Evaluating the Linguistic Range of ChatGPT: A Corpus Analysis

Mark Donnellan

Kindai University, Osaka, Japan
Email: donnellan@kindai.ac.jp

Abstract:

As authentic materials become increasingly valued in English language teaching, it is essential to evaluate the suitability of AI-generated content for pedagogical use. This study compares the linguistic range of ChatGPT-generated texts with a comparable corpus of authentic task-based EFL materials. Two corpora were constructed: one composed of human-authored listening texts, and another generated by ChatGPT using prompts based on the authentic scripts. Both corpora were analyzed using corpus tools to examine lexical variety, N-grams, and the most frequent words in each corpus. An initial comparison was also conducted using ChatGPT's self-analysis capabilities, followed by a conventional corpus analysis with established software. The results reveal that while ChatGPT produces grammatically accurate and topically relevant texts, its output lacks many features of authentic spoken language, such as informal discourse markers, spontaneous expressions, and lexical unpredictability. These findings highlight important limitations in the linguistic diversity and communicative realism of AI-generated content. Although ChatGPT may serve as a helpful resource for controlled practice and structured tasks, caution is advised when using its output as a substitute for natural language input. The study concludes with pedagogical recommendations for EFL educators and suggests directions for further research on refining AI use in language learning materials.

Keywords: ChatGPT, human-authored text, corpus linguistics, authentic language

1. Introduction

The emergence of AI-trained large language models (LLMs), such as ChatGPT (OpenAI, 2023), Claude (Anthropic, 2023), Gemini (Google DeepMind, 2023), and LLaMA (Meta, 2023), has sparked considerable discussion and debate within teaching and research communities. Although there are many LLMs available, each with its own strengths and limitations, this paper focuses on ChatGPT, a model developed by OpenAI. When prompted with the question "What is ChatGPT?", it describes itself as an LLM capable of understanding and generating natural language responses to a wide range of questions and tasks. According to OpenAI (2023), ChatGPT is designed to engage in dialogue, assist with problem-solving, and support users in academic, professional, and everyday contexts. This tool has divided opinions on the direction that teachers should take in relation to its usage with learners, with some teachers and researchers feeling that it would be better to limit or prohibit student usage and others feeling that we should embrace the technology. Within the discussion about the use of LLMs, one area of interest is the creation of pedagogical materials for English language learners. As educators explore the use of AI-generated texts, questions arise about how these compare to human-created content, particularly in terms of linguistic features. Corpus analysis provides a systematic method for evaluating ChatGPT's linguistic output against human-authored texts, enabling researchers to identify differences in lexical range, discourse features, and communicative authenticity that are critical in language learning contexts. This study reports on a small-scale corpus analysis of texts generated by

ChatGPT, comparing them to a comparable corpus of task-based EFL listening materials. The study is guided by two research questions:

1. What linguistic differences can be observed between ChatGPT-generated texts and comparable human-authored EFL listening materials?
2. What are the implications of these differences for language learning and teaching?

2. Literature Review

The rapid development of LLMs such as ChatGPT, Claude, Gemini, and LLaMA has prompted growing interest in their potential applications and limitations within education. While some educators express concern over issues such as academic integrity and student dependency (Cotton et al., 2023; Rudolph et al., 2023), others argue that such tools can support learning by offering personalized feedback, accessible practice opportunities, and even automated content generation (Kasneci et al., 2023; Qin & Wang, 2023). Among these educational applications, a promising area of exploration is the use of LLM-generated texts as instructional input for English language learners. However, questions remain regarding how such texts compare to human-authored materials in terms of linguistic complexity, diversity, and appropriateness for pedagogical use.

Several recent studies, such as those by Martínez et al. (2024) and Sandler et al. (2024), have assessed the linguistic characteristics of texts produced by LLMs, particularly ChatGPT. Martínez et al. (2024), for example, investigated the lexical diversity of ChatGPT outputs under different prompting conditions, finding that the model's lexical range varied depending on its assigned roles and system settings. Similarly, Sandler et al. (2024) compared thousands of ChatGPT-generated dialogues with human conversations using the EmpathicDialogues dataset, concluding that although ChatGPT often produces grammatically sound and topically relevant output, its language tends to be more homogenous and emotionally neutral than that of human speakers. These findings raise important considerations for EFL contexts, where exposure to naturalistic and varied language is crucial for learners' development of pragmatic competence and discourse awareness.

Beyond surface-level features such as vocabulary range, other studies have highlighted biases and stylistic constraints in ChatGPT's language. Fleisig et al. (2024), for instance, documented how the model demonstrates reduced fluency and increased stereotyping when interacting with non-standard English dialects, suggesting that its training data and optimization processes may favor dominant linguistic norms. From a pedagogical perspective, such tendencies could limit learners' exposure to authentic varieties of English and affect their attitudes toward global Englishes.

To analyze the language of ChatGPT rigorously, corpus linguistics offers a robust methodological foundation. Corpus linguistics is the systematic, empirical study of language through large, digitized collections of texts known as corpora. Using tools such as concordancers, frequency lists, and dispersion plots, researchers can examine patterns in lexis, syntax, and discourse across large datasets (Biber et al., 1998; McEnery & Hardie, 2012). In this context, a corpus-based comparison between ChatGPT-generated texts and human-authored EFL materials provides a transparent and replicable means of evaluating linguistic range and pedagogical suitability. Studies like Anthony (2023) and

Schoonjans (2023, 2024) have begun to explore the use of LLMs within corpus linguistics itself, suggesting both opportunities and challenges in integrating AI-generated data into traditional analytic frameworks.

Together, these strands of research underscore the need for empirical studies that compare AI-generated and human-created instructional texts, particularly in EFL settings. Understanding how LLM outputs differ from curated, task-based materials can help inform decisions about their classroom use and clarify the implications of relying on such models for content development.

3. Methods

For this study, two corpora were created: a corpus of authentic texts, and a corpus of texts created by ChatGPT. The former was composed of texts from task-based EFL materials which purported to be authentic. These texts were used with publisher permission. For ChatGPT to create comparable texts for the ChatGPT corpus, the researcher prompted ChatGPT with information about each of the authentic texts which ChatGPT responded to by creating texts. The prompting included specific guidance for ChatGPT regarding the length of the dialogue, which was matched to that of the corresponding human-authored text. Additional parameters included the target proficiency level (as specified in the textbook), the number of participants, and the contextual information necessary for generating a comparable dialogue. In some cases, follow-up prompts were provided to supply further context when the initial output was deemed too dissimilar from the original human-authored dialogue. A corpus analysis was carried out on the two corpora using Sketch Engine (Kilgarriff et al., 2014). The aim of this analysis was to elucidate the range and variety of language that ChatGPT could produce and how the language it produced compared to authentic language. The analysis focused on three aspects of the data: the number of unique words (type count), the most frequent words (keywords), and recurring multi-word sequences (n-grams). It should be noted that the small-scale nature of the data in the corpora and the fact that the sole focus is listening texts is a limitation. Nonetheless, the researcher feels that this approach can generate significant insights into how language produced by ChatGPT compares to that produced by humans in pedagogical contexts.

4. Results

The ChatGPT corpus contained a greater number of unique words, which was somewhat surprising. However, the human-authored corpus featured a significantly higher number of N-grams. This, the author argues, indicates greater authenticity—an important factor in language learning, particularly for developing listening skills. A more detailed explanation of these findings is provided in the following three sections.

4.1 Unique words (type count)

Table 1 presents the total number of words in each corpus, the number of unique words, and the ratio of unique words to total words. The human-authored corpus contained 18,539 words, of which 2,493 were unique, yielding a ratio of 0.1344. In contrast, the ChatGPT corpus contained fewer total words (15,776) but a higher number of unique words (3,153), resulting in a ratio of 0.1998. While this result

was somewhat unexpected, the data presented in the following sections offer more significant insights. The implications of these findings are discussed in Section 5.

Table 1. Number of Unique Words (Type Count)

	Words	Unique Words	Ratio
Human-created Corpus	18539	2493	0.1344
ChatGPT Corpus	15776	3153	0.1998

4.2 Keywords

Figure 1 and Figure 2 present the 50 most frequent words in each corpus. Some words appear frequently because they were central to the topic of the human-authored texts. For example, the word *frendshamen*—a non-English term—featured prominently due to its repeated use in a dialogue where two speakers discussed its meaning.

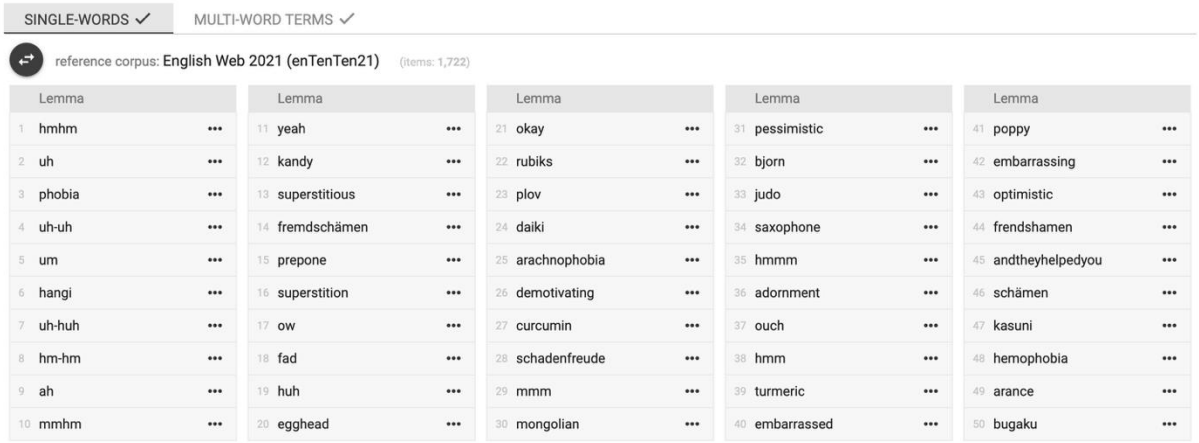
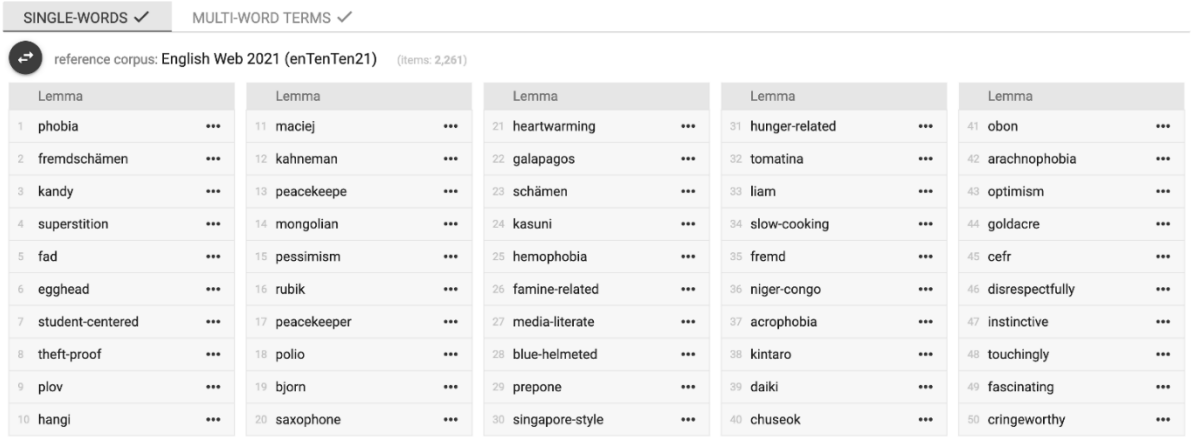


Figure 1. Keywords (Human Corpus)

Excluding such topic-specific items, the human-authored corpus provides clear evidence of the messiness and spontaneity characteristic of spoken language, including discourse markers, fillers, and informal lexis.



SINGLE-WORDS ✓ MULTI-WORD TERMS ✓

reference corpus: English Web 2021 (enTenTen21) (items: 2,261)

	Lemma		Lemma		Lemma		Lemma		Lemma
1	phobia	11	maciej	21	heartwarming	31	hunger-related	41	obon
2	fremdschämen	12	kahneman	22	galapagos	32	tomatina	42	arachnophobia
3	kandy	13	peacekeepe	23	schämen	33	liam	43	optimism
4	superstition	14	mongolian	24	kasuni	34	slow-cooking	44	goldacre
5	fad	15	pessimism	25	hemophobia	35	fremd	45	cefr
6	egghead	16	rubik	26	famine-related	36	niger-congo	46	disrespectfully
7	student-centered	17	peacekeeper	27	media-literate	37	acrophobia	47	instinctive
8	theft-proof	18	polio	28	blue-helmeted	38	kintaro	48	touchingly
9	plov	19	bjorn	29	prepone	39	daiki	49	fascinating
10	hangi	20	saxophone	30	singapore-style	40	chuseok	50	cringeworthy

Figure 2. Keywords (ChatGPT Corpus)

In contrast, the ChatGPT corpus contained more vocabulary typically associated with formal writing rather than informal speech, which suggested in a noticeable lack of authenticity.

The keyword analysis from the two corpora supported the argument that ChatGPT falls short in replicating the features of authentic texts. The implications of this finding are discussed in Section 5.

4.3 Recurring multi-word sequences (N-grams)

The analysis of 3-word and 4-word N-grams further highlighted the contrast between the two corpora and provided additional evidence of greater authenticity in the human-authored corpus.

	N-gram	Frequency	N-gram	Frequency	N-gram	Frequency	N-gram	Frequency
1	a lot of	25	14 the number of	8	27 in the future	7	40 would you like to	6
2	I do n't	17	15 you like to	8	28 you like to go	6	41 I thought it	6
3	around the world	14	16 do n't have	8	29 a big city	6	42 On the other hand	6
4	a little bit	12	17 what do you	8	30 often do you	6	43 it was n't	5
5	it was a	11	18 what kind of	8	31 I thought it was	6	44 part of the	5
6	like to go	11	19 in New Zealand	8	32 thought it was	6	45 like to go to	5
7	to go to	11	20 do n't know	8	33 there are many	6	46 a lot more	5
8	go to the	10	21 a long time	8	34 I want to	6	47 the United States	5
9	when I was	10	22 I did n't	7	35 was a little bit	6	48 that there are	5
10	in the world	9	23 I do n't know	7	36 would you like	6	49 and you can	5
11	one of the	8	24 the other hand	7	37 was a little	6	50 to be honest	5
12	I tried to	8	25 you want to	7	38 I went to	6		
13	do you think	8	26 you do n't	7	39 On the other	6		

Figure 3. N-grams (Human Corpus)

The human-authored corpus contained language chunks that are typical of spontaneous, authentic conversation. These included expressions such as *on the other hand*, *what kind of*, and *a little bit*. The corpus also featured a greater number of contractions, and the analysis identified a total of 50 N-grams.

	N-gram	Frequency			N-gram	Frequency	
1	Thank you for	19	•••	11	the part where	5	•••
2	Thank you for listening	13	•••	12	and I hope	5	•••
3	you for listening	13	•••	13	very scared of	5	•••
4	Have you ever	9	•••	14	you for your	5	•••
5	I hope you	8	•••	15	is n't it	5	•••
6	not just about	7	•••	16	Thanks for sharing	5	•••
7	It was a	7	•••	17	next time you	5	•••
8	we do n't	6	•••	18	some people are	5	•••
9	in the world	6	•••	19	Thank you for your	5	•••
10	part of the	6	•••	20	Is n't that	5	•••

Figure 4. N-grams (ChatGPT Corpus)

The ChatGPT corpus contained significantly fewer N-grams than the human-authored corpus, with only 20 identified. Moreover, these N-grams appeared less conversational, including phrases such as *thank you for listening*, which resembled the kind of language typically used in formal academic presentations rather than in casual dialogue.

The results presented in this section suggest that ChatGPT may fall short in replicating the linguistic features of human-authored texts. The implications and significance of these findings are discussed in the following section.

5. Discussion

The results of this study indicate that ChatGPT can produce linguistically rich and grammatically accurate texts, yet these often lack the qualities that make human speech both authentic and pedagogically valuable. The wider vocabulary and cleaner structure of ChatGPT-generated texts may be beneficial for controlled practice or reading comprehension tasks. However, for developing listening skills—especially at higher proficiency levels—exposure to the unpredictability and nuances of authentic spoken language is essential.

Educators using ChatGPT to generate materials should therefore consider the specific linguistic goals of their learners. If the objective is to expose learners to natural conversational features, unedited AI output may fall short. Conversely, if the goal is to provide clear and structured input, ChatGPT can serve as a useful tool. A balanced approach may involve combining AI-generated materials with human-authored texts or using AI output as a base for adaptation and refinement.

The findings suggest several practical implications for educators and materials developers. First, an awareness of the linguistic profile of AI-generated texts is essential. Teachers should evaluate whether

these texts align with the communicative needs of their students, particularly in listening and speaking contexts. Second, prompt design is crucial. This study employed basic prompting, with additional guidance provided in some cases where the output diverged significantly from the intended model. More nuanced prompting—potentially involving iterative refinement or stylistic constraints—may yield output that more closely resembles human-authored language. Third, further research is warranted. This pilot study relied on small corpora and a limited set of prompting strategies. Future studies could expand the scope to include other generative AI tools, additional genres (e.g., academic writing or classroom dialogue), and varied learner tasks (e.g., summarization or vocabulary acquisition). Moreover, the corpora used in this study were extremely small; larger datasets would allow for more robust and generalizable conclusions.

6. Conclusion

This study explored the linguistic differences between human-authored and ChatGPT-generated listening texts through a small-scale corpus comparison. The findings indicate that while ChatGPT offers lexical variety and structural clarity, it lacks many of the features that characterize authentic spoken language, including informality, spontaneity, and discourse-level nuance. These features are essential for learners aiming to develop listening skills in real-world contexts.

Teachers and materials developers should approach AI-generated content critically, balancing its strengths with the need to maintain exposure to realistic language use. With careful prompting and thoughtful integration, tools like ChatGPT can support language education, but they should not be viewed as a replacement for authentic, human-authored materials.

Declarations and Acknowledgement:

Early versions of this data were presented orally at EUROCALL 2023 and CamTESOL 2025.

References

Anthony, L. (2023). *Integrating large language models (LLMs) into a corpus analysis toolkit*. https://www.researchgate.net/publication/373819543

Anthropic. (2023). *Claude*. https://www.anthropic.com

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

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating? Exploring the potential of ChatGPT in higher education. *Innovations in Education and Teaching International*. https://doi.org/10.1080/14703297.2023.2190148

Fleisig, E., Smith, G., Bossi, M., Rustagi, I., Yin, X., & Klein, D. (2024). Linguistic bias in ChatGPT: Language models reinforce dialect discrimination. *arXiv preprint arXiv:2406.08818*. https://arxiv.org/abs/2406.08818

Google DeepMind. (2023). *Gemini*. https://deepmind.google/technologies/gemini/

Kasneci, E., Sessler, K., Betschart, S., Lampert, C. H., & Kaestner, C. A. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography, 1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Martínez, G., Hernández, J. A., Conde, J., Reviriego, P., & Merino-Gómez, E. (2024). Beware of words: Evaluating the lexical diversity of conversational llms using chatgpt as case study. *ACM Transactions on Intelligent Systems and Technology*. https://arxiv.org/abs/2402.15518

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

Meta. (2023). *LLaMA: Large Language Model Meta AI*. https://ai.meta.com/llama/

OpenAI. (2023). *ChatGPT*. https://openai.com/chatgpt

Qin, J., & Wang, W. (2023). Exploring the potential of ChatGPT for automated content generation in English language education. *Computers and Education: Artificial Intelligence, 4*, 100123. https://doi.org/10.1016/j.caeai.2023.100123

Rudolph, J., Tan, S., & Tan, C. (2023). A review of ChatGPT in education: Opportunities, challenges, and ethical considerations. *Education and Information Technologies, 28*, 5951–5976. https://doi.org/10.1007/s10639-023-11726-x

Sandler, M., Choung, H., Ross, A., & David, P. (2024). A linguistic comparison between human and ChatGPT-generated conversations. *arXiv preprint arXiv:2401.16587*. https://arxiv.org/abs/2401.16587

Schoonjans, S. (2023). ChatGPT: Friend or foe (to corpus linguists)? *ResearchGate*. https://www.researchgate.net/publication/372391900

Schoonjans, S. (2024). Analysing language data with ChatGPT-4 in corpus linguistics. *SSRN*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5126316