Research Article

Mengxuan Zhang and Peter Crosthwaite*

More human than human? Differences in lexis and collocation within academic essays produced by ChatGPT-3.5 and human L2 writers

https://doi.org/10.1515/iral-2024-0196
Received June 27, 2024; accepted December 22, 2024; published online January 23, 2025

Abstract: ChatGPT and similar generative AI models are increasingly being integrated into educational practice, including second language (L2) writing. However, AI-generated writing may significantly differ from that of developing L2 writers, raising potential issues over academic integrity. With AI detectors still underused in many contexts due to accuracy concerns, it is essential to determine if and how AI-generated texts are similar or different to those of L2 writers. This study compares lexical and collocation choices between essays on the same topic generated by ChatGPT (version 3.5) and human L2 writers using corpus analytic techniques. R statistical software was used for analysis and comparison. Results suggest that when tasked with producing academic essays under the same topic, ChatGPT-3.5 excels in generating texts with formal and complex vocabulary suited for academic and technical themes, while human L2 writers tend to focus on personal and social issues, using more varied and context-rich vocabulary. Accordingly, educators seeking to detect AI use should recognise the distinct linguistic and cultural experiences L2 writers bring to their writing, while students using AI need to understand the nature of the machine-like clarity and directness of expression as produced by ChatGPT as compared with their own typical production.

Keywords: artificial intelligence; ChatGPT; generative AI; L2 writing; lexical diversity; collocations

***Corresponding author: Peter Crosthwaite**, School of Languages and Cultures, The University of Queensland, South Brisbane, QLD, Australia, E-mail: p.cros@uq.edu.au. https://orcid.org/0000-0002-1482-8381
Mengxuan Zhang, School of Languages and Cultures, The University of Queensland, South Brisbane, QLD, Australia

1 Introduction

Generative AI (GenAI) and Large Language Models (LLMs) have now become a widely discussed topic, already attracting the interest of billions of users (Rafique et al. 2024). GenAI is an artificial intelligence (AI) technology capable of generating new content including text, images, audio, or video, rather than merely analysing or processing existing data (Hutson 2021). LLMs allow GenAI to generate linguistic content similar to humans by learning from extensive training data (Steiss et al. 2024). Notable generative AI models include the GPT (Generative Pre-trained Transformer) series, DALL-E, and VQ-VAE (Ramesh et al. 2022).

Since 2022, the advent of ChatGPT has propelled the development of large-scale natural language generation systems. ChatGPT, developed by OpenAI, is a generative AI model based on the GPT architecture (Wu et al. 2023). ChatGPT can understand prompts and generate text in natural language, excelling in various complex linguistic tasks e.g., answering questions, providing suggestions, and composing texts. ChatGPT has already been widely applied across various fields including education, customer service, content creation, and research assistance. At the end of January 2023, the public version of ChatGPT surpassed 100 million monthly active users within just two months of this launch (Wu et al. 2023). Currently, ChatGPT offers the freely accessible GPT-3.5/4o versions as well as other premium versions available through a paid subscription (Khan et al. 2024). The success of ChatGPT has led many AI companies to follow suit, igniting an “AI revolution”. Early users shared their experiences on social media, with most holding positive attitudes (Wilson et al. 2022).

Multiple studies have shown that ChatGPT can serve as a powerful auxiliary tool in academia, potentially significantly enhancing the teaching and learning of academic writing skills. Due to its extensive training data, ChatGPT can provide instant feedback on grammar, lexis and coherence, easily extract key concepts, and even offer relevant citations (Kung et al. 2023). In a recent experiment, ChatGPT was prompted to produce a high-quality physics essay that received a first-class score in the UK’s higher education system (Yeadon et al. 2023). However, this has also raised ethical concerns regarding academic plagiarism and issues around AI authorship in academic settings (Stokel-Walker 2022). According to a comprehensive SWOT (Strengths, Weaknesses, Opportunities, Threats) analysis of ChatGPT’s impact on education (Farrokhnia et al. 2024), despite strengths in terms of information access, reducing teaching workloads, and personalised learning, weaknesses of ChatGPT include limited understanding of text topics, inability to critically evaluate information, and a lack of higher thinking skills. Threats encompassed risks to academic integrity, increased plagiarism, and contextual limitations.

As previous research has primarily focused on employing qualitative methods conveying multiple end-user perspectives on the efficacy of GenAI for academic writing (Barrett and Pack 2023; Li et al. 2024), there is a distinct gap in quantitative studies comparing AI-generated academic texts with those of human L2 writers. While research on collocation frequently relies on the intuitive judgments of L2 writers, which can introduce subjectivity due to individual variations (Xia et al. 2023), GenAI applications employ complex statistical processes to generate collocates in text. However, these lack transparency, complicating the replication of specific outputs. If "linguistic knowledge is essentially collocational knowledge" (Nation 2013, p. 321), it is possible that limitations in training data or model weightings could mean the lexical and collocational choices made by AI-generated texts may not reflect the language typically employed in authentic writing or conversation (Crosthwaite and Baisa 2023), or they may not be register or contextually appropriate. In contrast, human L2 writers draw upon diverse personal, social, and linguistics experiences, enriching their writing with depth and authenticity of lexis. Likewise, studies on L2 collocation have revealed student writing is characterised by a narrower and more repetitive use of multi-word expressions (e.g. Hoang and Crosthwaite 2024). Subsequent adjustments to the training data, model weightings or prompts used may be necessary to mitigate differences between human L2 writing and that of GenAI if L2 writers employ GenAI to aid in the academic drafting or editing process. Yet, while studies exploring the nature of GenAI academic vocabulary are now coming to the fore (e.g. Jiang and Hyland 2024), the field still lacks research investigating lexical and collocation differences between texts generated by ChatGPT and human L2 writers. In particular, it is important to understand the key differences in the lexical choices made by developing L2 writers (bearing in mind both their development as academic writers and second language users), as compared with error-free generative AI production if practitioners are to target teaching materials and/or utilise generative AI applications in the L2 writing classroom. It is also important for students to understand these differences when using GenAI as part of the writing process, for example at the editing/revision stage of their work, as seen in Allen and Mizumoto (2024). Understanding differences in lexis generated by GPT versus that of human L2 writers may also help educators of academic writing more easily identify AI-generated work, aiding in detection of plagiarised content in the absence of reliable detection software (Jiang and Hyland 2024). Accordingly, the following research question is addressed by the current study:

> *What are the differences in lexis and collocation between academic essays produced by ChatGPT-3.5 compared with human L2 writers?*

2 Literature review

2.1 Generative AI in education

Generative AI (GenAI) in this study refers to software tools utilising Large Language Models (LLMs) trained on extensive text data that are able to generate human-like text, answer questions, and perform various language-related tasks with high accuracy (Kasneci et al. 2023). These tools demonstrate capabilities in producing intelligent-sounding responses to users' natural language prompts and are proficient in generating coherent and contextually relevant text reflecting advanced natural language processing abilities (Stokel-Walker 2022). ChatGPT as a prominent GenAI tool, offers numerous advantages by providing real-time access to vast amounts of information, thus facilitating the acquisition of (nearly) up-to-date and accurate data for literature and research (Wang et al. 2023).

The application of GenAI has gained significant attention in educational research as a tool to assist teachers and educators in creating more flexible, efficient, inclusive, engaging, and customised learning environments. For instance, GenAI can collaborate with teachers to accomplish tasks that are challenging for teachers to complete alone, such as grading a large volume of student assignments and providing the necessary support and scaffoldings for learners, including real-time feedback (Chan and Hu 2023). Human-machine collaboration not only reduces the workload of teachers but also enhances the effectiveness and efficiency of education, enabling learners to receive a more personalized educational experience. Cooper (2023) highlighted that educators can gain a deeper understanding of the learning process by incorporating GenAI across multiple domains. e.g., providing a mentor for each learner, helping learners and teachers acquire new 21st-century skills, enhancing data interaction, enabling the use of global classrooms, and supporting lifelong learning. Yang and Kyun (2022) provided similar findings and conclusions through an analysis of 25 empirical research papers on AI-supported language learning, namely that: (a) AI provides and supports more effective language learning and collaboration among learners; and (b) combining AI tools with teacher guidance can produce better learning outcomes and achieve higher aims.

2.2 (mis)Applications of generative AI for L2 writing

AI has also demonstrated itself as a potentially powerful tool in supporting L2 writing and pedagogy, encompassing a range of tools such as automatic translation, error correction, and automated scoring programs, all of which are now widely used in L2 writing classrooms (Stokel-Walker 2022). The impressive capabilities of GenAI writing tools in generating intelligent text in response to user prompts have potential for creating high-quality texts, increasing learners' interest in L2 writing and helping them to achieve higher learning outcomes (Zhao 2023). Previous research has emphasized that incorporating GenAI technology into the classroom significantly promotes students' L2 writing via providing instant, personalised feedback (Atlas 2023). Wang et al. (2023) highlight that even many academic practitioners already use GenAI to summarize literature, discuss and revise manuscripts, draft research proposals, and identify research gaps. For instance, Kung et al. (2023) and Cotton et al. (2024) found that ChatGPT assists researchers in rapidly collecting and analyzing relevant literature. Chan and Hu (2023) noted that integrating GenAI into L2 writing offers benefits such as reducing learners' writing time, costs, frustration, and anxiety, providing quick assessments through timely feedback, and predicting learners' future performance. ChatGPT also excels in L2 writing assessment, providing more objective and faster scoring than humans, with Xia et al. (2024) finding ChatGPT achieved an accuracy rate of 84.38 % when scoring various types of essays.

However – and rightly so – educators remain concerned about potential academic integrity issues, and the debate on how to integrate GenAI tools to enhance English writing skills while adhering to academic integrity principles is ongoing. For example, Guo et al. (2023) discovered that the text generated by ChatGPT-3.5 is remarkably similar to human-written text. In an experiment, participants unfamiliar with ChatGPT were asked to distinguish between answers provided by ChatGPT and those generated by humans. More than half of the participants could not differentiate between the two and found ChatGPT's answers to be more detailed and useful. This challenge to distinguish between GenAI generated outputs and learners' original writing is crucial in maintaining academic integrity (Cotton et al. 2024). Studies suggest that traditional university assignments are in danger as ChatGPT can now generate high-scoring essays and provide correct answers to problem questions (Al-Zahrani 2023; Crawford et al. 2023). This challenge is compounded by the fact that the ethical and acceptable boundaries of using ChatGPT in academic writing remain unclear to both students and staff (Tlili et al. 2023). Given the rapid evolution of generative AI tools and their increasing integration into educational contexts, it is imperative to adopt a critical perspective towards these technologies and educate leaners on their appropriate and ethical use, as underscored by recent research (e.g., Guo et al. 2023; Tlili et al. 2023).

Specifically, regarding lexis, studies have found that GPT exhibits notable performance in lexical choices and their application in generated academic texts. For instance, Li et al. (2024) analysed the vocabulary usage of GPT in academic English witing and discovered that it excels in terms of lexical diversity and precision. Additionally, AlAfnan and Mohdzuki (2023) examined GPT-generated academic texts highlighting the model's high consistency and accuracy in employing technical terms and academic vocabulary. Advances in LLM technology are also closing the gap, given ChatGPT-4 shows significantly higher lexical diversity when compared with ChatGPT-3 (Herbold et al. 2023). Enhancements in GPT's ability to express politeness and handle deixis within texts have also improved the naturalness and contextual awareness of its generated content (Rafique et al. 2024). Nonetheless, despite these advancements, ChatGPT's performance remains inferior to that of human writers in terms of lexical diversity, nuanced collocational usage and sensitivity to contextual appropriateness (e.g. Jiang and Hyland 2024).

However, almost all research to date has focused on the differences between GenAI-produced writing and that of monolingual L1 writers (e.g. Bašić et al. 2023; Li et al. 2024). Studies comparing differences between GenAI-produced writing and those of developing L2 academic writers are far less prominent. Examples include the impact of ChatGPT's automatic text generation on undergraduate students' perceptions and behaviours in learning L2 writing skills (Yan 2023), or studies focusing on the use of ChatGPT as a tool for influencing and guiding responses in higher-education students' L2 writing (Zou and Huang 2023). Understanding differences in texts produced by L2 writers compared with GenAI is important because educators need to identify the specific advantages and disadvantages of AI-generated content compared to L2 human writing if they are to use it correctly as a more effective writing aid, supplementing human teaching and thereby enhancing the overall quality of L2 writing education. Furthermore, this understanding helps set appropriate expectations for the use of AI in academic contexts, ensuring that L2 students and educators can make informed decisions when integrating these technologies into their learning and teaching processes.

3 Methodology

This study employs a quantitative research method, employing R studio and the R programming language to extract and analyse lexical items and collocates across a corpus of academic essays produced by ChatGPT and human L2 writers for comparative analysis.

3.1 Data

The data is divided into two sources: one containing L2 human-produced academic writing (L2 human corpus) and the other GenAI-generated academic writing (GPT corpus). In the L2 group, essay topics are sourced from a corpus of 30 argumentative essays each on different topics taken from an undergraduate level academic writing course at an Australian University. The course is designed to improve undergraduate students' L2 academic writing skills, while the essay task is designed to encourage students to incorporate critical thinking into their writing, using evidence and citations concisely and clearly to articulate their positions. Feedback for each essay was provided at the planning stage by English-speaking academic writing instructors, so as to help students develop their drafting, revising and critical thinking abilities. Students are required to use formal vocabulary and academic language to successfully complete the task. Although the students' names and ages are not disclosed, they are all undergraduates, aged between 18 and 20 years. They are non-native English speakers with diverse first languages and an English proficiency level ranging from upper intermediate to advanced (CEFR B2-C1), with IELTS scores between 6.5 and 7.0.

Essay topics cover various fields according to the students' own selection. In selecting the texts for analysis, care was taken to diversify the range of topics to prevent overlap and to cover various fields, as listed in Appendix 1. By including a wide range of topics, any bias that could emerge from lexical variation tied to a single subject (e.g., specialised terminology in a particular field) is reduced. Each essay in the corpus follows APA 7th edition citation and referencing format with an average length of 1,170 words excluding references.

ChatGPT-3.5 was employed as an additional source of data, comprising 30 essays generated by ChatGPT-3.5 using the following prompt: "Create a 1,000-word academic text about [topic], using APA 7th edition citations excluding references and without subheadings". A new chat window was opened for each essay. The topic element was taken verbatim from the equivalent topic from the L2 corpus via prompting to ensure comparability. The generated texts were copied into Microsoft Word documents to facilitate analysis. The average length of the ChatGPT-3.5-generated corpus was 896 words with the longest essay being 1,100 words, not including reference lists.

The analysis of lexical and collocational patterns in both the L2 and GPT corpora was conducted using R statistical software. As an open-source, high-level programming language and environment, R is highly flexible and supports faster processing on all major operating systems, including Linux, Windows, and Mac, without requiring licensing compliances (Kumar and Tyagi 2024). While corpus linguistics software such as AntConc and Sketch Engine are commonly employed in linguistic studies (Zih et al. 2020), R provides improved flexibility for automating processes, conducting custom statistical tests, and visualising the results (Schweinberger 2023), facilitating detailed comparisons between the lexical and collocational choices in each corpus.

3.2 Analysis

R Studio version 4.33 (2024.04+735, Posit Software, PBC) (within the R programming language) was employed as the instrument for all data analysis. R is a comprehensive programming environment encompassing complex natural language processing, statistical data analysis, data visualisation, and the creation of applications and websites (Schweinberger 2023; Verzani 2011).

Before importing the data into R Studio, the reference list from each text was manually removed and the file format was converted from "doc." to "txt." format. In R Studio, several essential data analysis packages were installed, including 'here', 'stringr', 'dplyr', 'ggplot2', 'quanteda', 'quanteda.textstats', 'tm (Text mining)' to process and promote the textual data with loading, searching, cleaning, annotating, and visualizing within a single tool (Arnold et al. 2019). Subsequently, the command 'library' was used to activate all analysis packages. Initially, the data paths for both groups were imported to identify the locations of the 60 academic essays. Then, the data were imported, and new filenames were assigned to each data group, with 'gpt' respesenting the GPT corpus and 'hmn' representing the L2 human corpus. The data for both corpora were reviewed to ensure successful importation. The next step involved cleaning all the data, conducted using the "qyanteda" package in R (Banks et al. 2018). This process assisted in converting all words in each text to lowercase, and numbers, functional words, prepositions (such as 'the', 'an', 'of', 'from'), pronouns (such as 'I', 'she', 'he'), and non-words (such as 'et al.' in citation) were removed. The most frequently used words in any corpus are often found together with function words that lack substantive meaning, making it essential to remove these function words at an initial stage. All remaining tokens were then compiled into a data frame named "texts_df", with the corresponding frequency of each token in both "gpt" and "hmn" recorded before the calculation of cross-corpora frequency, collocation and keyness according to the procedures below (Figure 1).

Figure 1 describes the analysis procedures within R programming language.

3.2.1 Frequency analysis

Frequency analysis is a method for extracting and comparing different terms, including but not limited to words, phrases, collocations, and sentences. In this study,

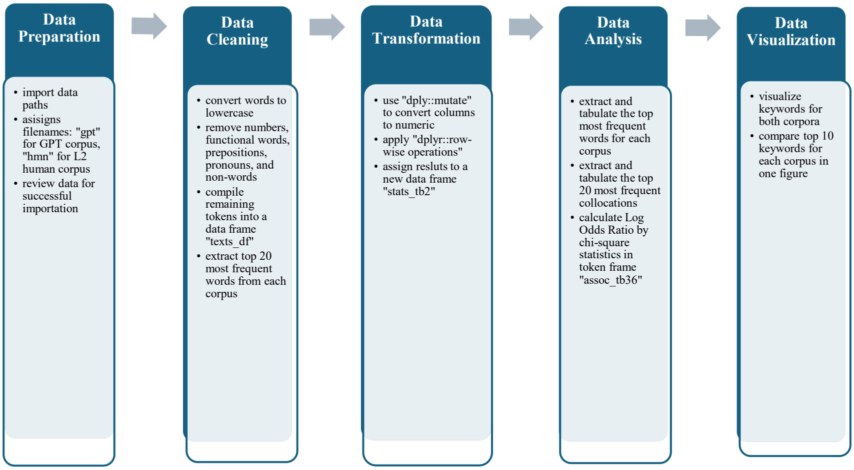
Data Preparation
• import data paths
• asisigns filenames: "gpt" for GPT corpus, "hmn" for L2 human corpus
• review data for successful importation
Data Cleaning
• convert words to lowercase
• remove numbers, functional words, prepositions, pronouns, and non-words
• compile remaining tokens into a data frame "texts_df"
• extract top 20 most frequent words from each corpus
Data Transformation
• use "dply::mutate" to convert columns to numeric
• apply "dplyr::row-wise operations"
• assign resluts to a new data frame "stats_tb2"
Data Analysis
• extract and tabulate the top most frequent words for each corpus
• extract and tabulate the top 20 most frequent collocations
• calculate Log Odds Ratio by chi-square statistics in token frame "assoc_tb36"
Data Visualization
• visualize keywords for both corpora
• compare top 10 keywords for each corpus in one figure

Figure 1: Overview of analysis procedures within R programming language.

word frequency was extracted as the first linguistic feature in this study, as identifying the most common words in a text forms the basis of textual analysis (Schweinberger 2023). These frequency data were organised into a frequency table, listing each word alongside its corresponding occurrence frequency. Given the vast number of words in each corpus, this study extracted and analysed the top 20 most frequent words. These words were tabulated to promote the analysis and comparison of disparities.

3.2.2 Collocation analysis

In linguistic and second language (L2) writing research, collocations are key to understanding both the syntactic structures and semantic nuances of a language, supporting learners' ability to produce fluent, native-like text (Hoover 2003). For L2 learners, recognising and using collocations enhances textual cohesion and fluency, which is particularly valuable when producing complex writing tasks such as essays. Identifying and comparing collocations across L2 and machine-produced texts allow us to explore the depth and range of L2 users' current productive knowledge of collocation while understanding their limitations in producing these complex multi-word units, unlike GPT which has access to a vast wealth of statistical collocational knowledge. In this study, the top 20 most frequent collocations in each corpus were extracted and analysed. Expanding our focus to the top 20, rather than a smaller 10 collocations as seen in previous research (cf. Daudaravicius 2010; Demir 2017), allows us to capture a broader range of collocational patterns and linguistic diversity given the small corpus sizes. This broader scope may more accurately reflect the range of lexical bundles and formulaic expressions that the L2 learners featured in the present study are able to produce given their current levels of proficiency.

3.2.3 Keywords

Regarding cross-corpus keyword calculation, the odds ratio (OR) is a crucial metric for determining differences across groups. An OR greater than 1 implies a strong likelihood of difference, an OR less than 1 implies a lower likelihood of difference, while an OR equal to 1 means there is no association between the exposure and the outcome. The formula for calculating the odds ratio is following:

$$\text{Odds ratio (OR)} = \frac{\text{odds of outcome in the exposed group}}{\text{odds of outcome in the non} - \text{exposed group}}$$

$$(\text{OR}) = \frac{O11 \times O22}{O12 \times O21}$$

The log odds ratio (Log OR) is the natural logarithm (log) of the OR. Log OR is widely used in statistical models, particularly in logistic regression analysis, due to its preferable mathematical properties compared to the raw OR. Log transformation aids in stabilising variance and linearising relationships between variables, which is beneficial for various statistical analyses. The formula for the Log OR is below:

$$\text{OR} = \frac{a \times d}{b \times c}$$

$$\log \text{OR} = \ln(\text{OR}) = \ln\left(\frac{a \times d}{b \times c}\right)$$

In the current study, the formula used is based on the Log OR, with an addition of 0.5 to each cell in a 2 × 2 contingency table. This adjustment, known as the Haldane Anscombe correction, addresses the issue of zero. Without this correction, an OR is involving a zero count would be undefined due to division by zero. Given that some terms in the generated data have zero counts in one corpus or another, applying an appropriate correction method is essential. The formula is used in this study:

$$\text{OR} = \log\left(\frac{(O11 + 0.5) \times (O22 + 0.5)}{(O12 + 0.5) \times (O21 + 0.5)}\right)$$

Due to the large volume of data, this study applied a filtering criterion with a *p*-value less than 0.001 to focus on statistically highly significant data. This criterion ensures that the likelihood of the observed association occurring by chance is extremely low, less than 0.1 %.

To visualise the Log OR as linguistic feature keywords, the top 20 Log OR keywords with $p < 0.001$ from each corpus were extracted and displayed in figures. The *Y*-axis represents tokens, while the *X*-axis indicated the association strength between tokens and Log OR.

4 Results

4.1 Word frequency

As previously mentioned, there are 30 different topics covering various fields, including the nuclear wastewater issue, education, AI, violence, the fashion industry, health, economy, and society (Appendix 1). As the students were free to choose their own topics, we could observe individual preferences and lexical tendencies in L2 academic writing based on students' personal experiences and perspectives.

Table 1 describes the highest frequency words in the L2 human corpus.

Table 1: Top 20 highest frequency tokens in 30 L2 human essays.

L2 human word frequency (top 20)		
Human	**Token**	**Countable**
1	Health	141
2	Sleep	135
3	Students	122
4	Children	112
5	Intelligence	90
6	Therapy	90
7	Journal	86
8	People	85
9	Art	83
10	Emotional	81
11	Global	73
12	Research	72
13	Vaccines	71
14	Human	71
15	Parents	70
16	Energy	68
17	Resources	68
18	Social	61
19	Study	59
20	Academic	58

The token “health” emerges as the most frequently mentioned word in the L2 corpus, appearing 141 times across 30 essays, equating to an average of at least 4.7 mentions per essay. Given the freedom students had to select their own topics, this frequency likely reflects a strong personal interest in health-related issues, possibly tied to the global relevance of health discussions in academic and societal contexts. Closely following are “sleep”, “students”, and “children”, each cited over 100 times, averaging more than three mentions per essay. This pattern indicates a strong focus on well-being within the essays, with related vocabulary including “intelligence” and “therapy” (mentioned 90 times), and “vaccines” and “human” (71 times). These words likely reflect broader societal concerns that are commonly discussed in academic settings, which supports the idea that L2 writers engage with socially relevant issues in their writing. The high frequency of words such as “students” and “children” also highlights a significant focus on educational themes and issues related to younger demographics. These results suggest that L2 writers frequently choose topics that are closely connected to their personally related factors, cognitive, affective (Kormos 2012), a trend observed in other studies on L2 composition.

Table 2 describes the highest frequency words in the GPT corpus.

Table 2: Top 20 highest frequency tokens in 30 ChatGPT essays.

ChatGPT-3.5 word frequency (top 20)		
ChatGPT-3.5	**Token**	**Countable**
1	Health	157
2	Challenges	112
3	Impact	107
4	Sleep	107
5	Energy	102
6	AI	99
7	Economic	94
8	Potential	94
9	Human	89
10	Development	85
11	Emotional	81
12	Academic	72
13	Addressing	71
14	Mental	71
15	Concerns	69
16	Contribute	68
17	Individuals	66
18	Essential	64
19	Students	64
20	Positive	63

In the GPT corpus, the most frequently mentioned word is the same as in the L2 human corpus (“health”), highlighting a shared focus on health-related topics, given the same topics as those from the human data were used to prompt GPT. Given this information, it is unsurprising that certain words like “health” are frequent in both corpora. However, the frequent appearance of abstract terms such as “challenges”, “impact”, and “energy” all appearing over 100 times in the GPT texts suggests a more generalised focus on societal and technological themes, contrasting with the more personal, human-centred themes in the L2 corpus.

The frequent mention of “AI (artificial intelligence)”, which appears 99 times, also underscores the importance of technology (namely it’s own) in GPT’s output as compared with L2 writers. Other words not appearing in the L2 corpus include “economic” and “potential”, “concerns”, “contribute”, “individuals”, “essential”, further indicating the GPT-produced texts’ different focus on addressing global issues rather than personal ones, despite sharing the same essay topics. Comparing the words of the two groups reveals that the human L2 writers are more focused on human-related aspects, such as education, therapy, and health, whereas the GPT corpus is more focused on relatively less-human content, including AI, economics, and technology. These differences suggest that even when prompted with similar topics, GPT’s training data and model prioritises more abstract, generalised content, while L2 writers tend to draw on their personal experiences and educational backgrounds.

Additionally, the top 20 words in the ChatGPT corpus generally have higher frequencies than those in the L2 human writing corpus, with each ranking differing by about 10 occurrences. This difference in frequency distribution reflects the varying focus and content generation strategies between the two groups. For example, despite “students” being a common term in both corpora, in the human group, “students” appeared 122 times, ranking third, whereas, in the GPT corpus, it appeared only 64 times, ranking second to last. This suggests that although both groups possibly only briefly touched upon the subject, human L2 writers might place a greater emphasis on discussions involving students - reflecting their personal experiences - while GPT-generated content has a broader thematic focus.

4.2 Collocations

All collocations listed in Tables 3 and 4 focus on bigrams or word pairs. These were chosen over longer collocation spans as shorter collocation occur more frequently in writing due to their flexible use across different contexts (Shin and Nation 2008),

Table 3: Collocations for L2 human corpus.

	Collocation	Count	Lambda	*Z*
1	Transformational leaders	15	6.581878	13.128988
2	Transformational leadership	11	6.039463	11.871059
3	Youtubes censorship	8	6.870197	9.994657
4	Youtube censorship	8	5.595665	10.833519
5	Nuclear effluent	8	6.722425	8.980217
6	Job satisfaction	8	8.141568	8.631254
7	Career growth	8	8.888261	6.003534
8	Artificial intelligence	8	8.88965	8.859977
9	Psychological society	7	8.50515	9.39683
10	Parents can	7	4.957989	9.381049
11	Mental health	7	8.04934	5.46441
12	Childhood obesity	7	7.405669	9.962692
13	Australian psychological	7	8.841795	9.151277
14	Nuclear wastewater	6	5.784348	8.974096
15	Job involvement	6	6.583115	9.672702
16	Digital tools	6	7.474051	9.11967
17	Content creators	6	4.693268	9.161273
18	Autism rate	6	6.938926	8.791304
19	Team members	5	10.652771	6.311643
20	Radioactive substances	5	8.252272	5.510587

providing a broader basis for our analysis. Table 3 describes the highest frequency collocations in the L2 human corpus.

As seen in the word frequency data above, the collocational content of these essays primarily focuses on leadership and work-related themes, as evidenced by the most frequent collocation is "transformational leaders", appearing 15 times, followed by "transformational leadership", which appears 11 times, revealing the importance placed on leadership within this L2 writing sample. This may reflect both academic and real-life interests of the students, as leadership is often a key topic in educational and professional development writing tasks, such as autobiographies or biographies where personal growth and leadership experiences are frequently highlighted (English 2006). "Job satisfaction" and "career development" discuss work and career planning issues, reflecting the writers' emphasis on professional aspirations.

Despite the variety of essay topics, health-related collocations such as "mental health" and "childhood obesity" also feature prominently aligning with previous research showing that L2 writers often draw on familiar, socially significant topics that reflect their language learning goals and self-efficacy beliefs, which influence their choice of personally of academically relevant subjects (Kormos 2012).

Table 4: Collocations for GPT corpus.

	Collocation	Count	Lambda	*Z*
1	Mental health	57	10.11613	7.095765
2	Renewable energy	39	10.122857	7.07466
3	Emotional intelligence	36	7.488245	19.104212
4	Vaccine hesitancy	30	8.373723	17.083199
5	Electric vehicles	29	9.238854	15.636198
6	Nuclear wastewater	28	9.333371	13.145297
7	Online dating	26	8.672825	15.120536
8	Public health	23	5.373373	15.703133
9	Sleep deprivation	23	9.165029	6.390532
10	Academic performance	22	6.921971	17.106692
11	Fiscal austerity	22	9.124817	10.510301
12	Childhood obesity	21	10.474156	12.239845
13	Crucial role	19	5.903993	16.720822
14	Fast fashion	19	11.223816	7.593572
15	Essay explores	18	10.21573	10.923345
16	Technical expertise	18	9.133943	12.41731
17	Climate change	17	9.680379	10.69595
18	Positive regards	16	9.756298	6.737236
19	Socioeconomic backgrounds	16	9.322161	12.258336
20	Employee motivation	15	7.621733	15.040136

Regarding collocation strength, a high *Z*-score indicates a high significance with a low likelihood of each word being found together by chance. In the L2 human corpus, the collocation “transformational leaders” exhibits the highest *Z*-score, indicating a statistically significant and strong association between these terms. Conversely, “team members” demonstrates the highest lambda value 10.65, suggesting a robust relationship despite its lower frequency. This pattern suggests the collocations with the strongest statistical associations are not necessarily those with the highest frequency. The data therefore indicate that L2 writers tend to rely on familiar word combinations while also making efforts to diversify their lexical choices. Overall, the lambda values of these 20 collocations are mostly concentrated around 6 to 8, indicating that their associations strengths are relatively similar.

Table 4 describes the highest frequency collocations in the GPT corpus. The data reveal some common collocations, such as “mental health”, “childhood obesity”, and “nuclear wastewater”, indicating that the vocabulary and collocations used in these areas are highly similar between the two corpora. However, there are significant differences between the two datasets. The content and collocations produced by GPT are more extensive and diverse, covering topics for health, renewable energy, vaccines, public health, academic writing, performance, fashion, and climate. Also, each collocation appears more frequently in the GPT data than in the L2 human writing data, with the lowest frequency GPT-produced collocation "employee motivation" appearing 15 times, the same as the highest frequency collocation in the L2 human data, "transformational leaders". This discrepancy may suggest that GPT-3.5, due to its training on extensive datasets, has the capacity to produce texts that consistently employ high-frequency collocations across a wide range of topics, whereas L2 writers may be more constrained by their familiarity with specific subject matter and depth of acquired vocabulary.

Regarding collocation strength, in the *Z*-score column, the highest score in Table is 19.10 for "emotional intelligence", far higher than the highest score of 13.12 in Table 3. The lambda values in Table 4 range from 5.37 to 10.65, with higher values indicating stronger associations between collocated words than those found in Table 3. These differences suggest how ChatGPT-3.5 and L2 human writers approach topics differently, reflecting distinct priorities and influences in their writing. GPT-3.5's extensive training data enables it to produce a greater variety of collocations with stronger internal associations, suggesting that it can deliver more lexically diverse content that L2 writers. The L2 writers' reliance on fewer collocations and topics more closely aligned with their experiences might indicate a tendency to engage deeply with familiar themes rather than exploring a broad range of subjects, as well as a limited range of productive collocations from which to draw on when writing.

It is important to note that these human authors are L2 writers, not L1 writers. This factor may explain many of the observed results, such as lower collocation frequencies and reliance on a smaller range of topics and collocations that the students are familiar with, given their limited L2 vocabulary. Also, in contrast to the data in the L2 corpus, none of the collocations in Table 4 are repeated, which suggests that GPT tends to diversify its collocational use across different contexts. This could potentially suggest GPT-generated texts tend to discuss all topics superficially without deep association or understanding of any particular field, relying solely on extracting content from its database when asked questions or given instruction rather than drawing on personal experience. The L2 writers' limited vocabulary and familiarity with certain topics result in lower collocation frequencies and a narrower range of discussed topics, contrasting with GPT's broader and more varied content generation.

4.3 Keywords

Table 5 describes the tokens and Log OR in the L2 human corpus.

Table 5: Log odds ratio of keywords specific to L2 human corpus (all $p < 0.001$).

	Token	Log odds ratio
1	Leaders	−0.98834
2	Youtube	−1.12277
3	Radioactive	−1.17545
4	Other	−1.20187
5	Marine	−1.27665
6	Job	−1.28441
7	Vaccines	−1.35541
8	Growth	−1.41925
9	Effects	−1.44484
10	Children	−1.46999
11	Censorship	−1.66186
12	Youtubes	−1.76212
13	Creators	−1.83139
14	Japan	−1.894
15	Autism	−1.98228
16	Employees	−1.99393
17	Body	−2.0229
18	Japanese	−2.0229
19	Possible	−2.0229
20	Videos	−2.04634
21	Different	−2.12342
22	Instance	−2.1231
23	Involvement	−2.1231
24	Elderly	−2.2142
25	Up	−2.24807
26	Seafood	−2.35941
27	Demographic	−2.49187
28	Transformational	−2.70635
29	Being	−2.63421
30	Explore	−2.63365
31	Material	−2.63365
32	Radiation	−2.63365
33	Shows	−2.63365
34	Food	−2.9729
35	Career	−2.75894

The Log OR values describe the relative frequency of these words generated by ChatGPT-3.5 compared to L2 human-produced L2 academic essays compared with those generated by ChatGPT 3.5. 35 tokens meet the high significance criterion for inclusion in Table 5. The Log OR values predominantly fall within the range of −1 to −3, with only one token below −1 and none exceeding −3. The more negative the value, the higher the association with the human group, so item 35 in Table 5 is the most likely to appear in the L2 data and not the GPT data. The word most closely associated with the human group and appearing most frequently is "food", with a Log OR of −2.972, it means that this token appears more significantly in human-produced content and has a higher association, whereas its frequency and association are lower in ChatGPT-generated content. Similarly, other words such as "youtube (−1.122)", "growth (−1.419)", "autism (−1.982)", "employees (−1.993)", "explore (−2.633)", and "children (−1.469)" are also more common in the L2 human corpus due to their connection with personal experiences, authentic context, and emotions. As previously mentioned, "leaders" also is associated with human L2 writing as compared with that of GPT.

Table 6 describes the tokens and Log OR in the GPT corpus. The data illustrate several trends in lexical selection by GPT that do not feature in L2 writing, indicated by positive Log OR. Notable examples e.g. the tokens "technical" and "initiatives"

Table 6: Log odds ratio of GPT corpus (all $p < 0.001$).

	Token	Log odds ratio
1	Educational	3.4210677
2	Technical	3.0553880
3	Play	2.9480120
4	Initiatives	2.9095022
5	Sustainable	2.8694533
6	Emotional	2.6926952
7	Vehicles	2.1698815
8	Addressing	2.1486006
9	Renewable	1.7380539
10	Ethical	1.6865393
11	Fiscal	1.5431336
12	Energy	1.5562788
13	Sleep	1.5449166
14	Crucial	1.4351039
15	Contribute	1.3255018
16	Essential	1.3079139
17	Individuals	1.2717868

with relatively high Log OR values indicate a strong association with GPT-generated texts and suggest a broader and more abstract lexical focus than that observed in the L2 corpus. Another significant token is “sustainable (2.869)”, indicating a high frequency of discussions on sustainability in ChatGPT-3.5 academic essays. Furthermore, tokens such as “vehicles (2.169)” and “renewable (1.738)” demonstrate that ChatGPT-3.5 frequently engages with subjects related to transportation and renewable energy.

When comparing the GPT-generated text table with the human-produced text table, several key differences are immediately apparent. In addition to the overall higher Log OR for tokens in the GPT corpus compared to the L2 human-writing corpus, the vocabulary in the ChatGPT table is mainly concentrated on themes such as technology and sustainability. In contrast, the L2 writer-produced text shows a prevalence of terms related to leadership, health, and specific real-world issues. Tokens such as “leaders”, “youtube”, “radioactive”, and “children” appear more frequently in human-produced texts as evidenced by their minus Log OR values in the GPT corpus. This suggests that human texts focus more on personal experiences, current events, and in-depth discussions of social issues. These distinctions are particularly relevant when considering that the human texts are produced by L2 writers. L2 writers might have a narrower range of vocabulary and less fluency compared to L1 writers, which could result in a stronger focus on familiar and concrete topics, such as leadership and health, and a greater reliance on real-world experiences and current events. This contrasts with ChatGPT-3.5, which has access to vast database of information and can generate content on a wide range of topics, including more abstract or technical themes.

Figure 2 visually describes the association strength between keywords and Log OR in the L2 human corpus. The horizontal axis represents the association strength (Log OR), while the vertical axis lists the keywords. Each point on the chart corresponds to a specific keyword and its Log OR value, indicating the frequency and importance of that keyword in L2 human-generated content compared to GPT-generated content. Keywords with more minus Log OR values are more strongly associated with L2 human texts, suggesting a higher frequency in human writing.

Figure 2 further illustrates the frequency and importance of certain keywords in the L2 human corpus relative to the GPT corpus. High association strength keywords include “food” and “career”, suggesting that these words are more common and frequently discussed in human-generated texts. These keywords reflect topics that are likely grounded in the real-life experiences or concerns of L2 writers. Keywords with medium association strength keywords such as “censorship” and “children”, while low association strength keywords like “leaders” and “youtube”, although more common in human texts, show less frequency difference compared with that of GPT-generated texts.

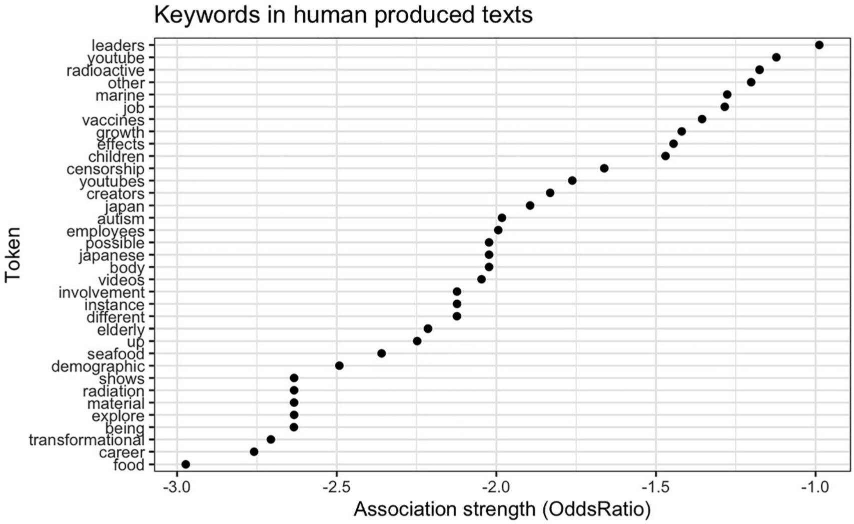

Figure 2: Keywords in human-produced texts.

Figure 3 describes the association strength between keywords and Log OR in the GPT corpus.

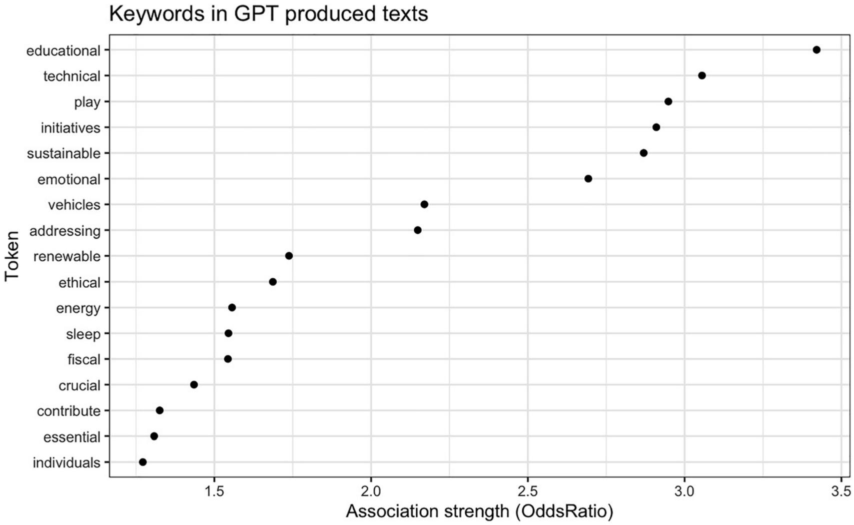

Figure 3: Keywords in GPT-produced texts.

Figure 3 displays keywords such as “educational” and “technical” have Log OR values exceeding 3.0, indicating that these tokens are used significantly more than in the GPT corpus compared to the human corpus. These two keywords represent the highest association strengths, highlighting ChatGPT’s focus on topics related to education, technology, and entertainment. Keywords with moderate association strengths include “play” and “initiatives”, suggesting a considerable amount of discussion on new projects or plans in ChatGPT texts. “individuals” and “essential” are examples of keywords with lower association strengths. While these tokens are still common in ChatGPT content, the frequency difference compared to human texts is less, and they do not stand out as much as the high-strength keywords. The Log OR values for the L2 human corpus are minus, with larger absolute values indicating higher association strength. Therefore, the keywords for the L2 human corpus are arranged from nearly −3.0, starting with “food” and “career”. Conversely, the Log OR values for the GPT corpus are positive, meaning that large values indicate higher association strength. Again, the variation in keyword association strengths highlights the thematic focus of each corpus. GPT-generated texts emphasise structured, formal discussions on education and technology, whereas L2 human writers tend to concentrate on topics grounded in authentic experiences and societal issues. This could be attributed to the limitations in L2 writers’ vocabulary and their preference for familiar and concrete topics, which contrast with ChatGPT’s ability to generate content on a broad range of subjects, including more abstract and technical themes. Understanding these distinctions provides a deeper comprehension of the characteristics of GPT-generated content and the focal points of L2 human-produced texts, offering valuable insights for L2 educators.

5 Discussion

This study investigated the lexical choices and collocations in academic essays produced by ChatGPT-3.5 and human L2 writers. Quantitative methods were employed, including detailed analysis of word frequency, collocations, keywords.

Several significant differences between the two groups have been identified. In terms of lexical choice, our study reveals significant differences in the thematic focuses and stylistic preferences between academic essays produced by ChatGPT-3.5 and those by human L2 writers. Both groups demonstrated a shared interest in health-related topics, which is consistent with prior findings suggesting that L2 writers often draw upon familiar and socially significant themes (Kormos 2012) which ChatGPT, when prompted, was able to replicate. However, despite writing on the same topics as human L2 writers, GPT’s lexis was more focused on broad, formal, and structured themes, tending to generate content centred on technical discussions particularly in the fields of education, technology, and sustainability. This finding supports the results from AlAfnan and Mohdzuki (2023), who highlighted GPT's strong performance in generating lexically precise academic texts with consistent use of technical vocabulary. In contrast, human L2 writers' essays were more concerned with personal and specific experiences and social issues, reflecting the writers' experiences and their relevance to academic pursuits. These results are consistent with findings from Ädel and Erman (2012), which show that L2 writers, despite having intermediate to advanced proficiency, tend to reply on simpler and more familiar lexical choices.

In terms of lexical diversity, our findings suggest that ChatGPT-3.5 produced texts with greater lexical diversity and formality compared to those produced by human L2 writers, which aligns with Yeadon et al. (2023) who noted that academic texts generated by GPT showed an increased diversity in generated content. This characteristic of GPT allows it to effectively generate sophisticated academic language; however, as noted in Lew's (2023) study, the language produced by GPT tends to lack emotional depth and personalisation. This limitation was evident in our analysis as well, where GPT-generated content demonstrated a polished but somewhat rigid style, in contrast to the more personalised and contextually nuanced content generated by human L2 writers. Yeadon et al. (2023) also noted that academic texts generated by GPT employ a broad range of vocabulary, aligning with Lew's (2023) hypothesis that ChatGPT can naturally mimic expert human academic writing. However, our study further suggests that even with the same topics and instructions, GPT's outputs demonstrate a certain degree of rigidity, irrespective of its powerful text generation capabilities and variability. This is in contrast to the lexical choices in L2 human-writing, aligning with the conclusions of Ädel and Erman (2012). Their analysis of high-frequency words in L2 texts showed that, despite the intermediate to advanced English proficiency levels (B2-C1) of the of the L2 corpus, these writers tend to overuse simple words or certain types of lexical bundles, unlike native speakers who utilise a broader range of lexical resources. This suggests that L2 writing is potentially more strongly influenced by various task-and context-related variables as compared with GPT's ability to cover a wide range of topics with ease.

In terms of collocation analysis, there were significant differences in how words were paired in essays from both L2 writers and ChatGPT. GPT content involved collocations on contemporary and forward-looking issues, such as mental health, sustainable development, and cognitive abilities. These findings are in line with Stokel-Walker (2022), who emphasised GPT's proficiency in generating content that reflects advanced natural language processing abilities and technical sophistication. Conversely, human L2 writers focused more on leadership, work-related issues, and environmental concerns, exploring themes directly linked to human experiences and societal challenges. These findings indicate that L2 writers, due to their limited vocabulary and fluency in a L2, typically rely on familiar and concrete topics (e.g. Hoang and Crosthwaite 2024). This reliance contrasts with ChatGPT's ability to generate content covering a broad range of subjects, including more abstract and technical themes. Therefore, while existing research suggests that versions of ChatGPT-3 and above perform well in drafting students-like essays (Herbold et al. 2023), the data in the present study still suggest significant differences in focus between human and GPT produced lexis. GPT-generated content is characterised by longer and more complex collocations more suitable for generating descriptive texts. In contrast, human L2 writers prefer shorter, simpler collocations that are more colloquial and accessible.

6 Conclusions

This study explores the lexical features in 30 academic essays on the same topics, created by ChatGPT-3.5 and L2 writers, focusing on word frequency, collocations, and keyword analysis. The results indicate significant differences in thematic focus and lexical choices between ChatGPT-3.5-generated and human-produced academic writing.

However, the study has limitations. Despite the significant patterns of variation revealed across the two corpora, the small size of the corpora could influence the significance of individual wordings discussed in the present study, which may be less significant in larger corpora. The study encompasses 30 different topics which may influence the selection of lexis and collocation strengths present across the two corpora as well as the selected themes arising. Future research could compare essays of similar topics so as to determine variation within the same thematic topics. A larger dataset could also provide more robust and comprehensive insights, reducing the potential impact of outliers or biases. Furthermore, human L2 writers come from diverse educational and cultural backgrounds, which influence their writing styles and vocabulary choices. The study may not have fully accounted for these individual differences, leading to potential biases in the results. Any biases or deficiencies in the data generated by ChatGPT-3.5 could affect the comparability between AI-generated and human-generated texts. For instance, half of the 30 essays generated by ChatGPT-3.5 did not meet the required 1,000-word which could impact the statistical derivation of keywords and collocation. The methods used in this research primarily focus on lexical features such as word frequency and collocations to compare and analyse these academic essays. Due to the lack of sentence-level comparison methods, the study may not fully assess the nuances of writing quality, coherence, and argumentation structure. These limitations need to be addressed in future research to provide a more detailed and comprehensive understanding of the capabilities and limitations of AI-generated academic content compared to human writing.

Limitations nonewithstanding, we have identified several gaps between academic writing generated by ChatGPT-3.5 and that produced by L2 writers. Our goal here is to suggest to L2 learners and practitioners that, as of now, GenAI tools like ChatGPT-3.5 still produces language quite different from that found in human L2 writing. The rigidity of GPT-generated academic lexis indicate that AI tools should complement rather than replace human input. Effective integration requires that educators and learners maintain a balance between using AI to enhance writing fluency and ensuring students continue to develop their own critical thinking, expressible abilities, and contextually appropriate writing skills. To maximise educational benefits, it is essential that both educators and learners adopt a cautious and critical approach when using ChatGPT, treating them as aids to supplement human effort rather than as autonomous writing solutions. L2 learners can benefit from the strengths of AI technology while still cultivating the depth and authenticity than their (very human) approach to writing provides.

Research ethics: The study completed an ethics waiver as no human subjects were involved.
Author contributions: The authors have accepted responsibility for the entire content of this manuscript and approved its submission. Zhang: data collection, analysis, drafting. Crosthwaite: conception, supervision, drafting.
Conflict of interests: None.
Research funding: None.
Data availability: Upon request.

Appendix 1: Topics for the essays in both corpora.

	Topics
1	*The release of nuclear wastewater into the ocean by Japan*
2	*Silver economics: challenges and opportunities in an aging global*
3	*Conditional positive regard: Is it a healthy way of parenting?*
4	*Navigating the boundaries: How YouTube censorship serves as guidance for creators*
5	*The relationship between leadership style and employee motivation*
6	*Vaccines do not cause autism*
7	*The problems of gestational surrogacy in the USA*
8	*Addressing The issues of intimate Partner Violence (IPV): How can we prevent IPV?*
9	*Art therapy should be promoted and implemented more in treating people's health*
10	*Reaching the ultimate child's well-being using authoritative parenting style*

(continued)

	Topics
11	*South Korean government should maintain fiscal austerity*
12	*Is the fast fashion industry worth keeping: The detrimental effects of fashionable trends in Indonesia*
13	*Technological resources in Norwegian primary schools: is the increasing presence of digital tools in education a positive development?*
14	*The impact of online dating on modern relationships*
15	*The importance of more renewable energy resources in Indonesia*
16	*How do celebrity endorsements influence people's purchasing decisions?*
17	*Schools and parents should manage and control high school students' smartphone usage during academic activities in both Australia and the U.S.*
18	*Childhood obesity: A comprehensive examination of causes, consequences, and interventions*
19	*Will AI be able to replace mental health professionals?*
20	*The sustainability of electric vehicles*
21	*Supporting students from Low socioeconomic backgrounds in Australia schools*
22	*What will happen if nuclear wastewater is discharged into the sea?*
23	*Balancing emotional intelligence and technical expertise in the modern workplace*
24	*Japanese release of Fukushima's nuclear contaminated water into Pacific Ocean*
25	*Possible implications of artificial intelligence in modern society*
26	*The influence of human activities on the greenhouse effect*
27	*Sleep deprivation's effect on adolescents' life*
28	*How much does lack of sleep affect academic performance in teenagers?*
29	*Artificial intelligence cannot completely replace human work*
30	*How to deal with vaccine hesitancy as a growing, global problem, focusing on the COVID-19 pandemic*

References

Ädel, Annelie & Britt Erman. 2012. Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes* 31(2). 81–92.

Al-Zahrani, Abdulrahman. 2023. The impact of generative AI tools on researchers and research: Implications for academia in higher education. *Innovations in Education & Teaching International*. 1–15. https://doi.org/10.1080/14703297.2023.2271445.

AlAfnan, Mohammad Awad & Siti Fatimah Mohdzuki. 2023. Do artificial intelligence chatbots have a writing style? An investigation into the stylistic features of ChatGPT-4. *Journal of Artificial Intelligence and Technology* 3(3). 85–94.

Allen, Todd & Atsushi Mizumoto. 2024. ChatGPT over my friends: Japanese English-as-a-foreign-language learners' preferences for editing and proofreading strategies. *RELC Journal*. 00336882241262533. https://doi.org/10.1177/00336882241262533.

Arnold, Taylor, Nicolas Ballier, Paula Lissón & Lauren Tilton. 2019. Beyond lexical frequencies: using R for text analysis in the digital humanities. *Language Resources and Evaluation* 53(4). 707–733.

Atlas, Stephen. 2023. *ChatGPT for higher education and professional development: A guide to conversational AI*. Kingston, RI, USA: University of Rhode Island. Available at: https://digitalcommons.uri.edu/cba_facpubs/548 (accessed 1 March 2023).

Banks, George, Hayley Woznyj, Ryan Wesslen & Roxanne Ross. 2018. A review of best practice recommendations for text analysis in R (and a user-friendly app). *Journal of Business and Psychology* 33(4). 445–459.

Barrett, Alex & Austin Pack. 2023. Not quite eye to AI: Student and teacher perspectives on the use of generative artificial intelligence in the writing process. *International Journal of Educational Technology in Higher Education* 20(1). 59.

Bašić, Željana, Ana Banovac, Ivana Kružić & Ivan Jerković. 2023. ChatGPT-3.5 as writing assistance in students' essays. *Humanities and Social Sciences Communications* 10(1). 1–5.

Chan, Cecilia K. Y. & Wenjie Hu. 2023. Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education* 20(1). 43–18.

Cooper, Grant. 2023. Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology* 32(3). 444–452.

Cotton, Debby, Peter Cotton & J. Rueben Shipway. 2024. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education & Teaching International* 61(2). 228–239.

Crawford, Joseph, Michael Cowling & Kelly-Ann Allen. 2023. Leadership is needed for ethical ChatGPT: Character, assessment, and learning using artificial intelligence (AI). *Journal of University Teaching and Learning Practice* 20(3). 1–19.

Crosthwaite, Peter & Vit Baisa. 2023. Generative AI and the end of corpus-assisted data-driven learning? Not so fast. *Applied Corpus Linguistics* 3(3). 100066.

Daudaravicius, Vidas. 2010. The influence of collocation segmentation and top 10 items to keyword assignment performance. *Computational Linguistics and Intelligent Text Processing*. 648–660. https://doi.org/10.1007/978-3-642-12116-6_55.

Demir, Cuneyt. 2017. Lexical collocations in English: A comparative study of native and non-native scholars of English. *The Journal of Language and Linguistic Studies* 13(1). 75–87.

English, Fenwick. 2006. Understanding leadership in education: Life writing and its possibilities. *Journal of Educational Administration & History* 38(2). 141–154.

Farrokhnia, Mohammadreza, Seyyed Kazem Banihashem, Omid Noroozi & Arjen Wals. 2024. A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education & Teaching International* 61(3). 460–474.

Guo, Biyang, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue & Yupeng Wu. 2023. How Close is ChatGPT to Human Experts? Comparison corpus, evaluation, and detection. *Arxiv Pre-print*. https://doi.org/10.48550/arxiv.2301.07597.

Herbold, Steffen, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva & Alexander Trautsch. 2023. A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports* 13(1). 18617.

Hoang, Hien & Peter Crosthwaite. 2024. A comparative analysis of multiword units in the reading and listening input of English textbooks. *System* 121. 103224.

Hoover, David. 2003. Frequent collocations and authorial style. *Literary and Linguistic Computing* 18(3). 261–286.

Hutson, Matthew. 2021. Who should stop unethical AI. *The New Yorker*. 15.

Jiang, Fen. & Ken Hyland. 2024. Does ChatGPT argue like students? Bundles in argumentative essays. *Applied Linguistics*. amae052. https://doi.org/10.1093/applin/amae052.

Kasneci, Enkelejda, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jurgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn & Gjergji Kasneci. 2023.

ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 103. 102274.

Khan, Mehreen, Muhammad Chaudhry, Muhammad Ahsan & Rameesha Ahmad. 2024. ChatGPT and the future of impact assessment. *Environmental Science & Policy* 157. https://doi.org/10.1016/j.envsci.2024.103779.

Kormos, Judit. 2012. The role of individual differences in L2 writing. *Journal of Second Language Writing* 21(4). 390–403.

Kumar, Lalit & Vivek Tyagi. 2024. Understanding the concepts of tools and techniques for data analysis using RStudio. In *Recent trends and future direction for data analytics*, 197–213. Hershey, Pennsylvania, USA: IGI Global.

Kung, Tiffany, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Victor Tseng, Rimel Aggabao, Giezel Diaz-Candido & James Maningo. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using Large Language Models. *PLoS Digital Health* 2(2). e0000198.

Lew, Robert. 2023. ChatGPT as a COBUILD lexicographer. *Humanities & Social Sciences Communications* 10(1). 704–710.

Li, Hanlin, Yu Wang, Siqi Luo & Cui Huang. 2024. The influence of GenAI on the effectiveness of argumentative writing in higher education: Evidence from a quasi-experimental study in China. *Journal of Asian Public Policy*. 1–26. https://doi.org/10.1080/17516234.2024.2363128.

Nation, Paul. 2013. *Learning vocabulary in another language*, 2nd edn. Cambridge: Cambridge University Press.

Rafique, Hina, Imran Nazeer & Jawaria Rehman. 2024. The impact of ChatGPT on language evolution: A linguistic analysis. *Journal of Education and Social Studies* 5(1). 56–68.

Ramesh, Aditya, Prafulla Dhariwal, Alex Nichol, Casey Chu & Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1(2). 3.

Schweinberger, Martin. 2023. Tree-based models in R. In *Lang. Technol. Data Anal. Lab.(LADAL)*. https://ladal.edu.au/tree.html#References (accessed 1 January 2025).

Shin, Dongkwang & Paul Nation. 2008. Beyond single words: The most frequent collocations in spoken English. *ELT Journal* 62(4). 339–348.

Steiss, Jacob, Tamara Tate, Steve Graham, Jazmin Cruz, Michael Hebert, Jiali Wang, Youngsun Moon, Waverly Tseng, Mark Warschauer & Carol Olson. 2024. Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction* 91. https://doi.org/10.1016/j.learninstruc.2024.101894.

Stokel-Walker, Chris. 2022. AI bot ChatGPT writes smart essays-should professors worry? *Nature*. https://doi.org/10.1038/d41586-022-04397-7.

Tlili, Ahmed, Boulous Shehata, Michael Adarkwah, Aras Bozkurt, Daniel Hickey, Ronghuai Huang & Brighter Agyemang. 2023. What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments* 10(1). 15–24.

Verzani, John. 2011. *Getting started with RStudio*, 1st edn. Sebastopol, California, USA: O'Reilly Media, Incorporated.

Wang, Xinghua, Linlin Li, Seng Chee Tan, Lu Yang & Jun Lei. 2023. Preparing for AI-enhanced education: Conceptualizing and empirically examining teachers' AI readiness. *Computers in Human Behavior* 146. 107798.

Wilson, Joshua, Matthew Myers & Andrew Potter. 2022. Investigating the promise of automated writing evaluation for supporting formative writing assessment at scale. *Assessment in Education: Principles, Policy & Practice* 29(2). 183–199.

Wu, Tianyu, Shizu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han & Yang Tang. 2023. A brief overview of ChatGPT: The history, status Quo and potential future development. *IEEE/CAA Journal of Automatica Sinica* 10(5). 1122–1136.

Xia, Detong, Yudi Chen & Hye Pae. 2023. Lexical and grammatical collocations in beginning and intermediate L2 argumentative essays: A bigram study. *International Review of Applied Linguistics in Language Teaching, IRAL* 61(4). 1421–1453.

Xia, Wei, Shaoguang Mao & Chanjing Zheng. 2024. Empirical study of large language models as automated essay scoring tools in English composition taking TOEFL independent writing task for example. *Arxiv Pre-print*. https://doi.org/10.48550/arxiv.2401.03401.

Yan, Da. 2023. Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Education and Information Technologies* 28(11). 13943–13967.

Yang, Hongzhi & Suna Kyun. 2022. The current research trend of artificial intelligence in language learning: A systematic empirical literature review from an activity theory perspective. *Australasian Journal of Educational Technology* 38(5). 180–210.

Yeadon, Will, Oto-Obong Inyang, Arin Mizouri, Alex Peach & Craig Testrow. 2023. The death of the short-form physics essay in the coming AI revolution. *Physics Education* 58(3). 35027.

Zhao, Xin. 2023. Leveraging artificial intelligence (AI) technology for English writing: Introducing wordtune as a digital writing assistant for EFL writers. *RELC Journal* 54(3). 890–894.

Zih, Hanane, Maha el Biadi & Zakirae Chatri. 2020. Evaluating the effectiveness of corpus linguistic software in analyzing the grammatical structure: LancsBox and AntConc as case studies. In *2020 6th IEEE congress on information science and technology (CiSt)*, 515–519, Agadir – Essaouira, Morocco.

Zou, Min & Liang Huang. 2023. The impact of ChatGPT on L2 writing and expected responses: Voice from doctoral students. *Education and Information Technologies* 29(11). 13201–13219.