple model versions and systematically vary prompt framing and generation parameters to map model variability. Finally, replication samples and analysis code should be preregistered and archived so that preprocessing decisions and extraction scripts can be inspected and reused.

References

Abdalhussein, H. F. (2022). Iraqi EFL learners' use of formulaic language in writing proficiency exams. *Journal of Language and Linguistic Studies, 18*(1), 1079–1093. https://doi.org/10.52462/jlls.240

Azalia, S., Widodo, P., Wiedarti, P., & Sudartinah, T. (2025). Lexical bundle structure and function of ChatGPT-generated essays: Corpus-based study of

advanced learners in Indonesia. *Journal of Applied Linguistics and Literature (JOALL), 10*(2), 515–533.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press. https://doi.org/10.1017/CBO9780511801686

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

Chen, Y.-H., & Baker, P. (2016). Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistics, 37*(6), 849–880. https://doi.org/10.1093/applin/amu065

Cortés, 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

Cortés, 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

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/10.1080/09296171003643098

Darweesh, A. D., & Kadhim, S. A. (2016). Iraqi EFL learners' problems in using conjunctions as cohesive devices. *Journal of Education and Practice, 7*(11), 169–180. (ERIC No. EJ1099615)

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports, 13*, Article 45644. https://doi.org/10.1038/s41598-023-45644-9

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). *A large-scale comparison of human-written versus ChatGPT-generated essays. Scientific Reports, 13*, Article 18617. https://doi.org/10.1038/s41598-023-45644-9

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

Javanbakht, Z. O. (2011). The impact of tasks on male Iranian elementary EFL learners' incidental vocabulary learning. *Language Education in Asia, 2*(1), 26–45.

Javanbakht, Z. O. (2014). Attitudes of Iranian teachers and students towards internship. *Journal of Applied Linguistics and Language Research, 1*(1), 50–64.

Jiang, F., & Hyland, K. (2025). Does ChatGPT argue like students? Bundles in argumentative essays. *Applied Linguistics, 46*(3), 375–391. https://doi.org/10.1093/applin/amae052

Lee, J., & Soylu, M. Y. (2023). ChatGPT and assessment in higher education. *An interview with A. Goel and S. Harmon. Centre for 21st Century Universities*.

Nesi, H., & Basturkmen, H. (2006). Lexical bundles and discourse signalling in academic lectures: Examples from the international corpus of English. *International Journal of Corpus Linguistics, 11*(3), 283–304. https://doi.org/10.1075/ijcl.11.3.04nes

Treffers-Daller, J. (2018). Back to basics: How measures of lexical diversity can help language teaching and research. *Language Teaching, 51*(2), 146–165. https://doi.org/10.1017/S0261444817000411

Wright, H. R. (2019). Lexical bundles in stand-alone literature reviews: Sections, frequencies, and functions. English for Specific Purposes, 54, Article 100783. https://doi.org/10.1016/j.esp.2018.09.001

Zhang, M., & Crosthwaite, P. (2025). More human than human? Differences in lexis and collocation within academic essays produced by ChatGPT-3.5 and human L2 writers. *International Review of Applied Linguistics in Language Teaching, 63*(1), 1–22. https://doi.org/10.1515/iral-2024-0196