A Corpus-Linguistics-Based Comparison of AI-Aided Writing and Students' Writing

Zheng Wang
Xiamen University Tan Kah Kee College, Xiamen, Fujian, China

Abstract: This study investigates the linguistic and stylistic differences between AI-aided writing and students' writing through a corpus-linguistics-based analysis. Two corpora were constructed, each consisting of 20 essays. The corpora were analyzed using AntConc 4.2.0 and compared against the Brown and Frown reference corpora to provide a benchmark for modern American English. Quantitative indicators such as type-token ratio (TTR), word length, sentence length, high-frequency words, and entropy were calculated to uncover distinctive linguistic features and patterns. The results reveal notable differences between the two corpora. AI-aided writing employs longer words compared to students' writing, indicating a more sophisticated vocabulary in AI-generated texts. In contrast, students' writing exhibits greater lexical variety and more syntactically flexible sentences. These findings suggest that AI-aided writing aligns more closely with the lexical sophistication of standard American English, as represented by the Brown and Frown corpora, while students' writing reflects a simpler and more conversational style. This study provides insights into the linguistic characteristics of AI-aided writing and its implications for education, language learning, and the evolving role of AI in writing practices.

Keywords: Corpus-Linguistics-Based; Comparison; AI-Aided Writing; Students' Writing

1. Introduction

1.1 The Application of Corpus Linguistics in Text Analysis

Corpus linguistics is a branch of linguistics that involves the study of language as expressed in samples (corpora) or "real world" text [1]. Corpus linguistics, which quantitatively describes language use is ultimately about “finding out about the nature and usage of language” [2]. This field utilizes large and structured sets of texts (corpora) to analyze lexical features, syntactical patterns, and sentence structures. By examining these corpora, linguists can gain insights into how language is used in various contexts, how it evolves over time. The primary goal of corpus linguistics is to provide empirical data to analyze linguistic features and text structures. This is achieved through the collection and analysis of large volumes of text, which can include written texts, or other forms of recorded documentation. These texts are then processed and analyzed using various corpus linguistics tools and techniques to identify patterns and trends [3,4].

Another key advantage of corpus linguistics is its reliance on authentic language data, as opposed to hypothetical examples. This allows for a more accurate understanding of language in use. Additionally, corpus linguistics can be applied to the analysis of a wide range of linguistic subfields, including syntax, semantics, pragmatics, sociolinguistics, and language acquisition.

1.2 The Development of AI-Aided Writing

Computer-aided writing, also known as artificial intelligence (AI)-aided writing, refers to the use of computer software and AI technologies to assist in the writing process. The development of AI-aided writing has seen significant advancements over the past few years, transforming how we approach content creation and editing. Computer-aided writing or AI-aided writing, based on massive datasets containing diverse text from various domains, such as open-access books, academic papers, news report, online forums, other publicly available text can even generate entire passages or articles based on given prompts, making them valuable for brainstorming and overcoming writer's block.

One of the key benefits of computer-aided writing is its ability to generate different styles of texts and enhance writing quality. Additionally, these tools can help students and those with limited writing skill generate and refine text more efficiently.

As AI technology continues to evolve, computer-aided writing tools are expected to become even more sophisticated, offering more personalized and context-aware assistance to writers. By automating routine tasks such as proofreading and formatting, writers can focus more on the creative and strategic aspects of their work. AI-aided writing is becoming increasingly popular and helpful, it is essential for us to identify the unique characteristics of Ai-aided writing and make better use of Ai-aided writing tool.

1.3 Comparison of AI-Aided Writing with Students' Writing

As it is mentioned above, the advent of AI-aided writing tools has transformed the way individuals approach the writing process, offering unprecedented support and efficiency. To understand the impact and nuances of AI-aided writing, it is essential to compare it with students' writing, particularly through the lens of corpus linguistics. Corpus linguistics, which involves the analysis of language data in large text corpora, provides a robust framework for examining the similarities and differences between AI-generated content and human-authored student writing.

By leveraging corpus linguistics, we can delve into aspects such as linguistic patterns, lexical features, stylistic choices, and error frequencies, offering a comprehensive understanding of how AI-aided writing aligns with or diverges from traditional student compositions. Therefore, it is necessary to use the tool of corpus linguistics to compare and find out how AI-aided writing aligns with or diverges from traditional student compositions.

2. Method

We collected essays about the following 4 topics, namely, "How to start a business successfully", "cultural differences of different countries", "what are essential factors for a meaningful life" and "how to avoid energy crisis?" To be exact, we collected 5 essays from AI-aided writing essays and 5 students' essays from each topic respectively.

Then the corpora were analyzed by Antconc 4.2.0, which provides a robust framework for examining the similarities and differences between AI-generated content and human-authored student writing.

They were subsequently compared against the better-known American-English Brown corpus and its 1990s updates Frown corpus. Totaling roughly one million words in collections of 500 text samples, these two reference corpora are popular among corpus linguistics researchers [5-7].

To empirically examine the similarities and differences between AI-generated content and human-authored student writing in detail. The main task was to calculate the quantitative indicators, such as types, tokens, type-token ratio (TTR), entropy [8-9], high-frequency words, small words proportion to reveal the linguistic features of AI-aided writing and students' writing.

3. Results

This section presents the basic information of our two corpora and the two reference ones and compares their lexical features and quantitative indicators.

3.1 Lexical Levels: Word Length and Small Word Proportion

Table 1. Basic Information about the Corpora

	Word Count	Average word length	language
AI Writing (AI)	16,313	5.99	Modern American English
Students' Writing (ST)	17,854	4.90	Modern American English

As Table 1 shows, the average word length of the AI is 5.99 letters while ST is 4.90. The difference in word length shows that AI-aided writing uses more sophisticated and complex vocabulary while words in students' writing are more compact and brief.

The proportion of small words (words consisting of no more than 5 letters) in the AI was calculated as roughly 69.9%, in contrast to the proportion of small words in ST 64.25%.

This indicates that small words are more prevalent in the ST. By contrast, in the Brown

Corpus, the proportion of small words was 69.9% while the proportion of small words in the Frown Corpus was 69.3%.

The much higher proportions of small words (64.25%) in students' writing indicate students' compositions, mostly composed of simple small words, are plain and easy to read. Quite obviously, students are just taught to learn to write well. It is understandable that compositions written by them are plain and simple.

In contrast, AI-aided writing (small words proportion of 46.99%,), with the help of diverse vocabulary database, can generate more sophisticated and complex words compared with students' writing. In addition, the proportion of small words of AI-aided writing are almost equivalent to Brown and Frown corpus, which means AI-aided writing fits the standard of modern American English.

Table 2. Lexical Features: Word Length and Sentence Length of AI and ST

	Brown	Frown	AI	ST
Tokens	1,191,332	1,241,887	1,934	2,599
Types	47,037	45,445	16,313	17,854
TTR	3.95	3.66	11.86	14.56
Ave. Word Length	4.47	4.39	5.99	4.90
Sentences	42,564	56,925	940	1160
Sd. Sent. Length	22.88	15,3	17.35	15.39
1-letter word	38,603	44,264	3	6
2-letter word	171,837	209,927	17	37
3-letter word	302,087	374,288	56	103
4-letter word	249,219	162,730	153	280
5-letter word	110,469	113,649	197	327
6-letter word	85,903	89,201	234	380
7-letter word	77,518	82,558	273	388
8-letter word	56,189	59,775	294	340
9-letter word	39,867	39,867	231	262
10+-letter word	26,897	26,897	476	476
Note: "Ave." is short for "average", "Sd." for "standardized", and "sent" for "sentence".				

Table 2 shows in detail the word and sentence lengths, TTR of our corpora, and the two reference corpora.

3.2 Quantitative Indicators: TTR and Sentence Length

As Table 2 shows, the type-token ratio (TTR) of the ST (14.56) far outnumbers that of AI-aided writing (11.86), revealing far greater lexical variation in the former than in AI-aided writing and indicating a better and diverse writing style [9-11]. Students' writing exhibits greater lexical variety, characterized by more diverse lexical types and fewer function words than AI-aided writing. The result shows that AI-aided writing write in a more mechanic way with more repetitive words and functional words.

Notably, in Table 2, the average sentence length of AI-aided writing is 17.35 longer than that of 15.39 in students' writing. This is attributed to the fact that the AI-aided writing tools or computer writing programs intend to use more syntactically complex and long sentences. In contrast, students, as beginners in language learning, tend to produce shorter and simple sentences. As a result, students' writing is syntactically flexible and brief.

3.3 Comparison of High-Frequency Words of AI-Aided Writing and Students' Writing

Tables 3 and 4 show the frequencies of the top ten words used in AI-aided writing and students' writing. In AI-aided wiring, *and, ranked as the top 1 most used word,* roughly doubles that of the words second most used, indicating AI writing tend to use more conjunction and function word to make the article more logic and coherent. Generally speaking, other top high-frequency words in AI-aided writing, such as "the", "of", "to", "in" and "is", most overlapped with those in students' writing, except for "your" and "can", which means that AI-writing tend to use more function words than students' writing.

On the other hand, *the,* ranked as the top 1 most used word in students writing, indicated the occasional misuse of article word "the" in students' writing. In other words, Chinese students tend to use article word *"the"* more often than it is grammatically needed.

Table 3. Top 10 High-Frequency Words of AI

Type	len	Rank	Freq	p(x)
and	3	1	1099	0.067369583
a	1	2	515	0.031569914
to	2	3	426	0.026114142
of	2	4	341	0.020903574
energy	6	5	326	0.019984062
the	3	6	314	0.019248452
in	2	7	297	0.018206339
can	3	8	246	0.015079998
is	2	9	210	0.012873169
your	4	10	200	0.012260161

Table 4. Top 10 High-Frequency Words of ST

Type	len	Rank	Freq	p(x)
the	3	1	1009	0.056513946
and	3	2	690	0.038646802
of	2	3	623	0.034894141
to	2	4	584	0.032709757
a	1	5	382	0.021395766
is	2	6	370	0.020723647
in	2	7	364	0.020387588
energy	6	8	323	0.018091184
life	4	9	188	0.010529853
it	2	10	171	0.009577686

3.4 Analysis of Entropy of AI-Aided Writing and Students' Writing

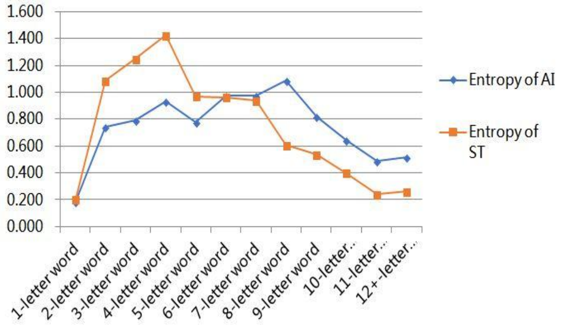

Figure 1. Entropy of AI and ST

To show in detail the difference in information density and dispersion (figure 1), we present how we calculated the "entropy" of in AI and ST. As a measure of information, entropy serves as an indicator of uncertainty, a measure of equilibrium or uniformity of language unit frequency distribution, and a measure of dispersion, etc. [9-11].

Overall, entropy of ST peaks in 4-letter word (1.400) while entropy of AI peaks in 8-letter word (1.110). This reveals the 4-letter words convey the most dense information in students' writing while the 8-letter words in AI-aided writing convey the most dense information, which also indicates that AI-aided writing uses longer and more complex words to express meaning than students' writing. In other words, textual analysis indicates that students' writing are mostly made up of 4-letter words while AI-aided writing is mostly composed 8-letter words. A highly possible explanation for the result would be that student are written in a conversational style, filled with simpler word and sentence.

From the analysis above, we can "reasonably define corpus linguistics as dealing with some set of machine-readable texts which is deemed an appropriate basis on which to study a specific set of research questions" [12].

4. Conclusion

In summary, corpus linguistics offers a wealth of linguistic insights through the systematic analysis of large datasets of student's writing and AI-aided writing.

The results of the study enabled us to arrive at the following conclusion: students' writing indicate students' compositions (small words proportion of 64.25%), mostly composed of simple small words, are plain and easy to read. In contrast, AI-aided writing (small words proportion of 46.99%,) can generate more sophisticated and complex words. The average word length of AI-aided writing is 5.99, longer than that of students wiring ST 4.90, which means AI-aided writing uses more sophisticated and complex vocabulary while words used in students' writing are more compact and brief.

In addition, the average sentence length of AI-aided writing is 17.35 longer than that of 15.39 in students' writing. This is attributed to the fact that the AI-aided writing tools or computer writing programs intend to use more syntactically complex and long sentences. As a result, our finding indicates students' writing is syntactically flexible and brief compared with AI-aided writing. The reason for this phenomenon is students, as beginners in language learning, tend to produce shorter and simple sentences and vocabulary to express their ideas. On the contrary, AI-aided writing, with the help of diverse vocabulary database, can generate longer, more sophisticated and complex words. A comparison of entropy of Students' writing and AI-aided writing further indicated small and simple words are more prevalent in students writing.

However, the type-token ratio (TTR) of students' writing (14.56) far outnumbers that of AI-aided writing (11.86), revealing far greater lexical variation in students' writing.

Last but not least, analysis of the top 10 high-frequencies of words used in AI-aided writing and students' writing shows AI writing tend to use more conjunction and function word to make the article more logic and coherent. The word *"the",* ranked as the top 1 most used word in students writing reveals the fact that Chinese students tend to use article word *"the"* more often than it is grammatically needed.

A corpus-linguistics-based comparison of AI-aided writing and students' writing helps reveal how AI-aided writing differs from students' writing in terms of word frequency, lexical diversity, sentence structure, and other linguistic features. This research has also implications for educators because such a study can highlight the strengths and weaknesses of students' writing compared to AI-aided writing.

Despite its merits, nevertheless and no doubt, there were confines that limited generalization of the results since the corpus was small and limited to only 40 essays. Therefore, it is recommended that the findings of this study should be considered as a starting point for further investigation

References

[1] Sinclair, J. (Ed.). (1991). Corpus, concordance, collocation. Oxford University Press.

[2] McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge University Press.

[3] Biber, D., Conrad, S., & Reppen, R. (1998). Corpus linguistics: Investigating language structure and use. Cambridge University Press.

[4] Hunston, S. (2006). Corpus linguistics. In K. Brown (Ed.), Encyclopedia of language & linguistics (2nd ed., pp. 234–248). Elsevier. https://doi.org/10.1016/b0-08-044854-2/00944-5

[5] Martinčić-Ipšić, S., Miličić, T., & Todorovski, L. (2019). The influence of feature representation of text on the performance of document classification. Applied Sciences, 9(4), 743. https://doi.org/10.3390/app9040743

[6] Mendhakar, A. (2022). Linguistic profiling of text genres: An exploration of fictional vs. non-fictional texts. Information, 13, 357. https://doi.org/10.3390/info13080357

[7] Fajri, M. S. A., & Okwar, V. (2020). Exploring a diachronic change in the use of English relative clauses: A corpus-based study and its implication for pedagogy. Sage Open, 10(4). https://doi.org/10.1177/2158244020975027

[8] Gries, S. T. (2010). Useful statistics for corpus linguistics. In A. Sanchez & M. Almela (Eds.), A mosaic of corpus linguistics: Selected approaches (pp. 269–291). Peter Lang.

[9] Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[10]Altmann, G., & Köhler, R. (2015). Forms and degrees of repetition in texts: Detection and analysis. Walter de Gruyter GmbH.

[11]Geluso, J., & Hirch, R. R. (2019). The reference corpus matters: Comparing the effect of different reference corpora on keyword analysis. Register Studies, 1(2), 209–242. https://doi.org/10.1075/rs.18001.gel

[12]Richards, B. (1987). Type/token ratios: What do they really tell us? Journal of Child Language, 14(2), 201–209. https://doi.org/10.1017/S0305000900012885.