25 - Examining ChatGPT Models as L2 Academic Spoken English Dialogue Partners: A Corpus Linguistics Approach

Koki Sekitani
sekitani.koki@toyoeiwa.ac.jp
Toyo Eiwa University, Japan

Masaaki Ogura
mogura@omu.ac.jp
Osaka Metropolitan University, Japan

Takeshi Sato
satoken@people.kobe-u.ac.jp
Kobe University, Japan

Abstract

This study examines the use of ChatGPT, with a primary focus on its capacity to generate academic spoken English. While previous research on ChatGPT for L2 learning has primarily concentrated on its writing abilities, this study aims to assess the quality of conversational data generated by ChatGPT. To this end, we compared discourse data from the Michigan Corpus of Academic Spoken English with output produced by the ChatGPT-4o model. Specifically, we assessed language use in both the corpus and ChatGPT in academic discourse between a non-native student and professors, focusing on lexical diversity, lexical sophistication, syntactic complexity, and readability. The analyses revealed that ChatGPT-generated discourse exhibited a more diverse lexicon and longer clauses. Additionally, different tendencies were observed between the discourses of the student and the professors, such as contrasting results in readability. The findings offer novel insights into enhancing AI's interactive capabilities with L2 users by aligning them more closely with the dynamics of human spoken communication.

Keywords: ChatGPT in academic discourse, L2 learning and AI-generated language, Lexical and syntactic complexity, AI as a conversational partner, Comparative corpus analysis

1. Introduction

This study examines the authenticity of the academic spoken English discourse generated by ChatGPT by comparing it to that of an academic spoken English corpus. ChatGPT has been rapidly and widely adopted by L2 learners to enhance their proficiency and fluency. A number of studies have investigated the impact of ChatGPT on L2 learning with respect to writing (e.g., Yang & Li, 2024). For example, Su et al. (2023) show that ChatGPT provides valuable feedback at various stages of the L2 writing process such as brainstorming, revision, and proofreading. Mizumoto et al. (2023) also point out that writing assessment with ChatGPT demonstrates comparable reliability to that of humans.

Nevertheless, some studies have pointed out the possibility that ChatGPT could serve as an L2 oral interaction partner for learners. This is because it allows learners to speak with reduced pressure and anxiety (Javier & Moorhouse, 2023; Hayashi & Sato, 2024). They will not worry about their L2 errors or negative feedback from interlocutors, which would otherwise lead to poor L2 performance (Tsiplakides & Keramida, 2009). Since ChatGPT will instantly provide elaborate and contextually relevant response within a very short time, it will simulate dynamic interaction experiences such as

[25] To cite this proceeding paper: **Sekitani, K., Ogura, M. & Sato, T. (2024).** Examining ChatGPT models as L2 academic spoken English dialogue partners: a corpus linguistics approach. In D. K.-G. Chan et al. (Eds.), *Evolving trends in foreign language education: Past lessons, present reflections, future directions. Proceedings from the 10th CLaSIC 2024* (pp. 282–288). Centre for Language Studies, Faculty of Arts and Social Sciences, National University of Singapore. https://doi.org/10.5281/zenodo.14504937

role-playing in different settings (Javier & Moorhouse, 2023). Such real-time interactions will be carried out without time and location constraints and without the risk of negative reactions from interlocutors who joined it even at midnight. The environment will provide personalized learning opportunities (Shi, 2024), which will allow for increased exposure to the target language. Such experiences will help learners to improve their contextual L2 knowledge and skills, as well as facilitate the improvement in L2 listening and speaking proficiency (Al-Khasawneh, 2023). ChatGPT just implemented voice input, although this service is currently available for paid subscribers. However, in the near future it will become widespread and popular among L2 learners and teachers and will also have a significant impact on the development of L2 listening and speaking skills.

In consideration of the potential of ChatGPT for L2 oral interactions across a variety of settings, our study aims to investigate the extent the spoken language generated by ChatGPT aligns with authentic spoken language. This study focuses on academic spoken discourse, in which suitable interlocutors are often difficult to find. To compare the two discourses, data were extracted from the Michigan Corpus of Academic Spoken English (MICASE), which contains a variety of spoken language interactions. Thus, this study examines the lexical and syntactic complexity, readability, and content of the language output from MICASE and ChatGPT-4o, a widely used, paid model as of October 2024.

2. Method

2.1 Corpus Data

A skit titled "Social Psychology Dissertation Defence" in MICASE was selected for analysis. The skit consists of 12,000 words and lasts 75 minutes. It includes spoken interactions between a doctoral candidate and four professors serving as reviewers. They pose questions to the postgraduate student about the research, and the student responds to each one. The criteria for this selection were as follows: the main speaker (Ph.D. student) is a non-native speaker of English from an East Asian country, which is close to the authors' country; the discourse contains several authentic interactions between the candidate and professors; the dissertation defence represents one of the most challenging L2 activities for academic purposes.

2.2 Data Generation Process for ChatGPT-4o

We generated ten spoken skits using the same prompt, aligned with the settings of the corpus data. These settings included the number of characters, the linguistic background (one speaker is a non-native English speaker), the geographical background (a university in an Anglophone country), the academic field (social psychology), the situation (a Ph.D. defense with interactions between students and professors), and the number of words in the discourse. Since ChatGPT-4o could not generate the entire discourse at one time, we prompted it to generate the discourse step by step. To avoid potential copyright infringement, the original corpus data was not included in our prompts.

2.3 Indices for Analysis

2.3.1 Lexical diversity

We used *Text Inspector*, a web-based lexical data analysis tool, to conduct statistical analyses of the skit data. This tool provides indices for measuring lexical diversity.

2.3.2 Lexical sophistication

Text Inspector was also employed to analyze the proportion of CEFR-level words in each skit.

2.3.3 Syntactic complexity

In addition to lexical complexity, we analyzed syntactic complexity. First, we counted the total number of words, clauses, and Analysis of Speech Units (AS-units; Foster et al., 2000). The AS-unit is an augmented version of the T-unit, an independent clause, or a dependent clause connected to or embedded in an independent clause (Foster et al., 2000). Examples of one T-unit are: (a) "I like birds," (b) "I liked the movie we saw yesterday," and (c) "If it rains tomorrow, I will go to see a movie." The AS-unit builds on the T-unit, including independent phrases that do not contain verbs, such as (d) "At the museum." Examples (a) to (d) all contain one AS-unit, whereas (e) "I have a bird and its name is Pupu," contains two AS-units because it has two independent clauses connected by the coordinating conjunction "and."

Then, we calculated the ratios of words per clause, words per AS-unit, and clauses per AS-unit. Higher values for these measures indicate more complex utterances. These three metrics were used to assess syntactic complexity.

2.3.4 Readability

We analyzed the data with three types of readability indices: Flesch Reading Ease, Flesch-Kincaid Grade, and Gunning Fog Index. The Flesch Reading Ease (FRE) is a readability metric for English texts. This metric evaluates how easy or difficult a text is to read. This index considers two factors: the number of syllables per word and the number of words per sentence. The FRE score ranges from 100 to 0. Higher values indicate easier readability, while lower values suggest more difficult texts. The formula for calculating FRE is expressed as follows:

$$FRE = 206.835 - (1.015 \times ASL) - (84.6 \times ASW)$$

ASL is the average sentence length (the number of words per sentence), and ASW is the average number of syllables per word.

The Flesch Kincaid Grade is a readability metric designed to indicate the level of education needed to understand a given text. Developed in the U.S., the score corresponds to a U.S. school grade level, meaning a score of 8.0 suggests that an eighth-grade student should be able to understand the text. The formula for the Flesch-Kincaid Grade Level is:

$$Grade\ Level = (0.39 \times ASL) + (11.8 \times ASW) - 15.59$$

The Gunning Fog Index calculates the readability of English texts by estimating the number of years of education required to understand a text. The Gunning Fog Index is calculated using the following formula:

$$Gunning\ Fog\ Index = 0.4 \times (Average\ Sentence\ Length + Percentage\ of\ Words\ with\ Three\ or\ More\ Syllables)$$

3. Results

3.1 Quantitative Index Comparison Analysis

To compare the utterances in a doctoral dissertation defense in MICASE with those generated by ChatGPT-4o, we first calculated the means and standard deviations for ten ChatGPT-4o samples of the student's and professors' utterances separately. We then conducted one-sample *t*-tests on indices of lexical and syntactic complexity for each set of utterances. Tables 1 and 2 present the descriptive statistics and *t*-test results.

Table 1 - Comparison of Lexical and Syntactic Complexity in Student Utterances: MICASE vs. ChatGPT-4 in Doctoral Dissertation Defenses

Index	Measurement	MICASE	ChatGPT-4o (*n*=10)		Difference	*t*	*p*	*d*
			M	*SD*				
Lexical diversity	Token count	2184	2442.10	1089.23	258.10	0.75	.473	0.24
	Type count	552	705.90	215.33	153.90	2.26	.050	0.72

	Type/token ratio	0.25	0.31	0.07	0.06	3.00	.015	0.95
	VOCD	97.95	129.03	10.95	31.08	8.97	<.001	2.84
	MTLD	54.90	115.07	11.80	60.17	16.12	<.001	5.10
Lexical sophistication	A1 type %	33.76	24.80	3.53	−8.96	−8.02	<.001	−2.54
	A2 type %	14.23	14.29	1.20	0.06	0.17	.872	0.05
	B1 type %	16.06	19.23	2.12	3.17	4.72	.001	1.49
	B2 type %	11.31	18.84	2.04	7.53	11.66	<.001	3.69
	C1 type %	4.20	6.38	1.16	2.18	5.91	<.001	1.87
	C2 type %	2.19	3.29	1.00	1.10	3.50	.007	1.11
Syntactic complexity	Word count	3058	2410.30	1069.66	−647.70	−1.92	.088	−0.61
	Clause count	471	273.80	120.05	−197.20	−5.19	<.001	−1.64
	As-unit count	319	144.50	59.83	−174.50	−9.22	<.001	−2.92
	Words/clause ratio	6.49	8.81	0.39	2.32	18.77	<.001	5.94
	Words/AS-unit ratio	9.59	16.57	0.98	6.98	22.43	<.001	7.09
	Clauses/AS-unit ratio	1.48	1.88	0.10	0.40	12.87	<.001	4.07
Readability	Flesch Reading Ease	54.55	31.44	7.62	−23.11	−9.59	<.001	−3.03
	Flesch-Kincaid Grade	12.52	13.88	1.21	1.36	3.56	.006	1.12
	Gunning Fog Index	15.95	17.27	1.41	1.32	2.97	.016	0.94

Table 2 - Comparison of Lexical and Syntactic Complexity in Professors' Utterances: MICASE vs. ChatGPT-4 in Doctoral Dissertation Defences

Index	Measurement	MICASE	ChatGPT-4o (*n*=10)		Difference	*t*	*p*	*d*
			M	*SD*				
Lexical diversity	Token count	6283	1280.30	667.88	−5002.70	−23.69	<.001	−7.49
	Type count	1075	428.20	142.43	−646.80	−14.36	<.001	−4.54
	Type/token ratio	0.17	0.37	0.08	0.20	7.57	<.001	2.39
	VOCD	91.83	99.69	7.64	7.86	3.25	.010	1.03
	MTLD	50.42	94.57	6.28	44.15	22.24	<.001	7.03
Lexical sophistication	A1 type %	24.14	31.38	3.71	4.58	6.17	<.001	1.95
	A2 type %	17.15	15.11	1.32	−2.04	−4.87	<.001	−1.54
	B1 type %	17.97	17.06	1.67	−0.91	−1.73	.119	−0.55
	B2 type %	13.79	16.85	1.52	3.06	6.37	<.001	2.02
	C1 type %	3.63	5.77	0.95	2.14	7.12	<.001	2.25
	C2 type %	3.90	2.80	0.93	−1.10	−3.73	.005	−1.18
Syntactic complexity	Word count	8515	1279.60	675.55	7235.40	−33.87	<.001	−10.71
	Clause count	1303	176.50	92.86	−1126.50	−38.36	<.001	−12.13
	As-unit count	741	120.40	58.40	−620.60	−33.61	<.001	−10.63
	Words/clause ratio	6.53	7.22	0.42	0.69	5.19	<.001	1.64
	Words/AS-unit ratio	11.49	10.46	0.83	−1.03	−3.95	.003	−1.25
	Clauses/AS-unit ratio	1.76	1.45	0.07	−0.31	−14.69	<.001	−4.65
Readability	Flesch Reading Ease	48	51.74	3.81	3.74	3.11	.013	0.98
	Flesch-Kincaid Grade	16.97	9.14	0.65	−7.83	−37.92	<.001	−11.99
	Gunning Fog Index	20.51	12.49	0.92	−8.02	−27.66	<.001	−8.75

The results indicate that ChatGPT-4o generally demonstrated greater lexical diversity in student responses compared to MICASE, as shown through Type/Token Ratio, VOCD, and MTLD metrics. A comparison of vocabulary level ratios by CEFR revealed that ChatGPT-4o used a higher proportion of lower-frequency, advanced vocabulary. Regarding syntactic complexity, ChatGPT-4o incorporated longer clauses and AS-units, using AS-unit structures that contained a greater number of dependent

and embedded clauses. For readability, the findings consistently indicate that ChatGPT-4o's responses were more challenging to read compared to the Michigan Corpus. The Michigan Corpus scored higher on the Flesch Reading Ease, while ChatGPT-4o achieved higher values on the Flesch-Kincaid Grade and Gunning Fog Index, which is considered to reflect greater textual complexity.

For professor responses, ChatGPT-4o again displayed higher lexical diversity across Type/Token Ratio, VOCD, and MTLD metrics. The CEFR-based comparison of vocabulary levels revealed a complex pattern: ChatGPT-4o utilized a greater amount of A1, B2 and C1-level vocabulary, whereas MICASE included more A2 and C2 levels. At the C1 level, ChatGPT-4o surpassed the Michigan Corpus, though the pattern reversed at the C2 level, where the Michigan Corpus had a higher proportion. In terms of syntactic complexity, ChatGPT-4o outputs tended to produce longer clauses. However, AS-units in MICASE were generally longer and contained a higher number of dependent and embedded clauses. Readability scores show that ChatGPT-4o achieved higher Flesch Reading Ease scores, while the Michigan Corpus scored higher on the Flesch-Kincaid Grade and Gunning Fog Index, which indicates that utterances of the professors were more difficult to read than those generated by ChatGPT.

3.2 Observations on Content

The following are comparative content analyses of key characteristics between ChatGPT-4o's generation in L2 academic spoken English contexts and MICASE.

3.2.1 Linguistic patterns and syntax

ChatGPT-4o's responses demonstrated a high frequency of participle clauses with adverbial meaning compared to those in MICASE. This allows for a more formal and concise expression of ideas that aligns with academic standards. Additionally, ChatGPT-4o interactions were notable for their minimal use of fillers and interruptions. Unlike MICASE, ChatGPT-4o exhibited almost no rephrasing, repetition, or interruptions. The use of fillers such as "um" was minimal. This results in a smooth and uninterrupted conversational flow. Furthermore, ChatGPT-4o outputs were grammatically accurate, in contrast to the natural, often imperfect spoken language observed in MICASE.

3.2.2 Discourse themes and research focus

The themes in ChatGPT-4o interactions often centered on international comparative research, with the United States frequently serving as a point of comparison. The discussion generally followed a consistent structure. The main chair guided the exchange with questions about research objectives, definitions of key terms, comparisons with existing theories, methodological details (such as sample size and selection methods), practical implications, and research limitations. This structured format ensures thorough and balanced discussions across topics, unlike MICASE, where spontaneous topic shifted and varied discourse patterns were more common.

3.2.3 Citation and research focus

In contrast to MICASE, ChatGPT-4o interactions did not include explicit citations or references to prior research. The discussions tended to focus on empirical studies, with most studies adopting a mixed-methods approach that combined qualitative and quantitative research. Quantitative findings were rarely referenced in responses, while qualitative case examples were commonly used to illustrate key points. When quantitative analysis was mentioned, it was mainly on methodological aspects rather than specific numerical results.

3.2.4 Interactional dynamics and questioning style

ChatGPT-4o interactions featured consistently logical and structured student responses, with students demonstrating full comprehension of all questions posed. Unlike in MICASE, there were no misunderstandings or requests for clarification, even for indirect questions such as "You mean…?" or "You are saying that…?" Notably, negative comments or criticisms were absent, and the defense always concluded successfully, with unanimous pass judgments. Once a pass was granted, the defense typically ended with a congratulatory statement, "Congratulations, Dr. [name]," to which the student responded with gratitude, often saying, "Thank you so much!"

Moreover, turn-taking was highly organized; the chair skillfully moderated to ensure balanced participation. Professors did not interject in each other' s questions or add remarks, in contrast to the more fluid and overlapping exchanges found in MICASE.

3.2.5 Student recognition

When the student's gender was unknown, the pronoun "they" was consistently used, though this usage may seem somewhat out of place during the pass/fail deliberations among professors.

4. Discussion

The findings of the present study indicate both the strengths and limitations of AI language models in mimicking human academic discourse. While ChatGPT-4o produces advanced language with high lexical diversity and complex syntax, it might not fully capture the naturalness and interactive nature of human communication. Even so, ChatGPT can produce text of sufficient quality and serve as a convenient tool for learners to study academic English. Overall, since it tends to produce English that is more complex than students' utterances but less complex than professors', it can be considered useful as learning material. Given that ChatGPT has been demonstrated to be an effective tool for L2 learning as long as users are aware of its advantages and disadvantages (Chen, 2024), the findings of the present study are important in understanding its strengths and potential challenges as an interlocutor in academic spoken discourse.

This study has limitations in that a small sample size of ChatGPT-4o outputs and its focus on specific academic contexts may affect the transferability of the findings. Future research should explore larger datasets and a variety of academic settings to further investigate the capabilities and scope of AI language models.

Acknowledgements
This research was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (Grant Number [24K04155]).

References

Al-Khasawneh, F. M. (2023). From text to tech: Investigating chatbots' role in enhancing oral proficiency in second language learners. Journal of Southwest Jiaotong University, 58(5), 601–610. https://doi.org/10.35741/issn.0258-2724.58.5.45

Chen, X. (2024). The application of ChatGPT in second language learning classrooms: Opportunities and challenges. Transactions on Social Science, Education and Humanities Research, 5, 132–137. https://doi.org/10.62051/hg1w7578

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. Applied Linguistics, 21(3), 354–375. https://doi.org/10.1093/applin/21.3.354

Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology, 32(3), 221–233. https://doi.org/10.1037/h0057532

Hayashi, K., & Sato, T. (2024). The effectiveness of ChatGPT in enhancing English language proficiency and reducing second language anxiety. In WorldCALL 2023 Conference Proceedings (pp. 201–208). https://doi.org/10.22492/issn.2759-1182.2023.23

Javier, D. R. C., & Moorhouse, B. L. (2023). Developing secondary school English language learners' productive and critical use of ChatGPT. TESOL Journal, e755, 1–9. https://doi.org/10.1002/tesj.755

Mizumoto, A., Shintani, N., Sasaki, M., & Teng, M. F. (2024). Testing the viability of ChatGPT as a companion in L2 writing accuracy assessment. Research Methods in Applied Linguistics, 3(2), Article 100116. https://doi.org/10.1016/j.rmal.2024.100116

Simpson, R. C., Briggs, S. L., Ovens, J., & Swales, J. M. (2002). The Michigan Corpus of Academic Spoken English. The Regents of the University of Michigan.

Shi, X. (2024). Advantages, challenges, and prospects of ChatGPT in oral English teaching. Transactions on Social Science, Education and Humanities Research, 4, 99–109. https://doi.org/10.62051/c49b2t84

Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. Assessing Writing, 57, 100752. https://doi.org/10.1016/j.asw.2023.100752

Tsiplakides, I., & Keramida, A. (2009). Helping students overcome foreign language speaking anxiety in the English classroom: Theoretical issues and practical recommendations. International Education Studies, 2(4), 39–44. https://doi.org/10.5539/ies.v2n4p39

Weblingua Ltd. (n.d.). Text Inspector [Web tool for analyzing text]. https://www.textinspector.com

Yang, L., & Li, R. (2024). ChatGPT for L2 learning: Current status and implications. System, 124, Article 103351. https://doi.org/10.1016/j.system.2024.103351