Infant and Child Development

WILEY

METHODOLOGICAL ARTICLE OPEN ACCESS

A Script and Tutorial for Using Rev AI's Automatic Speech Transcription

Margaret Broeren | Yuzhe Gu | Mark Pitt | Virginia Tompkins

Department of Psychology, Ohio State University, Columbus, Ohio, USA

Correspondence: Margaret Broeren (broeren.1@osu.edu)

Received: 25 September 2024 | **Revised:** 11 March 2025 | **Accepted:** 13 March 2025

Funding: The authors received no specific funding for this work.

Keywords: automatic speech transcription | codes for the human analysis of transcripts | python | Rev AI | speech to text | tutorial | verbal response

ABSTRACT

We introduce Speech Transcriber with Rev AI (STR) - a Python script that allows for easy interfacing with the Rev AI speech transcription service. Recent advancements in technology have led to increased accuracy and affordability of automatic transcription services, making them preferable over the laborious and time-consuming process of manual transcription. STR allows users to take advantage of speech-to-text transcription services to transcribe their own verbal response data. STR is partially tailored to child development researchers utilising the Codes for the Human Analysis of Transcripts (CHAT) though the code is generic enough to output unformatted transcriptions. STR allows transcription of single words and multi-speaker dialogues in 50+ languages. We describe STR, provide a tutorial for CHAT-formatted transcriptions, describe settings available for customising transcription and conduct a brief analysis of the efficiency and accuracy of transcription. Speech that was transcribed in over half an hour by trained transcribers was transcribed in less than two minutes (with ~90% accuracy) by Rev AI. Considering the additional time needed for error correction and CHAT formatting, we estimate that manual transcription takes twice as long as transcribing with assistance from STR.

1 | Introduction

Language interaction is an inherently social endeavour, used to share perspectives, share our past with others, teach and learn, demonstrate knowledge and tell stories. Verbal communication research is diverse, with many different goals depending on the discipline, spanning psychology, linguistics, sociology, education, speech language pathology and others. However, what is common among these approaches is the need to transcribe, "the process of transforming audio or video files into typed text in order to subject the text to coding and further data analysis" (Katz-Buonincontro 2022). This is often the most time-consuming aspect of language research. We introduce a software tool that reduces this burden by taking advantage of the availability of high-quality transcription from online providers. The tool should have broad applicability across the discipline of psychology and beyond.

We begin by reviewing the challenges of using transcription in research, namely, its tedious and time-consuming nature. However, advancements in automatic speech recognition have increased the viability of using automatic speech recognition technology to assist researchers with transcription; it has the potential to greatly reduce the amount of time and human effort devoted to this task. A challenge in using machine transcribers is setting up the programming platform and writing the code to interface with a transcription engine. The first aim of this paper is to describe Speech Transcriber with Rev AI (STR), a Python script with a graphical user interface (GUI) meant to make the process of interfacing with a transcription service easier for researchers. It is designed

Summary

- We introduce a software tool that reduces this burden (of hand transcribing speech) by taking advantage of the availability of high-quality (machine) transcription from online providers.
- Our analysis suggests that STR saves hand transcribers hours of work at a reasonably high accuracy rate of over 90%.
- STR proves a reliable tool for researchers interested in capturing speech of children and adults alike.

for researchers who want to explore or use machine transcription but wish to invest minimal time in doing so. An overriding goal in development was ease of use. As such, the GUI is designed for users who prefer to have minimal contact with the command line, programming and editing text files.

The second aim is to provide a tutorial demonstrating the use of Speech Transcriber with Rev AI (STR). The script can generate output in three formats, including the Codes for the Human Analysis of Transcripts (CHAT), which are described later. The third aim of this paper is to compare traditional manual transcribing to the output of STR, demonstrating the actual (not just estimated) time savings of using this novel method of automated transcription. In addition, we demonstrate the accuracy of transcription with STR.

1.1 | Challenges for Researchers Who Use Transcription

The largest challenge for language interaction researchers is the need to transcribe speech to text for further analysis. In their seminal work, Hart and Risley (1995) painstakingly transcribed by hand (i.e., with pen and paper) one-hour parent-infant interactions within 42 families over many months; the number of observations per family ranged from 23 to 30, meaning that some families produced 2½ years of monthly family language data. They reported that no one on the team took a vacation day for more than 3 years while they collected parent–child speech data and transcribed it. They first transcribed from audio tape to handwritten text, then converted handwritten text to computer files. It took them 3 years to transcribe and another 3 to convert the transcriptions, therefore delaying the analysis of their longitudinal data by 6 years. Researchers quantifying the average length of time to transcribe audio data estimate that it takes from three to 12 h to transcribe 1 h of language data (Hart and Risley 1995; Heilmann 2010; Hepburn and Potter 2021; ten Haven 1999) with longer transcription times for child data (Hart and Risley 1995). In other words, it takes three to 12 times as long to transcribe as the length of the recorded audio and this time must be accounted for in project timelines (Bailey 2008; Hepburn and Potter 2021). Researchers themselves describe transcribing as "boring and incredibly tedious" (Hart and Risley 1995, 38), "mind-numbing grunt-work" (Sidnell 2010, 23), and something to dread (Katz-Buonincontro 2022).

Accuracy of transcribing speech is another challenge for language researchers. Researchers typically aim to transcribe speech verbatim (e.g., Lüdtke et al. 2023; Wardell et al. 2021); yet the ways that people naturally use language make this difficult—overlapping speech, false starts, dysfluencies, fillers (e.g., um, uh), nonword vocalisations (e.g., laughter), and language variations and dialects (e.g., Bailey 2008; Lüdtke et al. 2023; ten Haven 1999). This, again, adds to the time-consuming nature of transcribing because of the additional time taken to insert these language devices amid grammatically correct utterances, which are easier to remember as the transcriber types compared to disruptions in this flow. This often requires repeated playbacks of recordings (Bailey 2008; Katz-Buonincontro 2022; Sidnell 2010).

Researchers must also tailor their transcriptions to fit their specific analysis tools, often following the guidelines of widely used transcription formats. One such format is CHAT (MacWhinney 2000), which is used by researchers around the world to transcribe human speech as part of the TalkBank system (https://talkbank.org/). TalkBank is the world's largest open-access source for language data, including data from thousands of researchers in over 42 languages. Originally developed for child language (CHILDES), it now includes 14 research areas (e.g., ASD Bank for language in autism), all of which use the CHAT transcription format. This database is a valuable resource, and researchers are encouraged to design studies (e.g., consent forms) in a way that makes sharing data ethical and feasible. To make these data sharable across researchers and compatible with the software used to analyse these files (CLAN), CHAT requires specific transcription formatting (e.g., how to indicate speakers, use of utterance delimiters). This formatting adds additional time to the process of converting speech to text. However, automating speech to text could greatly reduce the overall time to transcribe human language interaction using CHAT.

1.2 | Automated Transcription

Verbal responses are easy to collect given the range of available recording software and hardware. However, analysis can be tedious when it relies on manual transcription, making the response mode less appealing than it otherwise would be. Although researchers have devised solutions to automate transcription, there are still limitations. For example, in general, the accuracy of automatic speech recognition is lower for children's speech compared to adult speech (Potamianos and Narayanan 2003; Shahin et al. 2020). There also continue to be difficulties with transcribing speech in less-than-ideal situations, such as audio with background noise, overlapping speakers, or speakers with accents or dialectical variations; speaker diarisation (i.e., separation) is also difficult (Moore 2015). Two examples of automatic transcription applications that have been investigated by language researchers are Attila and Dragon, each with their own limitations.

Moore provided reliability information using IBM's "Attila" speech recognition engine (Soltau et al. 2010). However, the audio files used were not research data, but professional speech, such as U.S. Congressional hearings. Even with professional speech, word error rates (WER) were high. For example, WER for auto-transcription of the 2013 Benghazi Hearing was 29% that is, 29% of words were transcribed incorrectly (e.g., "guys out for a walk one night who decided" was transcribed "guys out for a walk when I decide"). Another program, Dragon Speech Recognition—Get More Done by Voice | Nuance (2016) uses

voice recognition to train the transcription software based on participants' voices (Wardell et al. 2021). However, accuracy is lower when using recorded audio files rather than using Dragon in real time, it does not include filler words (e.g., "um"), and the program does not offer diarisation (Wardell et al. 2021). Additionally, the (Dragon Naturally Speaking Voice Recognition Software 2023) is single-user and costly (the current price for the PC version is $699 US). Although Wardell et al. (2021) provide a tutorial for using Dragon for autobiographical interviews with adults, they did not report the reliability of the automated transcribing compared to manual transcribing.

Recent advances in automatic speech recognition, most notably end-to-end modelling (O'Shaughnessy 2024; Prabhavalkar et al. 2023), have led to high-quality and robust machine transcription of speech that overall has an acceptable accuracy rate and is used with confidence in applications ranging from increasing accessibility to media (e.g., web content), to medical dictation, to customer service call centres. Commercial transcription services are widely available, and competition has driven down pricing to the point where it may be most economical for researchers to pay for transcription. These services are increasingly making some of their speech-to-text models freely available for those with the resources and skill to implement them locally (Bhandari et al. 2024; Radford et al. 2022).

1.3 | Solution

Although transcribing has come a long way since early studies of child speech (e.g., Brown 1973; Hart and Risley 1995), transcribing language data is still a laborious process requiring further technological advancements to increase its efficiency (Lüdtke et al. 2023). In this paper, we present a solution to transcribing efficiency that addresses the challenges outlined above—a solution that dramatically decreases the time to transcribe while maintaining a high level of accuracy. A good, automated transcribing solution should also be able to handle multiple speakers and multiple languages (Lüdtke et al. 2023). Additionally, it should be able to accurately transcribe child speech, which is more difficult to transcribe than adult speech (Lüdtke et al. 2023; Stoel-Gammon 2001; Yeung and Alwan 2018) and to transcribe speech in less-than-ideal situations (e.g., when there is background noise or linguistically diverse speakers; Bolden 2015). It should also maintain privacy and confidentiality according to institutional data policies (Katz-Buonincontro 2022).

The overarching aim of this paper is to make automatic transcription accessible to a broad audience. Verbal responding is a natural and informative mode of data collection in research in psychology and many related disciplines. The field of child language data collection is vast—encompassing topics such as parent- or teacher-child shared reading (e.g., Chen et al. 2024) and parent–child reminiscing (Russell et al. 2024). As with Hart and Risley's (1995) work, researchers often collect language data as part of naturalistic observations of children's everyday lives, including parent–child and sibling interactions. Researchers are often interested in child language samples to assess morphosyntax in determining language disorders or to assess narrative abilities. Across the lifespan, researchers are also interested in adult speech, such as the sharing of autobiographical narratives. The field of conversation analysis (e.g., Sacks et al. 1974) is interested in capturing not just what people say, but how they say it. Experimental methods of language often require the transcribing of single-word naming or reciting sentences. Counselling sessions, interviews and focus groups must also be transcribed for later review.

Again, what these diverse approaches have in common is the need to transcribe the spoken language data collected. The audio files collected are the raw data from which researchers create quantifiable data to be analysed; an accurate transcription provides "easy access to the data", which audio and video cannot (Hepburn and Potter 2021, 38). Whereas in the past, researchers may have loosely transcribed speech or taken impressionistic notes, it is common practice now that researchers record interactions and transcribe them verbatim to ensure the accuracy of our data, which serves as the basis of further analysis (Katz-Buonincontro 2022). These analyses are diverse depending on the aim of the study, but the need to transcribe remains the common thread. Given the time cost per participant, many studies are underpowered given the smaller sample sizes of such work; researchers must balance the value of longer speech samples or more participants with the feasibility of transcribing such data (Lüdtke et al. 2023). Studies are often underpowered also due to the volume of variables quantified from language data. Unlike standardised language tests that typically produce a single data point (e.g., a standard score), language researchers may extract dozens of variables from each speech sample. The research tools available determine what researchers consider feasible (Beckwith et al. 1985). Moore (2015) points out that automated transcription could help to increase researchers' sample sizes by making the long task of transcribing much more feasible. The solution presented in this paper addresses this need.

2 | Speech Transcriber With Rev AI

Speech Transcriber with Rev AI (STR) is a Python script that transcribes a folder of audio files using the asynchronous machine transcription service provided by Rev AI (https://www.rev.ai/). The STR script can transcribe audio files ranging from single words to hours-long conversations. For simplicity, the script assumes all audio files are in the same recording format (e.g., *wav*). The transcriptions can be output as running text (e.g., one sentence per line or transcriptions separated by talker) in addition to one word accompanied by a transcription confidence score. Users run the script from a graphical user interface (GUI). A command-line (GUI-less) version of the script is also provided.

Retrieve STR from the GitHub repository: https://github.com/YuzheGu/Speech-Transcriber-with-RevAI. The repository includes the script (*str.py*), the command line version of the script (str_nogui.py), a configuration file that stores the settings for customising transcription (*transcription_config.ini*), and folders of audio files for practising transcription. The *README.md* file describes software requirements, setup procedures, script usage and an example use case. A folder to hold the output generated by following the examples is also included.

STR works by first checking that the variable settings in the configuration file are valid. If the configuration check passes,

then the script checks that (1) the audio file format (extension) is supported, (2) every audio file within the input folder has the same extension and (3) every audio file is at least two seconds in duration (a requirement), elongating and concatenating (see below) audio files when necessary; a temporary folder is created for storing these temporary audio files. Next, the audio files are uploaded to Rev AI for transcription, and the transcriptions are returned and saved in the output folder. On-screen messages update the user on the status of transcription.

2.1 | Rev AI Transcription Engine

Rev AI is an online platform that provides speech-to-text services. Our script utilises Rev AI's asynchronous machine transcription service. We chose to use Rev AI because of its ease of use, versatility, data security, low cost and accuracy. Setup is fast and easy, requiring only the creation of an account and the generation of an API key to begin transcription. Speech in 58 languages can be transcribed, but some transcription variables are only available for a few languages. Bilingual transcription is not currently available. The Rev AI engine is most suitable for transcribing individual words, a monologue from one speaker, or a conversation among multiple speakers. Although Rev AI has the capability to transcribe hundreds of file formats, we have restricted our script to six common extensions (*wav*, *mp3*, *ogg*, *opus*, *flac* and *webm*); the code can be easily modified to accommodate others. Rev AI is capable of processing low-quality (8 kHz sampling) audio, but 16 kHz is the lowest sampling rate recommended for use. Of particular interest to some researchers is the ability to provide a custom vocabulary (up to 6000 English phrases consisting of up to 12 words) to increase recognition accuracy. This service may be useful to users who wish to transcribe audio with many nonwords, unknown proper nouns, or to increase transcription accuracy if results without it are unsatisfactory.

Data security and privacy are maintained throughout transcription. Audio is encrypted in transit and in storage on Rev AI servers. Users have complete control of their files at all times, and files can be deleted on the server immediately after transcription. By default, files are deleted after 30 days. Users may modify this through their Rev AI account. HIPAA-compliant accounts are available for those requiring additional privacy.

The asynchronous machine transcription API that we utilised for this paper, called "Reverb Transcription" is charged at a rate of $0.20/h for English audio and $0.30/h for other languages rounded to the nearest second. Audio files that are less than 15 s will be charged as if they were 15 s in duration. As of February 2025, newly created accounts are credited with five hour of free transcription (https://www.rev.ai/pricing), which many individuals should find adequate to assess performance.

3 | Usage

This section contains a step-by-step guide to using STR. We first describe the process of obtaining an API token necessary for transcription. Second, we describe how to set up the environment needed to run the script. Next, we provide a GUI tutorial via an example use case, followed by a description of software settings. We end with some tips for transcription. The GitHub repository contains additional documentation and video tutorials of script usage.

3.1 | Setup

Follow these steps to set up the working environment:

Obtain an API Token: First, create an account at https://www.rev.ai/auth/signup. Once logged in, navigate to the "Access Token" tab on the left-hand side of the webpage. Click "Generate New Access Token". This token is required for transcription, so ensure that this token is copied and saved locally.

Set Up the Environment: STR requires Python ≥v3.8 (Van Rossum and Drake 2009). There are many versions available. Rev AI requires installation of Python v2.7+ or Python v3.4+ to use the Rev AI Python Software Development Kit (SDK). We suggest installing the latest version of Python (currently version 3.13). Users must also install two dependencies using pip3: *pydub* and *rev_ai* (Robert and Webbie 2018; Vaswani 2022). After opening a command prompt, enter and run the following commands separately: 'pip3 install pydub' and 'pip3 install rev_ai'. *pydub* is used to manipulate the audio files. *rev_ai* is the API that communicates with the server. If a user needs to transcribe audio file formats such as *mp3*, *ogg*, *opus*, *flac* and *webm*, they must also install *FFmpeg* (see https://www.hostinger.com/tutorials/how-to-install-ffmpeg for tutorial on installation based on your operating system; Tomar 2006). *FFmpeg* allows the script to read audio file formats other than wav.

Instal STR: Retrieve STR from the Github repository (https://github.com/YuzheGu/Speech-Transcriber-with-RevAI). To download it, select the green "Code" button at the top of the page. In the dropdown menu, click on "Download ZIP". The repository will then be downloaded as a ZIP file to the user's computer. Extract the contents of the ZIP file into a project folder. Those who prefer using Git should clone the repository into a project folder.

Create input and output folders in the project folder. Place the desired audio files for transcription inside the input folder. The output folder will contain the transcription output files once they are generated.

3.2 | GUI Tutorial for CHAT Users

We will now illustrate how to transcribe an input folder, conversation_*example*, containing *conversation.wav*, an audio file containing a brief two-person conversation. This tutorial assumes users prefer a CHAT-formatted output text file. The example folder is included within the GitHub repository. We also assume Python is installed and the user is familiar with using the command line.

Open the command prompt and use 'cd' commands to navigate to the project folder.[1] Execute the following command:[2] 'python3 str.py'. The GUI window will appear.

Edit the configuration variables. Enter the API token generated in Step 1 of *Setup*. If you would like to save your API token on

the current device, check the box after "save API token".[3] The remaining variables are pre-configured to transcribe the audio files within the example folder named "conversation_example" and save the generated output files into the folder named "example_output" (see Figure 1). Choose "CHAT" to generate a CHAT format output file and press the button to confirm your choice. Further modification of the configuration variables is not needed to transcribe the example conversation.

Click "Save & Transcribe". Transcription will begin.[4] Reminder: all audio files of the same extension type saved in this folder will be transcribed. Remove unnecessary audio files from the folder before transcribing.

Transcription messages will appear below the "Save & Transcribe" button that will inform the user of what the program is currently doing and when the program has finished transcription. Users will see "Configuration check passed." if all GUI entries have valid values. If there are errors in the GUI entries, error messages will appear and instruct the user where the errors are. Once the errors are fixed, click "Save & Transcribe" again to transcribe the audio. Note that users will need to have a positive balance in their Rev AI account to complete transcription.[5] Once all audio files are transcribed, the user will see the message "All transcription is finished", and the transcribed file(s) will appear in the output folder.

Once the program has finished, two files will be saved to the "example_output" folder. The first file, "conversation_transcription_(*today's date*).cha", consists of plain text consistent with the format of CHAT transcription guidelines (see Figure 2.) The second file, "conversation_transcription_(*today's date*).csv", consists of four columns: *filename*, *transcription*, *confidence* and *speaker*, as shown in Figure 3. Each row contains one word of the conversation that was transcribed. The *transcription* column contains the transcribed words. *filename* contains the name of the folder as well as the name of the audio file that the transcribed word appeared in: "conversation_example/conversation.wav". *confidence* contains a score (0 to 1) that denotes the engine's confidence in the transcription. *speaker* contains an integer (1, 2 and so on), which indicates the identity of the speaker in each row.

3.3 | Other Use Cases

Users who need to customise the transcription variables to better fit their audio (i.e., users who need to transcribe multiple short audio files) will choose and confirm the "customise"

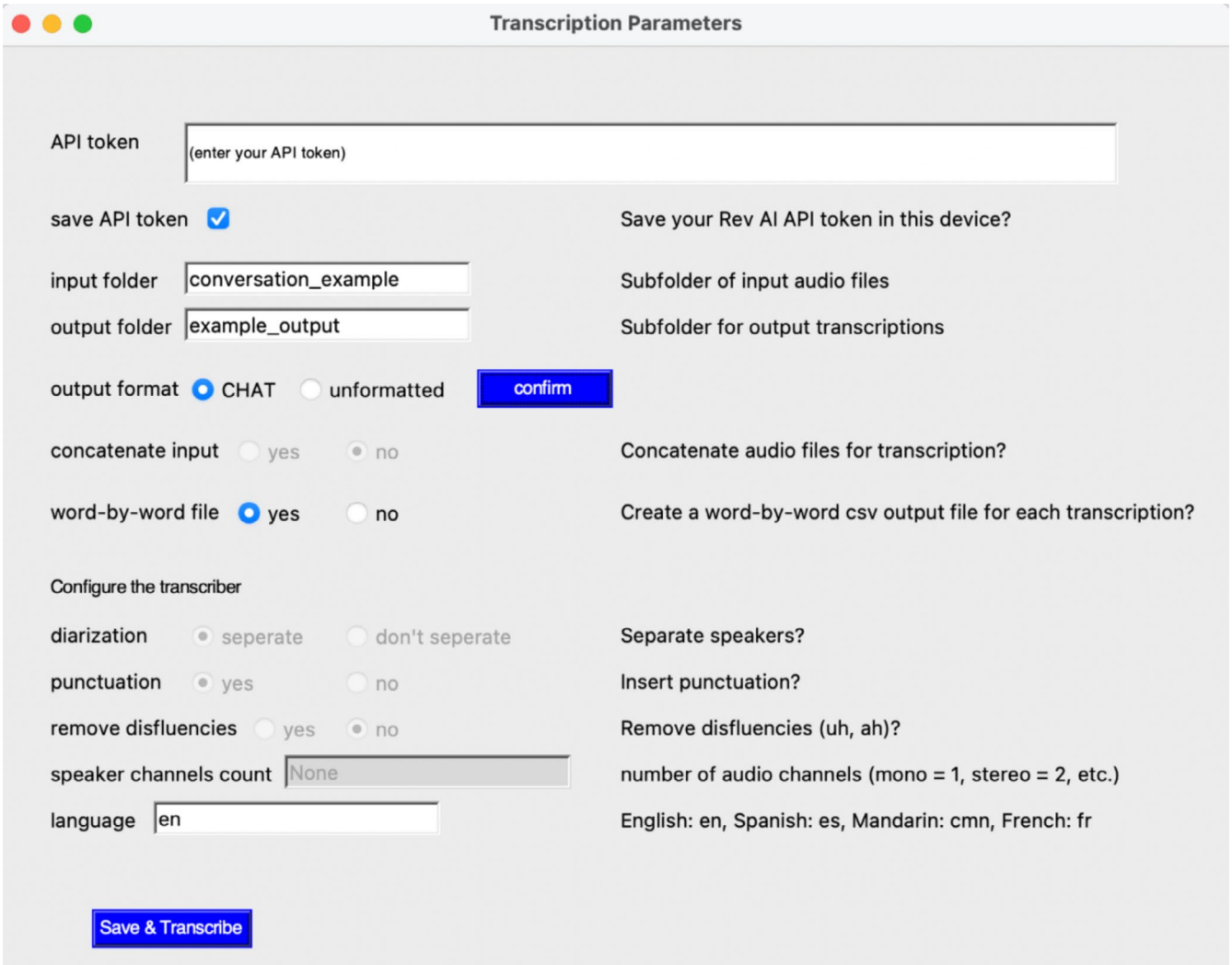

FIGURE 1 | Example GUI configuration.

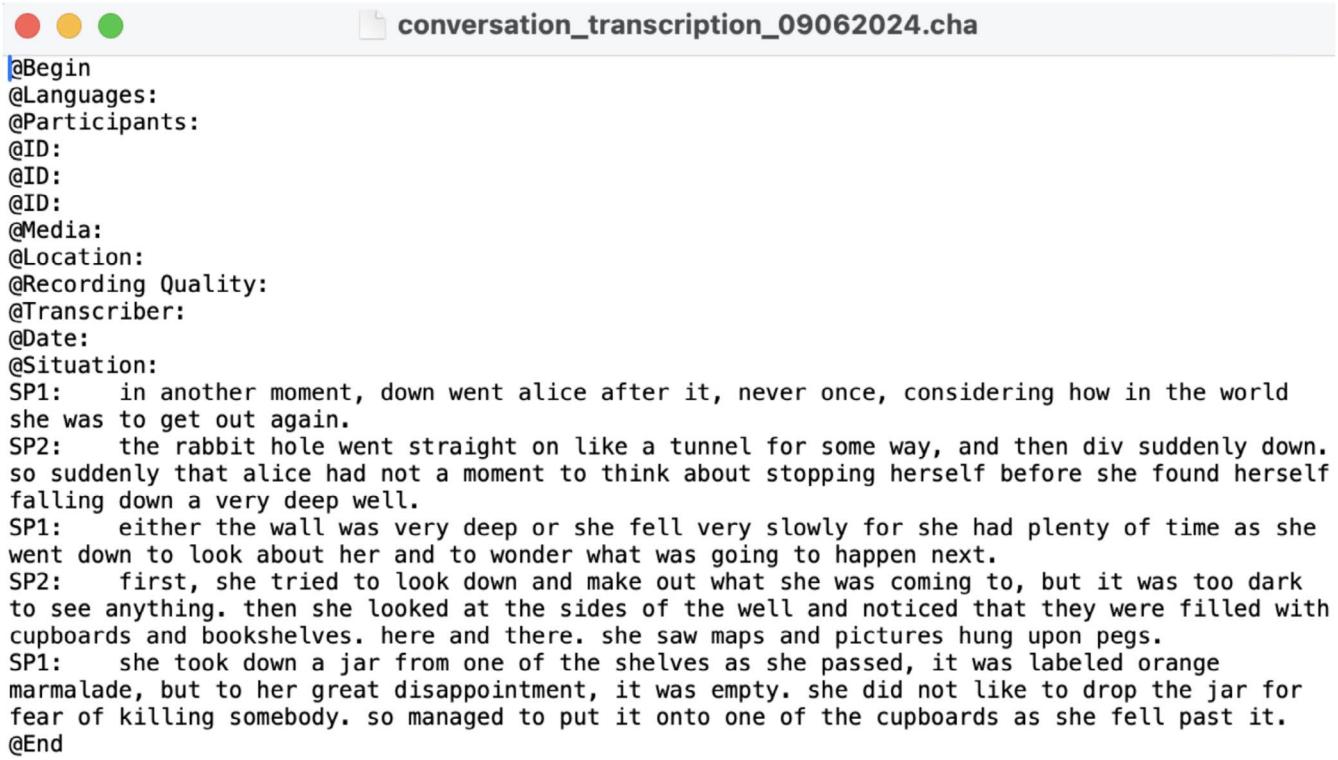

FIGURE 2 | cha Output for example conversation.

option in the GUI after inputting their API token and project folder names. Refer to the following lists of variable descriptions to know how to customise the transcription variables to best fit your audio:

3.4 | Transcription Variable Settings

3.4.1 | Optional Settings to Customise Transcription

- *concatenate_input—Default*: 'no'. Specifies whether audio files will be combined into one long audio file prior to transcription. Concatenation speeds up transcription and is less costly when individual sound files contain one or two spoken words. Specify 'yes' to concatenate your audio files. Specify 'no' to transcribe each audio file separately.
- *word-by-word file—Default*: 'yes'. Specify if a CSV file will be generated in addition to the text file. Specify 'yes' to have transcriptions also output in a CSV file. Specify 'no' to generate only text file(s). See README.md for details on output formats and examples.

3.4.2 | Transcription Settings

- *diarization—Set to 'separate' when the output format is specified as 'CHAT'. Default*: 'separate'. Specify whether transcription will be separated by speaker. Specify 'separate' if you would like the speakers to be identified (i.e., speaker 0, speaker 1, etc.). If speaker identity is not relevant, such as if the audio contains only one speaker, set to 'don't separate'. When set to 'don't separate', the transcription engine will not distinguish the speech among speakers. Note that speaker_channels_count must also be specified correctly for diarisation to succeed (see below).[6]
- *punctuation—Set to 'yes' when the output format is specified as CHAT. Default*: 'yes'. Specify whether the output will insert punctuation. Specify 'yes' if you would like the transcription to include punctuation; otherwise, specify 'no'. Only available for English.
- *remove_disfluencies—Set to 'no' when the output format is specified as 'CHAT'. Default*: 'no'. Specify whether the output will contain speech disfluencies that are recognised by the transcription engine ("um" and "uh"). Specify 'yes' if you would like your transcription to exclude disfluencies; otherwise, specify 'no'. Only available for English.
- *speaker_channels_count—Set to 'None' when the output format is specified as 'CHAT'. Default*: '1'. Specify the number of audio channels in the audio file. Mono = 1, Stereo = 2, (up to 8). Specify 'None' when there are multiple speakers in a single channel. The value of diarisation will be ignored if an integer value is provided here. Only available for English.
- *language—Default*: 'en'. Language of transcription. English = en, Spanish = es, Mandarin Chinese Simplified = cmn, French = fr. For a full list of languages available for transcription, see Rev AI supported languages: https://www.rev.ai/languages.

3.4.3 | Additional Transcription Variables

The following additional transcription variables are included as comments in str.py. Users may uncomment by removing the # symbol which precedes the needed lines and set them as needed:[7]

filename	transcription	confidence	speaker
conversation_example/conversation.wav	in	0.96	1
conversation_example/conversation.wav	another	0.99	1
conversation_example/conversation.wav	moment	0.99	1
conversation_example/conversation.wav	,	/	1
conversation_example/conversation.wav	down	0.98	1
conversation_example/conversation.wav	went	0.9	1
conversation_example/conversation.wav	alice	0.99	1
conversation_example/conversation.wav	after	0.99	1
conversation_example/conversation.wav	it	0.99	1
conversation_example/conversation.wav	,	/	1
conversation_example/conversation.wav	never	0.99	1
conversation_example/conversation.wav	once	0.96	1
conversation_example/conversation.wav	,	/	1
conversation_example/conversation.wav	considering	0.99	1
conversation_example/conversation.wav	how	0.99	1
conversation_example/conversation.wav	in	0.99	1
conversation_example/conversation.wav	the	0.99	1
conversation_example/conversation.wav	world	0.99	1
conversation_example/conversation.wav	she	0.99	1
conversation_example/conversation.wav	was	0.97	1
conversation_example/conversation.wav	to	0.99	1
conversation_example/conversation.wav	get	0.99	1
conversation_example/conversation.wav	out	0.99	1
conversation_example/conversation.wav	again	0.99	1
conversation_example/conversation.wav	.	/	1
conversation_example/conversation.wav	the	0.99	2
conversation_example/conversation.wav	rabbit	0.99	2
conversation_example/conversation.wav	hole	0.98	2
conversation_example/conversation.wav	went	0.99	2
conversation_example/conversation.wav	straight	0.99	2
conversation_example/conversation.wav	on	0.99	2
conversation_example/conversation.wav	like	0.99	2
conversation_example/conversation.wav	a	0.99	2
conversation_example/conversation.wav	tunnel	0.99	2
conversation_example/conversation.wav	for	0.99	2
conversation_example/conversation.wav	some	0.99	2

FIGURE 3 | csv Output for example conversation.

- *verbatim—Default*: 'True'. Specify 'True' if you would like to transcribe all utterance disfluencies (i.e., repetitions, stuttering, etc.), specify 'False'.
- *remove_atmospherics—Default*: 'False'. Specify 'True' if you would like to remove<laugh>, <affirmative>, etc.; otherwise, specify 'False'.
- *filter_profanity—Default*: 'False'. Specify 'True' if you would like to filter profanity; otherwise, specify 'False'.
- *custom_vocabularies—Default*: 'None'. Provide your own vocabulary to improve the transcription accuracy of these words (e.g., unique names, specific terminologies, etc.). Check custom vocabulary for details (https://docs.rev.ai/api/custom-vocabulary/#:~:text=Custom%20vocabularies%20are%20submitted%20as,to%201000%20for%20other%20languages).
- *diarization_type—Default*: 'standard'. Specify 'premium' if you would like to transcribe all utterances using the "premium" diarisation rather than the "standard".

3.5 | Best Practices and Tips

- Researchers should conduct a small test (a few short files or a snippet of a conversation) using their desired transcription configuration to ensure that the transcription is as expected before transcribing longer files.
- To maximise transcription accuracy, use only original recordings. Do not pre-process the audio before transcription (up sampling, filtering).
- If audio contains proper nouns, nonwords, or technical jargon that is unfamiliar to Rev AI's transcription

database (e.g., Rumpelstiltskin), it is in the researcher's best interest to utilise Rev AI's Custom Vocabulary API (https://docs.rev.ai/api/custom-vocabulary/). Users may submit up to 1000 phrases in the chosen language (up to 6000 English phrases) that are twelve words or fewer to the custom vocabulary API. By using this feature, you will add to the pool of words that Rev AI will pull from when transcribing your audio; therefore, increasing the chance that Rev AI will accurately identify custom vocabulary phrases.

- Conversations are transcribed verbatim, meaning revisions, repetitions and other disfluencies will be transcribed, which may not be desirable. Change the values of the *remove_disfluences* and *verbatim* variables to alter this behaviour.
- Confidence values are output alongside transcribed words within the CSV file. As described below, the authors have not found the confidence values to be a reliable indicators of transcription accuracy.
- *skip_punctuation*: Test this setting to ensure it performs as expected. Excessive and inaccurate punctuation may be added to output if set to FALSE. Only three punctuation labels are output by Rev AI: commas (,), periods (.) and question marks (?).
- Visit https://docs.rev.ai/api/asynchronous/best-practices/ to view Rev AI's official guidelines for best practices when using the asynchronous machine transcription service.

4 | Comparison of Manual Transcriptions With STR

Because of how laborious transcribing is, it is often completed by junior researchers with oversight by lead investigators (Bailey 2008). Our third aim was to compare the transcribing speed of junior researchers (undergraduate research assistants of the last author) to STR and using speech from a range of recording environments.

Transcription accuracy depends on the vocabulary and noise level in the recordings (Jette 2020). Although Rev AI reports an accuracy of 85.78% in their own broad testing, our limited testing with audio files from four labs found accuracy could be much higher, likely because of the recording quality (Jette 2020). A conversation from a counselling session using omni-directional standing microphones (took 2 min to transcribe 8731 words and cost $1.35[8]) yielded 97% accuracy. We also evaluated the transcription accuracy of recordings from a psychology experiment in German (took 2 min to transcribe 6600 words and cost $1.93[9]). The recordings consisted of many single sentences spoken by 14 participants. Transcription accuracy was 98.3%.

In a more challenging test, we evaluated the transcription of speech from a naming experiment conducted online (no control over microphone quality or the ambient recording environment) in which 82 English-speaking undergraduates repeated a monosyllable on each trial (11,800 recordings). Transcription accuracy was 85%. This is a harsh test of performance because of the lack of control of the recording setup and the lack of linguistic context (surrounding words) to assist in recognising less than 1 s of speech.

In a test of most relevance to the readership of this journal, two undergraduate students in the last author's child development laboratory who were already proficient in transcribing were selected to transcribe five audio files each in the traditional method. Audio files were of mothers and their preschool-aged children reading the book *Tell the Truth B. B. Wolf* (Sierra 2010) together. Importantly, students were blind to the purpose of their transcribing and were instructed to transcribe in a way that matched the STR output (e.g., transcribing was separated by speaker turn, not C-unit, as they were trained for other transcribing projects).

As shown in Table 1, audio files ranged from six to 11 min; students' time to transcribe the files ranged from about 30 min to over an hour and a half. Time taken to transcribe ranged from close to five to over 14 times the length of the audio file, with an average time of 7.7 times the length of the audio file. Although

TABLE 1 | Manual transcribing compared with speech transcriber with Rev AI (STR).

Transcriber and file number	Length of audio file	Manual transcription time	Transcription time/ audio length	Machine transcription time
1-File 1	0:06:01	0:36:00	5.98	0:01:22
1-File 2	0:11:09	1:12:00	5.40	0:01:21
1-File 3	0:06:17	0:40:00	6.37	0:01:11
1-File 4	0:06:25	0:32:00	4.99	0:01:27
1-File 5	0:08:47	0:43:00	4.90	0:01:35
2-File 6	0:08:31	1:33:00	10.92[a]	0:00:59
2-File 7	0:11:10	1:22:00	7.34	0:00:58
2-File 8	0:06:48	1:38:00	14.41[a]	0:01:04
2-File 9	0:08:50	1:36:00	10.87[a]	0:00:57
2-File 10	0:09:54	1:00:00	6.06	0:01:02

[a]Student transcribed across more than one session.

the file taking 14 times as long appears to be an outlier, overall, these times are consistent with estimates of prior researchers mentioned above (Hart and Risley 1995; Heilmann 2010; Hepburn and Potter 2021; ten Haven 1999). It should also be noted that the longer transcribing times were ones in which the student did not complete the transcribing at one time, which may affect transcribing speed. The table also demonstrates variability in student transcribers as one typically took longer than the other, though both had equivalent training and past experience in transcribing. In contrast, regardless of the audio length, Rev AI always transcribed these files in about one to one and a half min. Next, we examined the time required to "correct" the Rev AI files as compared to the transcriptions created by the student transcriber—specifically, we corrected words transcribed incorrectly, added speech that was missed by Rev AI, and corrected speaker changes. We focused on a long (11-min; File 2) and a short (6-min; File 3) audio file. We found that the time to make corrections was equivalent for both the long and the short audio file—about twice as long as the audio file itself (22 min for File 2 and 12 min for File 3). Thus, in direct comparison to the students' files of identical formatting but transcribed manually (72 and 42 min, respectively), this is a time saving of 50 min for the longer File 2 (69% faster) and a time saving of 18 min for the shorter File 3 (65% faster).

To determine the additional time needed to add the CHAT formatting to files in Table 1, the Rev AI-transcribed files of the same long (11-min; File 2) and short (6-min; File 3) audio files were edited to add formatting needed to pass Check in CHAT as well as standard formatting. Formatting included adding utterance delimiters, parsing turns into C-Units, noting repetitions and retracings, capitalising proper nouns and expanding word shortenings. The longer file added about 33 min of transcribing time; the shorter file added about 15 min. This time includes the time taken to make corrections described above, but in a separate pass of the Rev AI files; thus, it is not time *added* to making those corrections.

In both cases, the utilisation of Rev AI during transcription reduced the amount of time devoted to transcription by human researchers; manual transcribing took over twice as long as transcribing with Rev AI, which considers the additional time needed for error correction and CHAT formatting.

Next, student coders examined these files for accuracy of transcription. Students compared the manual transcriptions and the transcriptions generated by Rev AI word by word. They coded for accuracy in word transcription and speaker identification. Out of 11,509 total words transcribed across all 10 samples, approximately 90.34% of the words were correctly transcribed by Rev AI compared to the manual transcriptions. However, only about 4.63% of the total words were transcribed incorrectly due to mishearing (e.g., transcribing the word "gnome" as "no"). Approximately 1.75% of the total words were proper nouns that are not included in Rev AI's database of recognised words (i.e., author names and character names). In these cases, Rev AI would often transcribe the unknown proper noun as the most similar English word(s) in its database (i.e., "rumble stilt skin" for "Rumpelstiltskin"). These errors can be reduced through use of the custom vocabulary option. Other errors included inserting words which were not spoken in the audio, missing transcriptions of words that were spoken, and misidentification of homonyms.

Rev AI was fairly accurate at identifying the speakers within an audio file (approximately 95.25% of transcribed words). Speaker confusion would occur most often during child speech: Rev AI would miss child utterances altogether or mis-register a child utterance as the voice of the adult in the recordings.

In our tests of transcription accuracy, we found that transcription time varied little (range from 1 to 3 min) regardless of whether the sound files had a few hundred words (Table 1) or a few thousand (see above). Such trivial amounts of time mean that the real time–cost of using the transcription engine will be in post-processing transcriptions, such as correcting errors and adding CHAT formatting.

While the confidence score, which is output alongside each transcribed word within the CSV output file, sometimes points to inaccurate transcriptions, it does not reliably flag inaccurate transcriptions. In our testing, errors were sometimes found when a word had a corresponding confidence score of 0.94 (on a scale of 0 to 1) and lower. However, there were still many instances where words with a low confidence score were transcribed correctly. Additionally, we also encountered some transcription errors in which the corresponding confidence score was greater than 0.94. Ultimately, the authors cannot recommend the use of the confidence score to gauge the accuracy of transcriptions.

There is no way to estimate how many errors will be present within any given transcription. Accuracy depends on many factors, such as audio quality, sound file format, vocabulary and frequency of unfamiliar words in the audio. For example, in the parent–child shared reading described earlier (Tell the Truth, B. B. Wolf), Rev AI consistently transcribed unique character names incorrectly (e.g., Rumpelstiltskin, Miss Wonderly). Some users may find Rev AI's transcription accuracy sufficient for use in their research without any additional revisions, such as when detecting the presence of a set of key words. However, other researchers will likely need to review the outputs generated by this script before using the transcriptions in further analysis.

In either case, it is important to keep in mind that transcription engines are tools that are meant to ease the process of transcribing speech, not replace human transcribers entirely. The recommended use-case for this tool is to have a human transcriber review the output for accuracy and formatting errors before further analysis of the transcription is performed; for child data, particularly data with two interlocutors, we found that ensuring the accuracy of the spoken word and switches between speakers was mandatory. However, our script, and automatic transcription in general, makes it so that human transcribers can reduce the amount of effort put towards ~90% of the audio sample while saving the most effortful work on that < 10% of the audio that the automatic transcription service mis-transcribed.

5 | Discussion

The first aim of this paper was to describe (STR) and provide data demonstrating its accuracy and cost. The accuracy of speech

transcription (over 90%) was impressive compared to previous work that has considered error rates of 19%–29% to be "adequate" (Moore 2015). The second aim was to provide a tutorial of Speech Transcriber with Rev AI and guide potential users through the transcription process. The script's GUI provides users with an easy method of choosing transcription and output settings. For further ease of use, the script will automatically fill in certain transcription settings for those who indicate outputting CHAT formatted transcriptions.

This paper's third aim was to compare traditional manual transcription to the transcription output provided through STR on accuracy and time-saving potential. Our analysis suggests that STR saves manual transcribers hours of work at a reasonably high accuracy rate of over 90%. STR proves to be a reliable tool for researchers interested in capturing the speech of children and adults alike.

To provide a real-world implication of this tool, the last author (Tompkins 2019/2020) collected data on parent–child shared reading (similar to the book used for aim 3) to examine how shared reading changes with repeated readings and its relation to children's outcomes (e.g., social understanding). Within the study, 109 parents read a book with their 4–5-year-old child. Most parents read the book a second and third time. In total, this data set contains 45 h of recorded readings. In the current paper, we estimated that manual transcribing takes about twice as long as using Rev AI transcription services plus human correction and CHAT formatting. Since this study contains similar data to the ten transcriptions analysed earlier, we can use the average transcription time/audio length ratio (~7.7) to estimate that it will take 346.5 h to manually transcribe this dataset. If we assume that the use of Rev AI to assist in transcribing cuts the total time of transcription by half, the estimated time savings of using Rev AI in this study is approximately 173 h. Given that audio transcription is often completed by undergraduates with limited hours in the laboratory, this time savings could mean that investigators complete projects more efficiently or collect more robust samples.

6 | Conclusion

Although our test cases focused on parent–child shared reading, Rev AI has broad applicability regardless of the academic discipline, context of the language interaction, or behaviours under investigation. Speech-language pathologists might use Rev AI for initial transcriptions of language samples used in the diagnosis of language disorders; this could be particularly convenient in school settings. Conversation analysis may be used as a first pass on transcribing before adapting to Jeffersonian format. Similar to others who have reported on automated transcribing software (Moore 2015; Wardell et al. 2021), we acknowledge that the additional editing required after the initial transcript is time-consuming; however, Rev AI still saves a great deal of time compared to manual transcription.

Author Contributions

Margaret Broeren: writing – original draft, writing – review and editing, software, supervision, project administration, formal analysis, methodology, conceptualization, data curation, investigation, visualization, validation. **Yuzhe Gu:** writing – original draft, writing – review and editing, software, conceptualization, investigation, visualization. **Mark Pitt:** conceptualization, writing – original draft, writing – review and editing, software, investigation, resources, methodology. **Virginia Tompkins:** writing – review and editing, writing – original draft, conceptualization, data curation, formal analysis, investigation, resources.

Acknowledgements

We thank Holger Mitterer and Dan Strunk for providing recordings. We thank Heather Daly, Abigail Love, Makaylah Ekegren, Maggie Jiang, Eve Saltzman, Audrey Hutchison, Cameron Jones, Haoxuan Yuan, Emma Bookwalter, Min Feldman, Katherine Nelson, Lily Andrews, Rachel Grobman and Mohamad Ali Bamdad for their contributions to the project.

Ethics Statement

IRB approval was acquired prior to transcription scoring of the counselling session transcript. This study did not require additional ethics approval.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Peer Review

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/icd.70007.

Open Access Statement

With the exception of the recordings from the counselling sessions, data and materials are available upon request.

Endnotes

[1] For instructions on how to use 'cd' commands to navigate to a folder, see this tutorial: https://tutorials.codebar.io/command-line/introduction/tutorial.html.

[2] Some users may use 'python str.py' instead.

[3] Note that your API token is linked with the usage of your Rev AI account balance. We recommend that you do not save your API token on a public computer.

[4] Some Mac users may experience difficulties with the GUI being unresponsive to mouse clicks. To resolve this, click outside of the GUI window a few times, or move the cursor around, before clicking on the GUI again.

[5] Insufficient funds will result in the program closing without an error message.

[6] During revision of this paper, we learned of enhancements to rev.ai. The two of most relevance to readers are the addition of a "premium" transcription engine and public release of one of their transcription engines (https://github.com/revdotcom/reverb). We updated str.py to include engine type as an additional transcription variable but have not evaluated it. Our script utilises the "standard" version of diarization provided by the Rev AI API by default. If you would like to use the "premium" version, see *Additional Transcription Variables*.

[7] For a full explanation of these variables, visit: https://docs.rev.ai/api/asynchronous/reference/#operation/SubmitTranscriptionJob.

[8] At the time of transcription, asynchronous machine transcription cost $0.002/min.

[9] At the time of transcription, asynchronous machine transcription cost $0.002/min.

References

Bailey, J. 2008. "First Steps in Qualitative Data Analysis: Transcribing." *Family Practice* 25, no. 2: 127–131. https://doi.org/10.1093/fampra/cmn003.

Beckwith, R., L. Bloom, D. Albury, A. Raqib, and R. Booth. 1985. "Technology and Methodology." *Transcript Analysis* 2: 72–75.

Bhandari, N., D. Chen, M. Á. del Río Fernández, et al. 2024. "Reverb: Open-Source ASR and Diarization From Rev." *arXiv Preprint arXiv:2410.03930*.

Bolden, G. B. 2015. "Transcribing as Research: "Manual" Transcription and Conversation Analysis." *Research on Language and Social Interaction* 48, no. 3: 276–280. https://doi.org/10.1080/08351813.2015.1058603.

Brown, R. 1973. "Development of the First Language in the Human Species." *American Psychologist* 28, no. 2: 97–106. https://doi.org/10.1037/h0034209.

Chen, Y., N. Cabrera, C. Sudduth, and S. M. Reich. 2024. "Contributions of Mothers' and Fathers' Shared Book Reading With Infants at 9 Months to Language Skills at 18 Months in Ethnically and Socioeconomically Diverse Families." *Infant and Child Development* 33, no. 5: e2516. https://doi-org.proxy.lib.ohio-state.edu/10.1002/icd.2516.

Dragon Naturally Speaking Voice Recognition Software. 2023. "Computer Software."

Dragon Speech Recognition—Get More Done by Voice | Nuance. 2016. "Nuance Communications."

Hart, B., and T. R. Risley. 1995. *Meaningful Differences in the Everyday Experience of Young American Children*. Paul H Brookes Publishing.

Heilmann, J. J. 2010. "Myths and Realities of Language Sample Analysis." *Perspectives on Language Learning and Education* 17, no. 1: 4–8. https://doi.org/10.1044/lle17.1.4.

Hepburn, A., and J. Potter. 2021. *Essentials of Conversation Analysis*. American Psychological Association.

Jette, M. 2020. "The Podcast Challenge: Testing Rev.ai's Speech Recognition Accuracy. Rev."

Katz-Buonincontro, J. 2022. *How to Interview and Conduct Focus Groups*. American Psychological Association.

Lüdtke, U., J. Bornman, F. de Wet, et al. 2023. "Multidisciplinary Perspectives on Automatic Analysis of Children's Language Samples: Where Do We Go From Here?" *Folia Phoniatrica et Logopaedica: International Journal of Phoniatrics, Speech Therapy and Communication Pathology* 75, no. 1: 1–12. https://doi.org/10.1159/000527427.

MacWhinney, B. 2000. *The CHILDES Project: Tools for Analyzing Talk*. 3rd ed. Lawrence Erlbaum Associates.

Moore, R. J. 2015. "Automated Transcription and Conversation Analysis." *Research on Language and Social Interaction* 48, no. 3: 253–270. https://doi.org/10.1080/08351813.2015.1058600.

O'Shaughnessy, D. 2024. "Trends and Developments in Automatic Speech Recognition Research." *Computer Speech & Language* 83: 101538. https://doi.org/10.1016/j.csl.2023.101538.

Potamianos, A., and S. Narayanan. 2003. "Robust Recognition of Children's Speech." *IEEE Transactions on Speech and Audio Processing* 11, no. 6: 603–616.

Prabhavalkar, R., T. Hori, T. N. Sainath, R. Schlüter, and S. Watanabe. 2023. "End-To-End Speech Recognition: A Survey (arXiv:2303.03329)." *ArXiv*. https://doi.org/10.48550/arXiv.2303.03329.

Radford, A., J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. 2022. "Robust Speech Recognition via Large-Scale Weak Supervision." *arXiv:2212.04356 arXiv*. https://doi.org/10.48550/arXiv.2212.04356.

Robert, J., and M. Webbie. 2018. "Pydub. GitHub." http://pydub.com/.

Russell, S., A. L. Bird, K. Waldie, et al. 2024. "From Infancy to Eight: How Early Maternal Mental Health, Emotion Reminiscing, and Language Shape Children's Mental Health." *Development and Psychopathology*: 1–15. https://doi.org/10.1017/s0954579424000919.

Sacks, H., E. A. Schegloff, and G. Jefferson. 1974. "A Simplest Systematics for the Organization of Turn-Taking for Conversation." *Language* 50: 696–735.

Shahin, M., U. Zafar, and A. Beena. 2020. "The Automatic Detection of Speech Disorders in Children: Children, Opportunities, and Preliminary Results." *IEEE Journal of Selected Topics in Signal Processing* 14, no. 2: 400–412.

Sidnell, J. 2010. *Conversation Analysis: An Introduction*. Wiley-Blackwell.

Sierra, J. 2010. *Tell the Truth, B. B. Wolf*, edited by J. Seibold. Knopf Books for Young Readers.

Soltau, H., G. Saon, and B. Kingsbury. 2010. "The IBM Attila Speech Recognition Toolkit." *2010 IEEE Spoken Language Technology Workshop*: 97–102. https://doi.org/10.1109/slt.2010.5700829.

Stoel-Gammon, C. 2001. "Transcribing the Speech of Young Children." *Topics in Language Disorders* 21, no. 4: 12–21.

ten Haven, P. 1999. "Transcribing Talk in Interaction. A Practical Guide." In *Doing Conversation Analysis*. Sage.

Tomar, S. 2006. "Converting Video Formats With FFmpeg." *Linux Journal* 2006, no. 146: 10.

Tompkins, V. 2019/2020. Parent-Child mental state talk in repeated readings of the same storybook. The Ohio State University at Lima.

Van Rossum, G., and F. L. Drake. 2009. *Python 3 Reference Manual*. CreateSpace.

Vaswani, V. 2022. "Get Started With Speech Recognition in Python. Rev AI Developer Documentation."

Wardell, V., C. L. Esposito, C. R. Madan, and D. J. Palombo. 2021. "Semi-Automated Transcription and Scoring of Autobiographical Memory Narratives." *Behavior Research Methods* 53, no. 2: 507–517. https://doi.org/10.3758/s13428-020-01437-w.

Yeung, G., and A. Alwan. 2018. "On the Difficulties of Automatic Speech Recognition for Kindergarten-Aged Children." 1661–1665. https://doi.org/10.21437/Interspeech.2018-2297.