# Building the AVATAR Therapy Dialogues Corpus: the process of constructing a longitudinal corpus of three-way psychotherapeutic interactions

Sinéad Jackson,[1] Mark Huckvale,[1] Sandra Bucci,[2,3]
Moya Clancy,[4,5] Clementine Edwards,[6,7]
Miriam Fornells-Ambrojo,[8,9] Andrew Gumley,[4,5]
Gillian Haddock,[2,3] Thomas Jamieson-Craig,[7,10]
Jeffrey McDonnell,[8,9] Hamish McLeod,[4,5]
Alice Montague,[8,9] Mar Rus-Calafell,[6,11] Thomas Ward,[6,7]
Nikos Xanidis[4,5] and Philippa Garety[6,7]

**Abstract**

AVATAR therapy is an innovative form of relational therapy for the treatment of distressing auditory verbal hallucinations, or voice-hearing, targeted at reducing voice-related distress. AVATAR therapy involves the creation of a digital simulation of a single voice, termed an 'avatar', which is used in a series of three-way therapeutic dialogues.

[1] Department of Speech, Hearing and Phonetic Sciences, University College London, London, UK.
[2] Division of Psychology and Mental Health, School of Health Sciences, University of Manchester and the Manchester Academic Health Sciences Centre, Manchester, UK.
[3] Greater Manchester Mental Health NHS Foundation Trust and the Manchester Academic Health Sciences Centre, Manchester, UK.
[4] School of Health and Wellbeing, University of Glasgow, Glasgow, UK.
[5] NHS Greater Glasgow and Clyde, Glasgow, UK.
[6] Department of Psychology, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK.
[7] South London and Maudsley NHS Foundation Trust, London, UK.
[8] Research Department of Clinical, Educational and Health Psychology, University College London, London, UK.
[9] North East London NHS Foundation Trust, London, UK.
[10] Department of Health Service and Population Research, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK.
[11] Mental Health Research and Treatment Center, Faculty of Psychology, Ruhr-Universität Bochum, Bochum, Germany.
*Correspondence to*: Sinéad Jackson, *e-mail*: sinead.jackson.21@ucl.ac.uk

*Corpora* 2025 Vol. 20 (3): 363–381
DOI: 10.3366/cor.2025.0345

www.euppublishing.com/cor

This paper presents the AVATAR Therapy Dialogues Corpus, a specialised corpus containing orthographic transcriptions of AVATAR therapy sessions. We offer an overview of the corpus contents, and a detailed discussion of the design and construction of the corpus. We describe the processes and specialised tools created, transcription conventions, and mark-up designed to capture para-linguistic and non-speech features which may have clinical relevance. Finally, we discuss the potential of the corpus to provide a genuine innovation in clinical care, offering clinicians a data stream that could augment their understanding of patient experiences.

**Keywords**
AVATAR therapy, corpus building, corpus linguistics, psychotherapy, voice-hearing

## 1. Introduction

'The AVATAR Therapy Dialogues Corpus', also known as 'The AVATAR Corpus', is a specialised corpus containing orthographic transcriptions of AVATAR2 therapy sessions. AVATAR therapy is a novel therapy which aims to reduce the distress associated with voice hearing in schizophrenia and other psychoses. The AVATAR Corpus was designed to enable continuing research targeted at better understanding and evaluating AVATAR therapy, and processes of therapeutic change and transformation more broadly. The AVATAR Corpus exists as the only resource of its kind – a specialised, multi-party spoken corpus of a unique and innovative therapeutic context. This paper presents an account of the methodological decisions, procedures and tools used in its creation. Understandably, ethical concerns surrounding the collection and archiving of therapeutic data mean that specialised corpora of these interactions are rare, and accounts of the unique challenges and procedures to be followed in constructing such corpora are scarce. The following account may, therefore, be of use to readers interested in corpus-building, specifically building small, specialised corpora, multi-party dialogue corpora, and those working with psychotherapeutic data.

Research examining the mechanisms of therapeutic effect are crucial in the development of therapeutic interventions. Much of this research relies on transcripts of therapy sessions, yet the manual transcription process is both costly and labour intensive, often resulting in relatively small datasets which can limit the scope of analysis. Whilst recent advancements in the use of automatic speech recognition (ASR) technology to create spoken psychotherapeutic corpora offer promising alternatives (Miner *et al.*, 2020), with researchers working towards creating functional software specifically for this purpose (Flemotomos *et al.*, 2022), they are not without challenges. These methods continue to require refinement, primarily concerning accuracy and data protection. In addition, some findings indicate that such models may under-perform when processing clinical speech, compared to general conversation (Kodish-Wachs *et al.*, 2018). Our current capacity to fully explore the relationship between the structural and linguistic features of therapy sessions, and their impact on patient outcomes, therefore remains limited.

In response to the need for more comprehensive tools to analyse clinical interactions, efforts have been made to develop corpora that facilitate detailed linguistic and therapeutic research. The DAIS-C is one such example, developed to investigate the relationship between linguistic creativity and formal thought disorder in schizophrenia (Delgaram-Nejad *et al.*, 2023). In this paper, we present The AVATAR Corpus as a contribution to this endeavour, which has been designed to support inter-disciplinary research at the intersection of clinical psychiatry and linguistics. By including detailed information on turn length, annotation of paralinguistic and prosodic features such as laughter and pauses, as well as anonymisation codes that preserve relevant contextual information, the corpus aims to support a broad spectrum of research questions relating to both the clinical population it represents, the therapy type, and psychotherapy in general. The corpus was produced using a custom, four-step semi-automated procedure, making use of a combination of ASR and manual transcription. The method employed systematically applied closed sets of conventions, following recommendations on generating transcriptions for corpus studies from the field of applied linguistics.

We begin by presenting a brief overview of AVATAR therapy, followed by a description of the linguistic data which comprises the corpus. The paper concludes by considering future applications of the corpus.

## 2. AVATAR therapy

It is estimated that between 60 percent and 80 percent of individuals diagnosed with schizophrenia-spectrum disorders experience auditory verbal hallucinations in the form of hearing voices (Waters *et al.*, 2012). For many, voice-hearing is associated with social isolation and high levels of depression and distress (Han *et al.*, 2012). Therapeutic interventions which seek to shift the relational dynamic between voice-hearer and voice have delivered promising results (Dellazizzo *et al.*, 2022). These relational approaches aim to decrease the levels of distress associated with voice-hearing, helping to improve patient quality of life and general wellbeing. AVATAR therapy (Leff *et al.*, 2013), a specific intervention for the treatment of voice hearing in psychosis and the subject of multiple randomised controlled clinical trials in the UK (Craig *et al.*, 2018; and Garety *et al.*, 2024), is one such relational therapy.

During AVATAR therapy, voice-hearers engage in ‘face to face’ dialogues with a therapist-controlled, digital representation matching the auditory characteristics and associated imagery of their most prominent or distressing voice (Huckvale *et al.*, 2013). The therapy aims to re-balance the voice-hearer–voice relationship, increasing the voice-hearers’ sense of power and control (Ward *et al.*, 2020). In 2018, a randomised controlled trial (Craig *et al.*, 2018) found that AVATAR therapy resulted in a significant reduction in voice frequency and associated distress at twelve weeks when compared with supportive counselling. A second multi-centre clinical trial, AVATAR2 (Garety *et al.*, 2024), explored the efficacy of the therapy in brief and extended formats.

The AVATAR Corpus was built using data from AVATAR2 (Garety *et al.*, 2024), which tested two versions of AVATAR therapy: AVATAR Brief (containing six active dialogue sessions per course of treatment) and AVATAR Extended (containing twelve active dialogue sessions per course of treatment). Trial participants were individuals with a clinical diagnosis of Schizophrenia spectrum disorder (ICD10 F20–29) or Affective disorder with psychotic symptoms (ICD10 F30–39) who currently experience frequent and distressing voices. AVATAR Brief (AVATAR-BRF) takes a streamlined approach, with a focus on empowerment and building self-confidence as core targets of the treatment. AVATAR Extended (AVATAR-EXT) mirrors the early sessions of AVATAR-BRF but transitions into a second phase focussed on developing a shared understanding (or formulation), taking account of personal context and life history. This understanding informed Phase 2 dialogues which could consider the full range of treatment targets identified by Ward *et al.* (2020). See Garety *et al.* (2024) for a full description of the trial protocol, participant recruitment process and eligibility criteria.

Each session lasts approximately sixty minutes and consists of three parts: a pre-dialogue, an active dialogue, and a post-dialogue (Ward *et al.*, 2020). The AVATAR Corpus contains transcriptions of active avatar dialogues, the segment where the participant is communicating directly with the therapist-controlled avatar. For these dialogues, which can last anywhere from five to twenty minutes, the therapist and participant sit in separate rooms. Using specialised AVATAR system software, the therapist communicates with the participant either in their own voice, or through the avatar in an avatar voice. The result is a three-way dialogue between avatar, participant, and therapist (see Figure 1). The therapist determines the length of the active dialogues by beginning and ending a recording using the capabilities of the software. Most recordings start with the therapist advising the client that they are about to hear the avatar, and end with an invitation to finish the dialogue. Transcripts in the corpus contain written records of all speech produced by the three interlocutors (participant, therapist and avatar) for the duration of these recordings.

## 3. Corpus collection

The AVATAR Corpus contains 756 transcripts representing 100 trial participants: 53 AVATAR-BRF and 47 AVATAR-EXT. Session recordings were collected from the four UK research sites participating in the AVATAR2 therapy trial.

The building of the corpus happened concurrently with the AVATAR2 trial, and so data collection was dictated by availability. We set a target number of participants ($n = 100$), downloading the first twenty-five completed

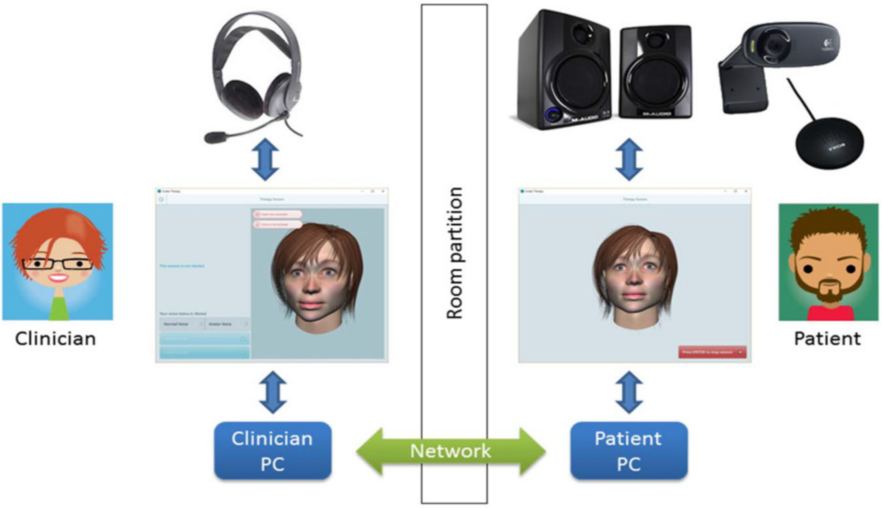


**Figure 1**: System configuration for AVATAR therapy.

| Site | AVATAR-BRF<br>$n = 53$ | AVATAR-EXT<br>$n = 47$ | Total<br>$n = 100$ |
|---|---|---|---|
| P01 | 13 | 14 | 27 |
| P02 | 14 | 12 | 26 |
| P03 | 13 | 9 | 22 |
| P04 | 13 | 12 | 25 |

**Table 1**: Structure of each sub-corpus by site and no. of participants.

| | AVATAR-BRF | *n* | AVATAR-EXT | *n* |
|---|---|---|---|---|
| No. of sessions | <6 | 8 | <12 | 29 |
| | 6 | 29 | 0.12 | 14 |
| | >6 | 16 | >12 | 4 |

**Table 2**: Number of sessions per participant.

sets per site of recordings for participants with >3 sessions as sessions were completed. Participants were excluded if they had completed fewer than three sessions or if the audio was distorted on more than three sessions. It should be noted that the actual number of sessions per participant did not always align with the six- and twelve-week course outlined in the trial protocol – a variation informed by clinical judgment and guided by specified criteria within the protocol. Tables 1 and 2 show the final number of participants per site and sessions per participant.

The AVATAR Corpus is divided by therapy type into two sub-corpora: AVATAR-BRF and AVATAR-EXT. Within each sub-corpus, participants are given a pseudonymised ID to which their sessions are linked. There is one transcript per session. Transcripts are saved in plain text format, and are grouped hierarchically by site, participant ID, and session number. For example, "P010001_1":

| | |
|---|---|
| P01 | Site identifier |
| 0001 | Participant ID |
| _1 | Session number |

This longitudinal component enables the tracking of individual progression, as well as variation across sessions and between therapy types – a possibility we hope will lead to valuable clinical insights in future studies. The following section offers a brief overview of the contents of the corpus, using the measures token count, turn count and turn length.

## 4. Corpus characteristics

The corpus contains a total of 929,567 tokens, and 62,305 unique turns: 18,498 avatar, 28,256 client, and 15,551 therapist. Token counts per session for each sub-corpus are shown in Table 3.

Each line in the transcripts is timestamped and contains a speaker identification label in angular brackets ('<C>', '<T>' and '<A>' for client, therapist and avatar). It is of interest to note that there were some participant files where the Avatar changed identity. In order to uphold labelling consistency, which was designed to more easily allow for future automated linguistic research – for example, the ability to easily isolate all avatar turns by identifying only lines containing the <A> label – this change was not noted in the speaker label. Participant files containing multiple avatar identities were instead noted in a User Guide which accompanies the corpus.

Figure 2 shows the distribution of speaker turns within each session, represented as a percentage. A turn is defined here as a change in speaker. As can be seen, the percentage of turns contributed by the therapist is greater in earlier sessions. For example, in Session 1 the therapist takes 38 percent and 39 percent of turns in –BRF and –EXT, respectively, compared to the avatar's 18 percent. By Session 3 the therapist and avatar converge, each taking 27 percent of turns in both therapy types. In subsequent sessions, the therapist-controlled avatar takes a greater proportion than the therapist speaking in their own voice, potentially indicating a shift in the interactional role of the therapist. The percentage of client turns remains relatively stable across sessions, ranging between 43 percent and 46 percent in both –BRF and –EXT.

Table 4 shows the mean turn length (calculated as the mean number of words spoken in each speaker's turn) of each speaker per session, in each sub-corpus. Though present, changes in turn length for both avatar and

| Session | AVATAR-BRF Tokens | AVATAR-EXT Tokens |
|---|---|---|
| 1 | 44,796 | 44,108 |
| 2 | 48,418 | 50,405 |
| 3 | 54,589 | 47,795 |
| 4 | 56,164 | 51,282 |
| 5 | 64,835 | 56,612 |
| 6 | 57,288 | 53,223 |
| 7 | 13,833 | 53,380 |
| 8 | 6,148 | 52,748 |
| 9 | | 50,595 |
| 10 | | 51,033 |
| 11 | | 36,207 |
| 12 | | 32,418 |
| 13 | | 3,690 |
| 1–13 | 346,071 | 583,496 |

**Table 3**: Token counts per session.

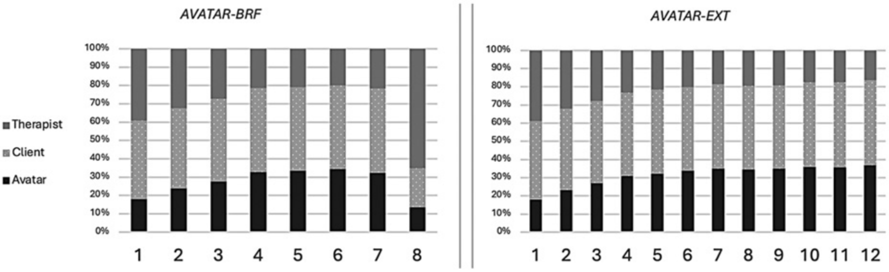


**Figure 2**: Percentage of turns taken by speakers across sessions.

therapist are small. Client turn length increases steadily across all sessions from fourteen to thirty-one words and seventeen to fifty-three words per turn in –BRF and –EXT, respectively. Considering the stable percentages of client turn counts, this could indicate that while the number of turns a client takes in each session does not vary, the amount of speech they produce increases. However, whilst a pattern is indicated, the large standard deviation for client turn lengths from Session 3 onwards points to a high degree of variation across participants.

| Session | AVATAR-BRF<br>mean turn length (SD) | | | Session | AVATAR-EXT<br>mean turn length (SD) | | |
|---|---|---|---|---|---|---|---|
| | Therapist | Avatar | Client | | Therapist | Avatar | Client |
| 1 | 23.4 (6.7) | 11.5 (1.7) | 14.4 (4.4) | 1 | 24.1 (7.8) | 11.4 (1.9) | 17.0 (9.6) |
| 2 | 23.7 (6.8) | 13.6 (1.9) | 17.4 (6.2) | 2 | 23.9 (7.7) | 13.3 (2.4) | 19.9 (12.1) |
| 3 | 23.4 (8.0) | 15.9 (4.0) | 23.1 (25.3) | 3 | 24.8 (5.3) | 15.5 (2.6) | 20.3 (10.2) |
| 4 | 23.0 (7.6) | 18.6 (5.1) | 24.5 (16.1) | 4 | 23.6 (6.6) | 17.6 (3.6) | 23.9 (12.6) |
| 5 | 22.3 (6.8) | 20.0 (6.2) | 27.2 (20.2) | 5 | 24.0 (9.5) | 17.7 (4.6) | 24.3 (14.8) |
| 6 | 22.7 (7.7) | 20.1 (5.6) | 29.6 (21.3) | 6 | 22.9 (7.2) | 18.6 (3.9) | 29.9 (19.0) |
| 7 | 22.3 (11.1) | 16.5 (5.5) | 26.6 (16.6) | 7 | 22.9 (9.2) | 18.7 (4.7) | 26.9 (13.9) |
| 8 | 23.8 (7.6) | 19.9 (4.5) | 31.4 (11.8) | 8 | 25.8 (12.4) | 18.9 (5.2) | 28.2 (17.3) |
| | | | | 9 | 22.4 (7.3) | 19.7 (5.4) | 31.9 (18.3) |
| | | | | 10 | 23.0 (9.9) | 19.5 (6.0) | 32.5 (23.0) |
| | | | | 11 | 20.2 (10.6) | 18.5 (4.3) | 32.9 (13.5) |
| | | | | 12 | 21.1 (11.7) | 18.4 (3.8) | 42.1 (26.6) |
| | | | | 13 | 19.4 (5.0) | 17.8 (1.1) | 53.6 (5.6) |

**Table 4**: Mean turn length per speaker across sessions. (Standard Deviation in brackets.)

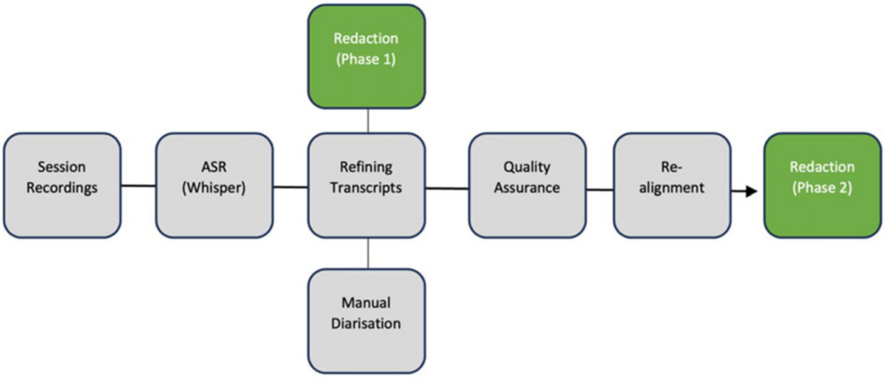


**Figure 3**: Dialogue transcription pipeline.

## 5. Corpus transcription

Compiling a written dataset of spoken psychotherapeutic interactions is not a straightforward task. The reliability and usability of a corpus as a research resource is dependent on transcription quality (Gablasova *et al.*, 2019), as the transcript serves as the basis for a systematic analysis of the data (Breiteneder *et al.*, 2006). In creating the corpus, we devised a custom-made five-step semi-automated procedure, making use of a combination of automatic speech recognition (ASR), computational methods and manual transcription, and capable of processing a relatively large amount of data whilst retaining an appropriate degree of linguistic sensitivity. As we hoped to create a resource that was suitable for computational, corpus and qualitative linguistic analysis, we aimed to maximise the inclusion of any linguistic phenomena which may have clinical relevance. Whilst the use of fully automated ASR methods would have been fast and cost-efficient, ASR models are currently incapable of identifying paralinguistic and non-speech features such as overlaps, laughter and hesitations (Umair *et al.*, 2022), all of which may be relevant to the analysis of therapeutic interactions (for an example, see Abbas [2020]).

We used an AI-trained ASR model, Whisper (OpenAI, n.d.), to generate first-pass transcripts of the recordings. These transcripts were then manually refined, diarisation (speaker identification) was added, and identifying personal information was noted. Quality assurance processes were performed, and the corrected transcriptions underwent a process to refine lexical and phonemic alignment. Finally, the previously identified personal information was redacted. This process is represented in Figure 3. Details of each stage are given below.

### 5.1 First pass automated transcription: Whisper

Whisper is an ASR model, selected for use in this project due to its ability to accurately transcribe different accents and audio files containing background noise, and, importantly, because it could be downloaded and run locally. This local operation preserved the privacy of the recordings.

Recordings made in the clinic were saved in stereo audio with therapist and avatar on one channel and client on the other, although there is often some cross-channel contamination. The audio was saved as MP3 files at 160kbit/s. For recordings made over video conferencing, audio was recorded on a single channel in MP4 format at 195kbit/s. To prepare the recordings for transcription by Whisper, they were converted to single channel recordings at 16,000 samples/sec in uncompressed audio.

Whisper processes the recordings in 30s chunks and concatenates the transcription of the chunks into a single file in VTT format. The VTT format is designed for subtitling, and each line contains up to 100 characters together with approximate timing information.

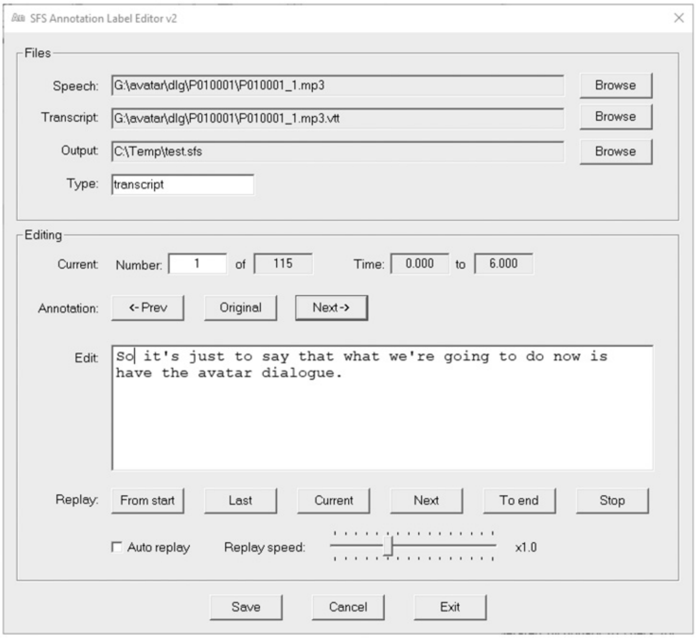
## 5.2 Manual editing and diarisation

The audio and VTT files for each recording were converted to the format required by the Speech Filing System (SFS) software (Huckvale, 2020), and SFS files containing the audio and the time-aligned lines of transcript were constructed. Tools within SFS then allowed the display of the aligned audio and transcription to help with checking for transcript accuracy and alignment.

Whilst SFS contains a tool for the correction of annotations, this was designed for phonetic annotations and was awkward to use for sentence level labels. A new tool was constructed which allowed manual correction of the transcript and alignment. This tool played the audio for each chunk found in the VTT file and made the transcription available for editing, as shown in Figure 4. The annotation tool has some optimisations for this application, including the ability to directly read MP3 and VTT files, with output to SFS files. It also allowed automatic replay for each transcribed portion, and the facility to replay the audio at a faster rate than normal.

Using the annotation tool, each of the session recordings were manually checked and corrected by a trained research worker. Mark-up of paralinguistic and non-speech features which had not been captured by the ASR model were added using closed sets of transcription conventions, discussed in more detail in Section 5.6. A full list of conventions can be found in Appendix A. Speaker labels were also inputted manually at this time.

We estimate that the manual correction of a fifteen-minute dialogue took around sixty minutes. The total elapsed time for this stage, encompassing diarisation, refining of orthographic transcriptions, and compiling a redaction key is estimated to be roughly 300 hours.



Speech: G:\avatar\dlg\P010001\P010001_1.mp3

Transcript: G:\avatar\dlg\P010001\P010001_1.mp3.vtt

Output: C:\Temp\test.sfs

Type: transcript

Current: Number: 1 of 115 Time: 0.000 to 6.000

Edit:

```
So it's just to say that what we're going to do now is
have the avatar dialogue.
```

**Figure 4**: Annotation editing

## 5.3 Quality assurance

To check the transcriptions, the following quality assurance processes were instigated:

(*i*) A lexicon was created for all the transcripts and sorted by frequency to identify low frequency terms that might be mis-spellings.

(*ii*) The lexicon was compared to the head words in two English pronunciation dictionaries, BEEP (Robinson, 1996) and CELEX (Baayen *et al*., 1995), to identify mis-spellings, non-English words and proper nouns.

(*iii*) The corrected transcripts were compared with the raw output of the ASR system to identify recordings with a particularly high number of changes. Overall, the number of changes in each transcript involved about 10 percent of the words. This process was initiated as we were interested in ascertaining the degree of accuracy of the ASR software used – information which may also be useful to researchers working with similar methods.

## 5.4 Refining alignments

The overall approach to alignment was to first generate phonemic transcripts by concatenating word pronunciations from dictionaries. Manual phonemic transcriptions were generated for 1,470 words found in the transcripts that

| Raw VTT output | After realignment |
|---|---|
| 00:24.520 --> 00:25.520<br>Okay.<br>00:25.520 --> 00:26.520<br>How are you feeling?<br>00:26.520 --> 00:29.160<br>Yeah I'm feeling good.<br>00:29.160 --> 00:31.720<br>Any anxiety or butterflies or anything? | 24.580 24.860 <C> Okay.<br>25.640 26.340 <T> How are you feeling?<br>27.040 28.100 <C> Yeah I'm feeling good.<br>29.000 31.060 <T> Any anxiety or butterflies or anything? |

**Table 5**: Example alignment refinement.

were missing from the BEEP and CELEX dictionaries. Next, a phone lattice was generated from the original audio recordings. This gives a vector of forty-five phone probabilities every 20ms through the signal. Lastly, a dynamic programming procedure was implemented to find the best alignment between transcript and lattice. See Table 5 for example input and output of the alignment refinement process.

To check the quality of the alignments, the average size of alignment shifts, and the largest single shift were collected and checked across all transcripts to find files where the alignment failed.

### 5.5 Redaction

Personally identifying information (PII) including names referring to people, addresses, places (such as hospitals, educational institutions, and so on), cities, and company names referencing local or personal companies were redacted. Countries and the names of large international conglomerates (for example Amazon or Google) were not redacted, as it was decided that it would not be possible to infer an individual's identity through reference to these widely known entities. Whilst this process was designed to anonymise the transcripts insofar as removing direct and indirect PII, it must be noted that it was not possible to remove discussion of personal or sensitive experiences due to the nature of the interactions.

The redaction process consisted of two phases. A key of PII for each participant was created as part of the initial editing. Once a set of transcripts had been edited, a dictionary of each unique word contained in the set was generated. The manually created key was then checked against the computer-generated dictionary. Once it was determined that any dictionary items referring to identifying information (as defined above) were accounted for in the key, an automated process was employed to search the transcripts and exchange noted words for a generic pseudonym. For example, a referenced person's name was substituted with "[Person$^{n}$]", numbered in order of appearance. Due to the complex nature of the interactions, it was not always possible to provide a straightforward substitution. For example, in some instances a referenced voice and a referenced person shared the same identity, by which we mean a client referred to a voice and a known person as being one and the same, attributing actions and behaviours to a singular voice/person entity. Whilst it may be possible (though not always) for a reader to disambiguate which is being referred to, the automated redaction procedure would not have been capable of correctly identifying which substitution is to be made. As a result, in these cases, the name was substituted with "[Person$^n$/Voice$^n$]". Users of the corpus must, therefore, be aware and make use of the redaction conventions when reading the transcripts. This difficulty highlights one of the limitations in using automated procedures to annotate such complex interactions. See Appendix B for details of these substitutions.

The final corpus consists of sets of redacted transcripts, linked by pseudonymised participant IDs. Users of the corpus will not receive access to the redaction key or any information created as part of the editing and redaction process.

### **5.6** Transcription conventions

Taking guidance from previously published research, we adopted a simple orthographic transcription, which accounted for paralinguistic and non-speech features such as interruptions, laughter, hesitations, filled and unfilled forms.

Following Atkins *et al.*'s (1992) recommendation, we used a closed set of permissible forms for non-standard spellings, in order to capture idiosyncratic pronunciation. For example, contractions such as *gonna*, *sorta*, etc., were transcribed as they were spoken, using pre-defined non-standard spellings. Non-standard words, for example *horribilize*, were transcribed as they were heard. Following Collins and Hardie (2022), we used a closed set of conventions for the mark-up of non-speech material using different flags for different phenomena. For example, square brackets were used for redactions and non-speech features, and braces were used for overlapping talk. For the mark-up of non-speech features, identifying abbreviations were enclosed in brackets to identify each marked phenomenon. For example, '[BR_yeah]', to indicate a word spoken as part of a sigh, and '[LG_yeah]' to indicate a word spoken whilst laughing. This formulaic approach has the benefit of generating transcripts which are suitable for analysis in various corpus linguistic software, whilst maintaining human readability. For example, square brackets as opposed to angular brackets were used so as not to be confused with the XML tags commonly used to structure corpora for use in popular corpus tools. Following Gablasova *et al.* (2019), hesitation sounds, filled pauses and backchannels were transcribed according to a closed set of permissible expressions which differentiated between major categories of these expressions (for example, *hmm* and *uhh*). See Appendix A for a full list of transcription conventions.

## 6. Ethics and availability

The corpus and an accompanying User Guide, including demographic information and participant outcome measures, will be made available for the purposes of non-commercial, ethically approved academic research through the KORDS Data Repository, hosted by Kings College London. Applications must be made to the research data team.[12]

All participants in the corpus have consented to the redacted transcripts of their session recordings being used in ethically approved research. However, certain participants have not consented to verbatim quotations being publicly shared or published. There are, therefore, limitations concerning the sharing of direct quotations from the corpus. On accessing the corpus, researchers will be provided with a list of transcripts from which quotations must not appear in written form in any publicly disseminated works. This includes (but is not limited to) scientific journals and academic presentations.

Ethics approval for this project was granted following review by the NHS/HRA London-Camberwell & St Giles Research Ethics Committee (REF: 20/LO/0657) on 10 June 2020.

## 7. Future directions for the corpus

In building the AVATAR Corpus, we aimed to create possibilities for research, enhancing the effectiveness of AVATAR therapy through the analysis of in-session language. The mechanics of therapeutic interactions carry a wealth of clinically relevant information. Analysis of linguistic patterns may be capable of revealing associations between linguistic markers and psychological phenomena, such as anxiety (Rook *et al.*, 2022), low self-esteem (Cheng *et al.*, 2023) and depression (Smirnova *et al.*, 2018) – symptomatology which may be useful to researchers hoping to better understand client experiences and progress. Speech characteristics may function as predictors or indicators of successful outcomes, and these insights could be used in the development of future training guidance for AVATAR therapists. For example, Knapton (2021) observed a relationship between the grammatical positioning of the self and the mind, and an individual's sense of agency, responsibility and blame, which they propose could be used to direct therapists towards areas of concern. In addition, in the course of a study using corpus linguistic methods to explore voice personification, Collins *et al.* (2023) propose that tracking changes in how individuals use semantic processes to attribute agency to the self and the voice could provide the basis for monitoring therapeutic progression in a longitudinal therapeutic context, which would be relevant to AVATAR therapy. Finally, datasets such as the AVATAR Therapy Dialogues Corpus enable the harnessing of machine learning and automated LLM processing: these methods have the potential to make therapy more personalised and efficient, helping clinicians to understand and help the people they are working with more effectively (Lutz *et al.*, 2022). We propose that the AVATAR Therapy

[12] At: research.data@kcl.ac.uk.

Dialogues Corpus will pave the way for novel studies concerning AVATAR therapy and relational therapies for voice-hearing more broadly, offering a significant contribution to the future development and implementation of this therapeutic model.

## Acknowledgments

The authors are grateful for the work of all the therapists on the AVATAR2 trial, and all the participants in the study who have graciously agreed that transcripts of their therapy sessions can be disseminated for research purposes. This work was supported by a Wellcome grant (no. 215471/Z/19/Z), awarded to King's College London as part of the AVATAR2 clinical trial.

## References

Abbas, N.F. 2020. 'Pragmatics of overlapping talk in therapy sessions', Journal of Language and Linguistic Studies 16 (3), pp. 1251–63.

Atkins, S., J. Clear and N. Ostler. 1992. 'Corpus design criteria', Literary and Linguistic Computing 7 (1), pp. 1–16.

Baayen, R.H., R. Piepenbrock and L. Gulikers. 1995. CELEX2 LDC96L14. (Dictionaries.) Linguistic Data Consortium.

Breiteneder, A., M.-L. Pitzl, S. Majewski and T. Klimpfinger. 2006. 'VOICE recording – methodological challenges in the compilation of a corpus of spoken ELF', Nordic Journal of English Studies S2, pp. 161–87.

Cheng, L., H. Mao and T. Zhang. 2023. 'Cognitive-pragmatic functions of mitigation in therapeutic conversations emphasizing rapport management', Frontiers in Psychology 14.

Collins, L. and A. Hardie. 2022. 'Making use of transcription data from qualitative research within a corpus-linguistic paradigm: issues, experiences and recommendations', Corpora 17 (1), pp. 123–35.

Collins, L., V. Brezina, Z. Demjén, E. Semino and A. Woods. 2023. 'Corpus linguistics and clinical psychology: investigating personification in first-person accounts of voice-hearing', International Journal of Corpus Linguistics 28 (1), pp. 28–59.

Craig, T.K., M. Rus-Calafell, T. Ward, J.P. Leff, M. Huckvale, E. Howarth, R. Emsley and P.A. Garety. 2018. 'AVATAR therapy for auditory verbal hallucinations in people with psychosis: a single-blind, randomised controlled trial', The Lancet Psychiatry 5 (1), pp. 31–40.

Delgaram-Nejad, O., D. Archer, G. Chatzidamianos, L. Robinson and A. Bartha. 2023. 'The DAIS-C: a small, specialised, spoken, schizophrenia corpus', Applied Corpus Linguistics 3 (3), 100069.

Dellazizzo, L., S. Giguère, N. Léveillé, S. Potvin and A. Dumais. 2022. 'A systematic review of relational-based therapies for the treatment of auditory hallucinations in patients with psychotic disorders', Psychol Med. 52 (11), pp. 2001–8.

Flemotomos, N., V.R. Martinez, Z. Chen, K. Singla, V. Ardulov, R. Peri, D.D. Caperton, J. Gibson, M.J. Tanana, P. Georgiou, J. Van Epps, S.P. Lord, T. Hirsch, Z.E. Imel, D.C. Atkins and S. Narayanan. 2022. 'Automated evaluation of psychotherapy skills using speech and language technologies', Behavior Research Methods 54 (2), pp. 690–711.

Gablasova, D., V. Brezina and T. McEnery. 2019. 'The Trinity Lancaster Corpus: development, description and application', International Journal of Learner Corpus Research 5 (2), pp. 126–58.

Garety, P.A., C.J. Edwards, H. Jafari, R. Emsley, M. Huckvale, M. Rus-Calafell, M. Fornells-Ambrojo, A. Gumley, G. Haddock, S. Bucci, H.J. McLeod, J. McDonnell, M. Clancy, M. Fitzsimmons, H. Ball, A. Montague, N. Xanidis, A. Hardy, T.K.J. Craig and T. Ward. 2024. 'Digital avatar therapy for distressing voices in psychosis: the phase 2/3 AVATAR2 trial', Nature Medicine 30 (12): 3658–68.

Han, K., J.K. Heo, S.O. Seo, M.Y. Hong, J.S. Lee, Y.S. Shin, J. Ku, S.I. Kim and J.J. Kim. 2012. 'The effect of simulated auditory hallucinations on daily activities in schizophrenia patients', Psychopathology 45 (6), pp. 352–60.

Huckvale, M. 2020. Speech Filing System (SFS). (Version 4.10) (Windows.) London: University College London.

Huckvale, M., J. Leff and G. Williams. 2013. 'Avatar Therapy: an audio-visual dialogue system for treating auditory hallucinations' in proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH) 2013, pp. 392–6. 25–29 August 2013. Lyon, France.

Knapton, O. 2021. 'The linguistic construction of the self in narratives of obsessive-compulsive disorder', Qualitative Research in Psychology 18 (2), pp. 204–26.

Kodish-Wachs, J., E. Agassi, P. Kenny and J.M. Overhage. 2018. 'A systematic comparison of contemporary automatic speech recognition engines for conversational clinical speech', AMIA Annu Symp Proc. pp. 683–9.

Leff, J., G. Williams, M.A. Huckvale, A. Arbuthnot and A.P. Leff. 2013. 'Computer-assisted therapy for medication-resistant auditory hallucinations: proof-of-concept study', The British Journal of Psychiatry: The Journal of Mental Science 202, pp. 428–33.

Lutz, W., B. Schwartz and J. Delgadillo. 2022. 'Measurement-based and data-informed psychological therapy', Annual Review of Clinical Psychology 18 (1), pp. 71–98.

Miner, A.S., A. Haque, J.A. Fries, S.L. Fleming, D.E. Wilfley, G. Terence Wilson, A. Milstein, D. Jurafsky, B.A. Arnow, W. Stewart Agras, L. Fei-Fei and N.H. Shah. 2020. 'Assessing the accuracy of automatic speech recognition for psychotherapy', NPJ Digital Medicine 3, 82.

OpenAI. (n.d.). Introducing Whisper. (Computer software.).

Robinson, T. 1996. BEEP Dictionary. (Dataset.) Cambridge: University of Cambridge.

Rook, L., M.C. Mazza, I. Lefter and F. Brazier. 2022. 'Toward linguistic recognition of generalized anxiety disorder', Frontiers in Digital Health 4, 779039.

Smirnova, D., P. Cumming, E. Sloeva, N. Kuvshinova, D. Romanov and G. Nosachev. 2018. 'Language patterns discriminate mild depression from normal sadness and euthymic state', Frontiers in Psychiatry 9, 105.

Umair, M., J.B. Mertens, S. Albert and J.P. de Ruiter. 2022. 'GailBot: an automatic transcription system for Conversation Analysis', Dialogue & Discourse 13 (1), pp. 63–95.

Ward, T., M. Rus-Calafell, Z. Ramadhan, O. Soumelidou, M. Fornells-Ambrojo, P. Garety and T.K.J. Craig. 2020. 'avatar therapy for distressing voices: a comprehensive account of therapeutic targets', Schizophrenia Bulletin 46 (5), pp. 1038–44.

Waters, F, P. Allen, A. Aleman, C. Fernyhough, T.S. Woodward, J.C. Badcock, E. Barkus, L. Johns, F. Varese, M. Menon, A. Vercammen and F. Larøi. 2012. 'Auditory hallucinations in schizophrenia and nonschizophrenia populations: a review and integrated model of cognitive mechanisms', Schizophr Bull 38 (4), pp. 683–93.

**Appendix A**: Transcription conventions and non-speech labels.

| | |
|---|---|
| Hesitation sounds/filled pauses, and yes/no sounds | |
| *uh*<br>*um*<br>*uhm*<br>*uhhm* | Thinking/filler/confusion<br>(depending on what the speech sounds closest to) |
| *mmm* | Demonstrating active listening |
| *hmm* | Demonstrating consideration/confusion |
| *mhmm*<br>*uh-huh* | Demonstrating affirmation/agreement (depending on what the speech sounds closest to) |
| *ahh* | Demonstrating realisation |
| *aagh* | Guttural sound demonstrating frustration/anger |
| Collocations | |
| *gonna* | *going to*, as in 'I'm gonna try to call it different names' |
| *y'know* | *you know*, as in 'despite, y'know, bumps in the road' |
| *tryna* | *trying to*, as in 'you're tryna let her understand' |
| *gotta* | *got to*, as in 'I've gotta be honest with you' |
| *wanna* | *want to*, as in 'do you wanna just say' |
| *sorta* | *sort of*, as in 'he just sorta said' |
| *kinda* | *kind of*, as in 'I hope you can kinda change' |
| *ain't* | *am not*, as in 'You ain't backing down today' |
| Other symbols | |
| dash | Word is spoken partially |
| [SIL] | Silence by current speaker not due to the other person talking occurring within a turn (minimum 1 second) |
| [LG] | Laughter that is not part of any word. |
| [LG_word] | Laughter that is part of a word |
| [BR] | Breaths/sighs |
| [GA] | Garbage noise that is not from the speaker (feedback, background noise, etc.) |
| [GA_word] | Garbage noise that occurs while the speaker is saying a word. |
| [?] | Transcriber has not understood which word is intended. |
| [word_?] | Transcriber has attempted to transcribe a word, but is unsure. |
| {word} | Overlapping talk between speakers |
| [OS_word] | Words spoken by an outside speaker |

## **Appendix B**: Redaction key

| Label | Description |
|---|---|
| [Therapist] | Therapist’s name |
| [Client] | Client’s name |
| [Avatar] | Avatar’s name |
| [Avatar/Person $^{n}$] | Labelling convention when the Avatar and a referenced person share the same identity. |
| [Voice $^{n}$] | Referenced voice with a different identity to the Avatar voice. |
| [Person $^{n}$/Voice $^{n}$] | Labelling convention when a referenced person and referenced voice share the same identity. |
| [Person$^{n}$] | Referenced person (numbered in order of appearance) |
| [Place$^{n}$] | Referenced place (numbered in order of appearance) |
| [Website] | Referenced website |
| [Company] | Referenced company |