American Journal of Qualitative Research
2024, Vol. 8 No. 2, pp. 153-160
https://doi.org/10.29333/ajqr/14487

AMERICAN JOURNAL OF QUALITATIVE RESEARCH
ISSN: 2576-2141

Using Chat GPT to Clean Qualitative Interview Transcriptions: A Usability and Feasibility Analysis

Zachary Taylor[1]
The University of Southern Mississippi, USA

ABSTRACT

One of the major inefficiencies in qualitative research is the accuracy and timeliness of transcribing audio files into analyzable text. However, researchers may now have the ability to leverage artificial intelligence to increase research efficiency through Chat GPT. As a result, this study performs feasibility and accuracy testing of Chat GPT versus human transcription to compare accuracy and timeliness. Results suggest that by using specific commands, Chat GPT can clean interview transcriptions in seconds with a <1% word error rate and near 0% syntactic error rate. Implications for research and ethics are addressed.

KEYWORDS: Chat GPT, artificial intelligence, qualitative research, educational research, transcription

To conduct qualitative research, educational researchers often need to transcribe audio files into analyzable text. Before automation, researchers listened to audio files on tape, pausing every few seconds, writing or typing the words verbatim, and repeating this process until the audio was transcribed. In the current age of automation, many transcription services have emerged, some free and some paid, through software vendors who can transcribe and clean audio files more efficiently than the researcher (Rev, 2023). These transcription services offer artificial intelligence (AI) audio transcription, increasing the efficiency of the transcription, including Microsoft Office's new AI transcription feature built into Microsoft Word in 2020 (Microsoft, 2023).

However, there are issues related to the reliance on AI-enabled transcription services and the accuracy of their audio transcriptions (Kvale, 2017; Wang et al., 2003; Xiong et al., 2018). For instance, word error rate (WER) is a measure of machine-transcribed text accuracy, originating in the field of natural language processing (NLP). WER measures the word-to-word differences between a reference or human transcription and a machine transcription, calculated by adding the number of word substitutions, word deletions, and word insertions and dividing that total by the number of overall words in a text (Wang et al., 2003; Xiong et al., 2018). User testing has consistently found that even advanced computational models for audio transcription are not as accurate as humans, as Microsoft's transcriptions have been found to produce a 16.5%-word error rate (WER), while Google Video and Rev transcription services only reduce WERs to 15.8% and 14.2% respectively (Rev, 2023).

[1] Corresponding Author: An Assistant Professor at the University of Southern Mississippi, USA.
E-Mail: z.w.taylor@usm.edu

Related to word error rate, machines also make syntactic errors while attempting to parse audio for transcriptions. Syntactic error rate (SER) is a similar calculation where machine-transcribed and human-transcribed phrases and sentences are compared to understand the accuracy between the two transcriptions. SER is calculated by dividing the overall number of syntactic errors in a text by the number of overall sentences in the text, with syntactic errors including misidentified phrases and clauses, missing words or extra words inserted into a single phrase, and incorrect syntactic placement, such as moving a prepositional phrase to the end of a clause when it appears at the beginning of the clause in the human transcription (Bock, 2011; Minkoff & Raney, 2000). Prior studies have found that machine transcriptions often commit syntactic errors when speakers within a transcription are learning a language or have difficulty speaking, commenting on the quality of the audio file necessary for machine transcription and the nature of machine transcriptions to fill the gaps in text when human beings may not speak smoothly or fluently (Bock, 2011; Minkoff & Raney, 2000).

Typically, it is assumed that if audio is clear enough, human beings make few errors (Kvale, 2017), while some researchers have viewed audio transcription as an art form, discouraging researchers from outsourcing audio transcription and potentially limiting the researcher's closeness and familiarity with the transcribed text (Hennessy et al., 2022). Additionally, prior studies of WER and SER have suggested that depending on the speaking ability of the interviewee within the audio file, it may be difficult for machine transcriptions to produce accurate text (Bock, 2011; Minkoff & Raney, 2000). As a result, in many instances, the use of AI-assisted voice recognition technology has resulted in the phenomena of "what is gained in speed is lost in accuracy" (MacLean et al., 2004, p. 114), with many researchers still preferring human transcriptions for qualitative research (Hennessy et al., 2022; Kvale, 2019).

As transformative as AI-assisted voice recognition may have been for qualitative researchers and their work, the advent and public launch of Chat GPT may push the boundaries of qualitative research efficiency even further. Hailed as "world-changing" artificial intelligence (OpenAI, 2022, para. 10), OpenAI launched Chat GPT in November 2022, immediately transforming how human beings do work. Since its launch in November 2022, Chat GPT has learned to conduct formerly human-intensive tasks such as writing computer code, composing music, writing essays, solving mathematical and chemistry problems, and a plethora of other AI-assisted activities (Agomuoh & Larsen, 2023). From here, it seems logical to attempt to extend Chat GPT's usage into the qualitative research space, specifically as an AI-assisted method of cleaning interview transcriptions.

As a result, this study performs usability tests of Chat GPT to measure its ability to clean interview transcriptions belonging to three different qualitative projects previously conducted by the researcher across several measures of text transcription accuracy. Given these aims, this study will answer the following questions:

RQ1: Compared to human transcription cleaning, how accurate is Chat GPT across word error rate (WER), syntactic error rate (SER), and punctuation error rate when cleaning interview transcriptions?

RQ2: Compared to human transcription cleaning, how fast is Chat GPT when cleaning interview transcriptions?

RQ3: Does altering the Chat GPT command influence the cleaning accuracy or speed of Chat GPT when cleaning interview transcriptions?

Methodology

This study is the first to evaluate Chat GPT's ability to clean interview transcriptions in the context of educational research. Therefore, this study's methodology draws from tenets of both qualitative research and computational linguistics to conduct a systematic evaluation. Then, limitations will be addressed in the Conclusion section of this study.

Research Design

This study employs a quantitative linguistic research design (Johnson, 2008), augmented by artificial intelligence in the form of Chat GPT. In quantitative linguistics, researchers seek to analyze text and transform that text into numbers, in effect quantifying the qualitative data (Johnson, 2008). For example, this study employs a common formula—counting every word in a document or word count—to quantify text to measure its length. In education and literacy studies, researchers expanded the field of quantitative linguistics to encompass readability formulas, transforming text into grade-level measures of the complexity of a text (Crossley et al., 2011). Over time, these readability measures have been validated across hundreds of studies in many different educational settings (Crossley et al., 2011). For instance, studies in higher education have applied a quantitative linguistics research design to explore the readability levels of admissions and financial aid information for prospective student audiences (Taylor, 2019a, 2019b). As a result, this study also engages with quantitative linguistic research design (Johnson, 2008) to quantify interview transcription texts and explore how Chat GPT transforms text through its text-cleaning processes.

Gaining IRB Approval and Gathering Data

First, institutional review board (IRB) approval was gained for three different qualitative research projects in 2021, 2022, and 2023, which have expiration dates in 2026, 2027, and 2028. After gaining IRB approval, a research team member conducted one-hour, semi-structured, one-on-one interviews with several different types of individuals about topics related to college student financial aid, public scholarship, and writing experiences in graduate programs. Interviews were conducted by six different researchers, demarcated by R1-R6 in Table 1. Participants across these three studies were unique individuals and were either all Ph.D. students or individuals who held Ph.D.s and worked full-time in higher education, either as faculty members or administrators.

Audio Transcription and Random Selection

Microsoft's free, AI-assisted audio transcription service was used to transcribe audio files, as the researcher's home institution required all research files to be stored on the Microsoft 365 OneDrive cloud storage system. Because both Microsoft Word and OneDrive are housed within Microsoft's 365 cloud system, using Microsoft's AI-assisted audio transcription service was both allowable by my institution and cost-effective. Then, the researcher randomly selected approximately 600-word excerpts from the different interviews, as the March 2023 version of Chat GPT allowed limited input and output by characters and words. The researcher was intentional that no random sample of text included any identifying information, protecting the identity of the participants of each study. This process resulted in 15 total texts that could be cleaned by Chat GPT.

Baseline Unclean Transcription Analysis

To analyze the unclean Microsoft Word transcriptions, the researcher used the Editor feature within Word to calculate word count, grade level readability, and words per sentence. It was important to gather these baseline measures to understand whether Chat GPT would change the word count, grade level readability, or words per sentence of each transcription, each of which could alter the cohesion and clarity of the transcription. Then, the researcher manually analyzed each transcription for word error rate (WER) and syntactic error rate (SER) by listening to the audio recording of the original interview and checking the Chat GPT-cleaned transcript word-by-word. WER is defined as the number of incorrectly transcribed words divided by the overall words in a transcription (Kvale, 2017; Rev, 2023). SER is defined as the number of complete sentences with a syntax error divided by the overall number of sentences in a transcription (Campbell et al., 2014). As a result, the researcher listened to the audio file and counted each word and each complete sentence. This provided a baseline word count and sentence count for subsequent WER and SER analyses. These baseline statistics are located in Table 1.

Table 1

Linguistic description of unclean educational interview transcriptions (N=15)

	Interview Topic	Word Count	Readability Level*	Words Per Sentence	Word Error Rate	Syntactic Error Rate
Text 1, R1	Writing	650	4.7	10.6	1.5%	21.7%
Text 2, R2	Writing	699	4.7	9.8	1.7%	31.8%
Text 3, R2	Writing	600	3.6	6.7	<1%	32.5%
Text 4, R2	Writing	590	5.6	14.2	<1%	26.8%
Text 5, R1	Writing	601	3.5	10.0	1.1%	20.3%
Text 6, R1	Public Scholarship	561	6.3	10.4	1.6%	17.6%
Text 7, R3	Public Scholarship	597	5.6	10.8	1.1%	7.2%
Text 8, R3	Public Scholarship	452	4.8	11.3	<1%	11.4%
Text 9, R4	Public Scholarship	381	6.1	13.8	2.1%	15.4%
Text 10, R4	Public Scholarship	586	7.0	15.0	2.7%	15.4%
Text 11, R5	Financial Aid	369	5.5	14.1	<1%	7.7%
Text 12, R5	Financial Aid	384	4.9	10.9	<1%	8.5%
Text 13, R5	Financial Aid	507	5.5	14.4	<1%	11.4%
Text 14, R6	Financial Aid	610	5.2	13.5	<1%	15.5%
Text 15, R6	Financial Aid	542	6.3	15	<1%	8.3%

*Note: Readability levels generated by the Flesch-Kincaid Grade Level Test; calculates a grade-level score of text by weighing average sentence length and average number of syllables per word.

Using Chat GPT to Clean Transcriptions

To clean the Microsoft Word transcriptions using Chat GPT, the researcher entered the command "Clean this transcription:" followed by the full text of the unclean transcription. In prior pilot studies, the researcher used various commands such as "Correct this transcription:" or "Fix errors in this transcription:" but the researcher found that "Clean this transcription:" was best at maintaining the same semantic and syntactic features of the spoken sentences of the interviewees, rather than synthesizing ideas and replacing semantic or syntactic features of the spoken sentences, resulting in inauthentic and inaccurate cleaned transcriptions. However, technology keeps advancing, and AI technologies such as Chat GPT will keep improving with human use, training, and fine-tuning. Moreover, these technologies do not disclose their algorithms or specifically how they use user-inputted text to train their machines and perform large language modeling. Given this issue, utilizing Chat GPT is a limitation in itself, as researchers do not have access to Chat GPT's code or algorithms, limiting how researchers can best understand the tool's capabilities and mechanisms.

Results

Chat GPT Cleaned Transcription Analysis

To analyze the Chat GPT-cleaned transcriptions, the researcher again used the Editor feature within Word to calculate word count, grade level readability, and words per sentence. Then, the researcher manually analyzed each transcription for word error rate (WER) and syntactic error rate (SER) by listening to the audio recording of the original interview and checking the Chat GPT-cleaned transcript word-by-word, repeating the pre-treatment (Chat GPT cleaning). This was performed by manually reading each transcript, listening to the audio, and checking the Chat GPT-cleaned file. These statistics are located in Table 2.

Table 2
Linguistic description of Chat GPT cleaned educational interview transcriptions (N=15)

	Interview Topic	Word Count	Readability Level*	Words Per Sentence	Word Error Rate	Syntactic Error Rate
Text 1, R1	Writing	614	7.8	16.1	<1%	5.3%
Text 2, R2	Writing	593	8.5	16.9	<1%	2.9%
Text 3, R2	Writing	506	6.6	9.7	0%	0%
Text 4, R2	Writing	521	7.3	17.3	<1%	3.3%
Text 5, R1	Writing	593	5.0	13.1	<1%	2.2%
Text 6, R1	Public Scholarship	542	6.4	10.2	<1%	3.8%
Text 7, R3	Public Scholarship	557	6.1	9.4	<1%	0%
Text 8, R3	Public Scholarship	145	7.8	16.1	0%	0%
Text 9, R4	Public Scholarship	312	7.3	16.8	<1%	5.9%
Text 10, R4	Public Scholarship	521	8.2	18.6	<1%	3.6%

Text 11, R5	Financial Aid	368	5.5	14.1	0%	0%
Text 12, R5	Financial Aid	380	5.1	10.8	<1%	0%
Text 13, R5	Financial Aid	499	5.3	13.8	0%	0%
Text 14, R6	Financial Aid	602	5.2	13.3	<1%	2.2%
Text 15, R6	Financial Aid	544	6.3	15.1	<1%	0%

*Note: Readability levels generated by the Flesch-Kincaid Grade Level Test; calculates a grade-level score of text by weighing average sentence length and average number of syllables per word.

Discussion, Implications, and Conclusion

Ultimately, evidenced by the baseline transcription data in Table 1 and the Chat GPT-cleaned transcription data in Table 2, this study finds that Chat GPT may present a considerable opportunity for qualitative researchers to make their data collection processes much more efficient. Moreover, ChatGPT did not make a single punctuation error, suggesting that this technology may automatically write grammatically correct sentences with accurate punctuation but may be limited in other ways. However, comparing unclean and Chat GPT-cleaned transcriptions, Chat GPT often made transcriptions shorter by-word count after cleaning redundant words and sentence fragments, but Chat GPT also made transcriptions more difficult to read by grade level by connecting sentence fragments, resulting in much longer words per sentence measurements. Additionally, Chat GPT was not able to clean transcriptions perfectly--there were still word errors and syntactic errors in several transcriptions, signaling that Chat GPT may not be a perfect replacement for human listening and transcription.

Moreover, this study suggests that speaker clarity of both interviewer and interviewee plays a role in Chat GPT transcription cleaning quality, as prior research has already demonstrated that speaker clarity affects transcription quality (Hennessy et al., 2022; Kvale, 2007; MacLean et al., 2004). For instance, although from an objective perspective, Researcher 5 (R5) from this study's texts 11 through 15 was a very clear speaker, and their transcriptions were noticeably cleaner than other researchers after only Microsoft's AI-assisted transcription. After the Chat GPT cleaning, Researcher 5's transcriptions were nearly flawless. From here, unclear speakers or speakers who often paused and added filler words (ex: ah, hmm, like, um, oh) presented additional challenges for Chat GPT's transcription cleaning ability. Therefore, as has been proven in the past, audio quality and speaker clarity are still highly valued elements of recording qualitative data, even if artificial intelligence can assist in cleaning audio transcripts.

However, this study was limited in several ways. First, as of March 2023, Chat GPT could not accept an input or create an output of more than 500-600 words or 2,000-2,500 characters. Although the researcher could not detect the specific limit of Chat GPT in providing output, I needed to make sure that the random selections of unclean transcriptions did not cut off mid-sentence and included both interviewer questions and interviewee responses. As a result, this study's sample size is relatively small, given the time-intensive nature of manual error checking for both unclean and Chat GPT-cleaned transcriptions. Additionally, this study only analyzed data from three different studies across three topics, six researchers, and fifteen interviewees. As a result, future studies should expand upon this one and further explore how Chat GPT can assist qualitative researchers in transcribing and cleaning data. Moreover, in the future, if Chat GPT allows audio file uploads, perhaps Chat GPT will outperform Microsoft's current free AI-assisted transcription service and make the qualitative data collection process even more efficient.

In addition to technical limitations, there are also considerable ethical concerns. For decades, qualitative research has been viewed and practiced as an intensely human-centered method, with human beings manually performing the work of the interviewer, including transcribing audio files and cleaning text transcriptions (Kvale, 2007). Data in this study suggests that machines—including AI tools and large language models such as Chat GPT—may be able to do the work that humans have done at nearly the same level but much more efficiently. As a result, a broader discussion in all fields utilizing qualitative research methods should begin with considering how artificial intelligence can be integrated into human-centric work, if it belongs at all. Some scholars in certain fields may consider machine-treated texts inauthentic or inappropriate for qualitative work, while others may welcome the monotony and tedium that Chat GPT eliminates when performing audio transcription and text cleaning tasks.

Ultimately, this study demonstrates that Chat GPT may represent an incredible efficiency for qualitative researchers who need to clean interview transcriptions. However, as this study suggests, human beings still have a place in qualitative data-cleaning processes, and prior research has questioned the morals and ethics of completely replacing human transcription (Hennessy et al., 2022). As a result, our collective future will continue to be shaped by artificial intelligence, and in the realm of research, Chat GPT may shape the work that qualitative researchers do, freeing them to do more pressing, more important, more humanistic work than cleaning audio transcriptions.

References

Agomuoh, F., & Larsen, L. (2023, September 28). What is ChatGPT? Here's how to use the AI chatbot everyone's talking about. *DigitalTrends*. https://www.digitaltrends.com/computing/how-to-use-openai-chatgpt-text-generation-chatbot/

Bock, K. (2011). "How much correction of syntactic errors are there, anyway?" *Language and Linguistics Compass, 5*(6), 322–335. https://doi.org/10.1111/j.1749-818X.2011.00283.x

Campbell, J.C., Hindle, A., & Amaral, J.N. (2014). Syntax errors just aren't natural: Improving error reporting with language models. In *Proceedings of the 11th Working Conference on Mining Software Repositories (MSR 2014).* Association for Computing Machinery, New York, NY, USA, 252–261. https://doi.org/10.1145/2597073.2597102

Crossley, S. A., Allen, D. B., & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language, 23*(1), 84–101. https://eric.ed.gov/?id=EJ926371

Hennessy, M., Dennehy, R., Doherty, J., & O'Donoghue, K. (2022). Outsourcing transcription: Extending ethical considerations in qualitative research. *Qualitative Health Research, 32*(7), 1197–1204. https://doi.org/10.1177/10497323221101709

Johnson, K. (2008). *Quantitative methods in linguistics*. Wiley-Blackwell.

Kvale, S. (2007). *Transcribing interviews: Doing interviews*. SAGE Publications. https://doi.org/10.4135/9781849208963

MacLean, L. M., Meyer, M., & Estable, A. (2004). Improving accuracy of transcripts in qualitative research. *Qualitative Health Research. 14*(1), 113–123. https://doi.org/10.1177/1049732303259804

Microsoft. (2023). Transcribe audio files. *Microsoft Office 365*. https://support.microsoft.com/en-us/office/transcribe-your-recordings-7fc2efec-245e-45f0-b053-2a97531ecf57

Minkoff, S. R. B., & Raney, G. E. (2000). Letter-detection errors in the word the: Word frequency versus syntactic structure. *Scientific Studies of Reading, 4*(1), 55–76. https://doi.org/10.1207/S1532799XSSR0401_5

OpenAI. (2022). Introducing: Chat GPT. *OpenAI*. https://openai.com/blog/chatgpt

Rev. (2023). *Rev automatic audio transcription*. Rev. https://www.rev.com/services/auto-audio-transcription

Taylor, Z. W. (2019a). Six easy steps: Do aspiring college students understand how to apply for financial aid? *Journal of Student Financial Aid*, *48*(3), 1–17. https://doi.org/10.55504/0884-9153.1643

Taylor, Z. W. (2019b). Writing dollars into sense: Simplifying financial aid for L2 students. *Journal of Student Affairs Research and Practice, 56*(4), 438–453. https://doi.org/10.1080/19496591.2019.1614937

Wang, Y.-Y., Acero, A., & Chelba, C. (2003). Is word error rate a good indicator for spoken language understanding accuracy. *2003 IEEE Workshop on Automatic Speech Recognition and Understanding*, 577–582. https://doi.org/10.1109/ASRU.2003.1318504

Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., & Stolcke, A. (2018). The Microsoft 2017 conversational speech recognition system. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, 5934–5938. https://doi.org/10.1109/ICASSP.2018.8461870

Notes on Contributor

Dr. Zachary Taylor is an assistant professor at the University of Southern Mississippi. Dr. Taylor has worked in education for 14 years as a pre-college counselor, financial aid consultant, assistant director of admissions, and admissions analyst, specifically aiming to serve low-income students, students with disabilities, and students of color.

ORCID

Zachary Taylor, https://orcid.org/0000-0002-6085-2729