Original Article

Automated generation of 'good enough' transcripts as a first step to transcription of audio-recorded data

Methodological Innovations
May-August 2018: 1–14

DOI: 10.1177/2059799118790743
journals.sagepub.com/home/mio
SAGE

Christian Bokhove and Christopher Downey

Abstract

In the last decade, automated captioning services have appeared in mainstream technology use. Until now, the focus of these services have been on the technical aspects, supporting pupils with special educational needs and supporting teaching and learning of second language students. Only limited explorations have been attempted regarding its use for research purposes: transcription of audio recordings. This article presents a proof-of-concept exploration utilising three examples of automated transcription of audio recordings from different contexts; an interview, a public hearing and a classroom setting, and compares them against 'manual' transcription techniques in each case. It begins with an overview of literature on automated captioning and the use of voice recognition tools for the purposes of transcription. An account is provided of the specific processes and tools used for the generation of the automated captions followed by some basic processing of the captions to produce automated transcripts. Originality checking software was used to determine a percentage match between the automated transcript and a manual version as a basic measure of the potential usability of each of the automated transcripts. Some analysis of the more common and persistent mismatches observed between automated and manual transcripts is provided, revealing that the majority of mismatches would be easily identified and rectified in a review and edit of the automated transcript. Finally, some of the challenges and limitations of the approach are considered. These limitations notwithstanding, we conclude that this form of automated transcription provides 'good enough' transcription for first versions of transcripts. The time and cost advantages of this could be considerable, even for the production of summary or gisted transcripts.

Keywords

Interviews, transcription, qualitative data, automated captions, technology, automatic speech recognition

Introduction

A number of key methods in social science involve some form of transcription of audio. As the UK data archive posits, transcription work is a time-consuming process that often is outsourced to external transcribers (UK data archive, 2017). In projects involving multiple rounds of data analysis, it is important to follow standardised guidelines while transcribing. This is one of the reasons why large-scale qualitative projects often release a separate detailed transcription and translation manual. There are various types of transcript of audio recordings for research purposes, depending on the degree of detail required in the transcription process, from capturing additional information such as pauses and intonation, through to the production of condensed or essence transcripts where some of the information captured in the raw audio recording is deliberately omitted from the transcript. Regardless of the type of transcript required, it is commonly accepted that the process of transcription is likely to require multiple rounds of engagement with the audio file (Paulus et al., 2013). We posit therefore that the generation of transcripts, utilising technology such as automated speech recognition (ASR) tools embedded in web-based auto-captioning

Southampton Education School, University of Southampton, Southampton, UK

Corresponding author:
Christian Bokhove, Southampton Education School, University of Southampton, Highfield, Southampton SO17 1BJ, UK.
Email: C.Bokhove@soton.ac.uk

services, could provide a useful first draft for use in later cycles of the process of transcription, provided that the quality of the automated transcript is sufficient to serve as a foundation for further editing, addition and improvement.

This article, therefore, revisits the recommendation from the UK data archive (2017) not to automate the transcription of recorded interviews, as these tools 'all require a great deal of training and calibration to be able to recognise a particular voice, accent and dialect'. It also considers whether recent advances in technology might allow us to utilise freely available web-tools to quickly come to 'good enough' first drafts of transcripts. We contend that such a workflow might significantly reduce the time and costs involved in the transcription process. This may well lead to significant gains for researchers working in fields for which gaining grant funding for research projects can be particularly challenging. This may include gains for those researching within controversial fields for which access to external research funding may be very limited. It should also reduce the overall costs associated with transcription, and so help avoid the need to compromise on the scale and scope of a project during the research design stage, if this is in an effort to balance the costs of the research within the limits of available funding.

The article engages with a number of relevant themes from the literature, giving an overview of some of the challenges in the transcription process from the point of view of a researcher. The review also considers literature related to some of the technical and educational aspects of the use of automated captioning, as well as the use of voice and speech recognition tools for the purposes of transcription in a variety of contexts including non-research related settings. This is followed by a proof-of-concept exploration of an approach to the use of auto-captioning technologies to generate automated transcripts. Three different audio sources are used as raw data; captured from three different environments. One is from a school classroom setting, one from a public hearing with multiple speakers in a larger space, and finally audio data from a one-to-one interview setting. Full details of the process, via freely available web-based auto-captioning and basic caption-processing tools, is provided as an example of an ASR approach that might be used to generate automated transcripts. The quality of the resulting transcripts is tested through calculation of a percentage match between the automated transcript and a manually produced transcript using a common software tool for originality checking (Turnitin, 2017). An analysis of some of the most common and persistent mismatches observed between the automated and manual transcripts is provided to consider whether these present obstacles in the production of automated transcripts that diminish the utility of the transcripts as a 'first draft' effort. Finally, the challenges and limitations of the approach, via auto-captioning software tools, are considered before conclusions are drawn as to the potential of the technique and whether the proof-of-concept was successful. The ethical aspects, for example, are key to consider.

Relevant literature

The theme of this article covers two main areas of methods research associated with preparation of transcripts from audio recordings. First, the process of transcription itself, second, the role of automated production of transcripts. This section aims to present a non-exhaustive overview of some of the aspects involved in producing automated transcripts.

Transcribing interviews

Like the UK data archive (2017), research method textbooks usually describe transcription as a time-consuming process, forming part of the qualitative research realm (Cohen et al., 2007). As interviews can be immensely rich in data and detail, verbatim transcripts are considered to convey these meanings best (Cohen et al., 2007: 462). The requirement to produce verbatim transcripts can also be seen as an act of respect for the participants in a study: they are devoting valuable time to the research, and therefore, it is only reasonable to record every word. However, the time-consuming nature of the transcription process has caused others to refer to the 'fetish of transcription' (Walford, 2001: 92). With a suggested ratio of five to one – 5 hours to transcribe 1 hour of interviews, Walford (2001) suggested it is particularly costly in terms of time. Punch and Oancea (2014) subscribe to a rule of thumb of needing at least 4 hours for every one hour. Audio recordings are heavily contextualised, which means that transcripts 'inevitably lose data from the original encounter' (Punch and Oancea, 2014: 367), requiring translation from one set of rule systems (oral and interpersonal) to another rule system (written). Kvale (1996: 166) highlights this by pointing out that the prefix trans in *tran*script indicates a change of state or form, in essence a selective transformation. Lee (1993) refers to the issue of 'transcriber selectivity', while Kvale (1996:) holds the view that 'transcripts of interviews, however detailed and full they might be, remain selective, since they are interpretations of social situations' (p. 163). In this view, transcripts might be considered to be already interpreted data; the transcript acts as a 'screen' between the researcher and the original situation of the recording (Kvale, 1996: 167). The ultimate consequence of this is that there can be no single 'correct' transcript, rather only transcripts that are more or less useful for the research. Taking this relative notion of transcripts even further, transcripts can be said to be 'decontextualized, abstracted from time and space, from the dynamics of the situation, from the live form, and from the social, interactive, dynamic and fluid dimensions of their source; they are frozen' (Kvale, 1996: 367).

Taking both aspects at face value, namely the time-consuming nature of transcription, and 'transcriber selectivity', we propose that an automated transcription procedure might (a) save a lot of time and (b) be more detached from the context. Of course, the latter point is potentially the more contentious. On the one hand, a 'dumb' algorithm would, by definition, refrain from introducing subjectivity into the process,[1] but on the other hand, it is the context that provides meaning to a

transcript. We follow this path in the understanding that a researcher always will need to follow up any data processing phase, like transcription (whether automated or not), through application of professional judgement. Kvale (1996: 163) raises the issue of transcriber reliability, indicating that in social science research transcribers can employ different styles and rendering of the transcription wording. Kvale (1996: 174) adds that any attempt to include non-verbal cues in the transcript, such as indicators of tone, mood and pauses, or other responses such as laughter or giggling would only serve to exacerbate issues of 'intersubjective reliability'. It is likely that the production of such detailed and high-quality transcripts would require multiple phases of transcription to layer in detail and provide opportunities for checking coverage and content. In a large-scale study utilising qualitative data collection methods, involving multiple members in the research team, transcription can produce thousands of pages of transcribed material (Hunt et al., 2011), but even though the authors specifically report the challenges of managing large-scale qualitative datasets, virtually no space is given to the discussion of the process of transcription other than the need to capture separately the interviewer's perception of a respondent which can be lost in steps as early as the transcription process. Hunt et al. (2011: 9–10) point out that all members of the team, including 'all senior researchers' will read significant numbers of full transcripts and go back to listen to the raw audio recordings in the process of analysis.

Given these challenges, we assert that transcription in the research process will always be a trade-off between available time or means, and the quality of the transcript. A 'better' transcript, with 'better' being defined as the most complete and trustworthy account of a media file, will be more costly. Given this trade-off, perhaps the automated transcription of audio can assist as a first or early step in the process, providing a sufficiently 'good enough' first version. The use of such automated services could serve as a useful 'first draft' transcription of audio data that would then form the foundation for what Paulus et al. (2013) refer to as 'cycles' or 'rounds' of transcription, which usually require the transcriber to engage with the audio recording on multiple occasions in order to capture all the required elements from the raw data (pp. 96–97). We believe construction of a 'first draft' would also be relevant for a range of transcription types, whether the aim be production of a verbatim transcript (capturing features of speech and non-speech), transcription types utilising a notation system (e.g. Jefferson, 2004) to represent specific features of recorded talk, or even for condensed or other gisted transcripts, allowing the researcher to make choices as to what features to omit from the transcript such as repetitions, false starts and other features of natural speech (for types of transcription see Paulus et al., 2013: 96–101). Auto-captioning applications are one means by which researchers can gain access to sophisticated tools for the automated recognition of speech. We have focused on this route to obtaining a 'first draft' transcript because auto-captioning is widely and freely available as part of the panoply of resources that support making the huge volume of audio-visual content available on the web more accessible to diverse audiences. The process of auto-captioning can also carry with it some additional technical advantages that we believe a researcher may be able to exploit in order to enhance the utility of the resulting transcripts. We will discuss these additional benefits in more detail towards the end of the article.

Literature on automated captioning

The production of automated captions for videos with audio tracks, entail the automatic recognition of speech in the audio data, providing automatic subtitles for the audio belonging to a video. The literature on auto-captioning can be divided under two broad themes, namely literature related to various technical aspects of automated captioning, and literature focusing on the use of captioning to support students with additional educational needs, and thus supporting teaching and learning.

Technological developments. In their brief review of the history of ASR, Juang and Rabiner (2004) describe how as far back as the 1930s Bell Laboratories proposed a model for speech analysis and synthesis. Major advances in the statistical modelling of speech in the 1980s led to widespread application of ASR in situations where a human–machine interface was needed (Juang and Rabiner, 2004). With ever increasing technological improvements (for an overview, for example, see Ramírez and Górriz, 2011) software solutions became available. Initially, this was in the form of stand-alone software like Dragon Naturally Speaking or IBM's ViaScribe, later as part of other mainstream, sometimes web-based programmes. By 2009,[2] the first version of video provider YouTube's captioning system was introduced, exemplifying a trend towards online solutions, like IBM's hosted transcription service. The development of speech recognition systems has been rapid since then. All major commercial speech recognition systems are based on deep learning (e.g. see Deng and Yu, 2014). In recent years, both Microsoft and IBM, with 5.9% and 5.5%, respectively, approached the word error rate for humans, seen to be around 5% (Fogel, 2017; Tarantola, 2016). These advances have also influenced captioning functions in YouTube. Deep learning algorithms, for example, in 2017 for captioning sound effects (Chaudhuri, 2017), have been used to further improve YouTube's speech recognition quality. Whenever a new video is now uploaded to YouTube, the new system runs and tries to identify captions, including sounds. Given these rapid technological developments it was expected that perhaps the initially sub-optimal experiences with speech recognition as a tool for transcription of audio might have improved.

Methods of obtaining captions. There are several popular ways to obtain captions. A first option is to ask professional companies to do this. This takes substantial time and is often accompanied by considerable costs (Dubinsky, 2014; Johnson, 2014). A second option, used for more than a decade through tools like Media Access Generator (MAGpie), Subtitle workshop and Amara, is to manually make a subtitle file that can be used in combination with the video. Recently, YouTube has managed

to integrate these features within their own service. Finally, there is the option of using auto-captioning services, with YouTube year-on-year improving the quality of this feature. Another example is Synote (Wald, 2010) interfacing with text-to-speech functionalities. In this scenario, text-to-speech software generates the captions without human intervention (Fichten et al., 2014). This option is the quickest option available, but there is debate on the accuracy of this approach (Bennett et al., 2015; Parton, 2016). As technology progresses, it can also be expected that accuracy will increase. Furthermore, *the degree* of accuracy required from transcription of audio data depends upon the nature of the transcript and their purpose.

Accuracy of captions. Nevertheless, even for a first draft, accuracy might be a problem (Johnson, 2014), but the reaction to the severity of the inaccuracy is mixed, ranging from 'devastating' (Anastasopoulos and Baer, 2013), 'a barrier to communication' (Parton, 2016), 'humorous' (Clossen, 2014) to 'a fairly good job' (Suffridge and Somjit, 2012). An often heard recommendation is to start with auto-captions and then edit to reduce the number of errors and fix any timing issues (Clossen, 2014; Johnson, 2014). In most cases, it is acknowledged that although manual transcription might be superior, it is not realistic to think the same amount could be done, because of time and money constraints. Automated captioning can fail to accurately convey the intended message (Barton et al., 2015; Johnson, 2014). Recently, however, it has also been noted that performance of the relevant algorithms seems to improve (e.g. Liao et al., 2013). One way to further speed up and improve the accuracy of the transcription process for interviews is to listen to the interview and repeat what was said using voice recognition software (VRS) that has been trained to recognise a specific voice. This is the method used for live subtitling/captioning for television and court reporting as well as for supporting deaf people in meetings. The process is known as 'respeaking', 'shadowing' or 'parroting', but still involves at least the same time the recording lasts. In this article, we hypothesise that the accuracy of automated captions might support the transcription process.

Supporting teaching and learning. The potential of automated captioning to support teaching and learning for students with special educational needs, including second language users, is well recognised (e.g. Collins, 2013). Captioning can be seen as supplementing video-based materials, for example, in the context of a foreign language instructional tool (Dahbi, 2004). This can also be a selection of words for captioning, so-called 'key word captioning'. Students' understanding of the video content can be increased, even if the complexity of the captioned key words surpasses the reading level of the student (Ruan, 2015). A study of deaf students (Shiver and Wolfe, 2015) suggested that a large contingent of students preferred to watch videos with automated captioning than with no captions. Parton (2016) studied the use of captions in relation to deaf students. She notes the role captions play in improving accessibility of video resources, highlighting the legal obligations that (higher) education institutions have to make materials accessible. Lewis and Jackson (2001) have demonstrated that script comprehension of deaf and also hearing impaired students was greater with captioned videos. Bain et al. (2005) describe key advances in audio access that have occurred since 2000, mentioning the intention to create real-time access for students who are deaf and hard of hearing, without intermediary assistance. They utilise a tool called Viascribe to convert speech recognition output to a viable captioning interface. Federico and Furini (2012) also focussed on students with some form of additional need (e.g. hearing impaired, dyslexic and English as a Second Language (ESL)), proposing the use of off-the-shelf ASR software. Wald (2005) seems to broaden the target audience, acknowledging that ASR can 'assist those who require captioning or find notetaking difficult, help manage and search online digital multimedia resources and assist blind, visually impaired or dyslexic people by augmenting synthetic speech with natural recorded real speech' (p. 1) or even more general 'anyone who needs to review what has been said (e.g. at lectures, presentations, meetings etc.)' (Wald and Bain, 2007: 446). Ranchal et al. (2013) also extend the potential benefits of ASR for students who have difficulty taking notes accurately and independently, particularly for non-native English speakers and students with disabilities.

A proof-of-concept: voice and speech recognition tools in the literature. There is limited literature on the use of VRS and ASR tools to aid transcription. Any attempts to utilise VRS tools on raw audio from multivoice interviews usually result in expression of exasperation due to woefully low accuracy rates (Dempster and Woods, 2011; Dresing et al., 2008; Evers, 2011). Some authors have utilised an approach in which the researcher simultaneously listens to the original voice recording while dictating/reciting into VRS trained to recognise the researcher's own voice, in a manner analogous to the 'respeaking' approach to captioning described above. Such articles usually report very small-scale, personal comparisons of the time commitment required to conduct the VRS dictation method versus standard listen-and-type transcription approaches (Matheson, 2007). Johnson (2011) has argued that traditional listen and type requires less time than simultaneous dictation via VRS. To further support the transcription process, sometimes custom applications are used; for instance, Roberts et al. (2013) managed data in Synote, a freely available application enabling synchronisation of audio recordings with transcripts and coded notes.

Some researchers (e.g. Evers, 2011) have questioned the need to transcribe audio recordings at all, now that it is possible to add analytical codes directly onto raw digital files for all sorts of media, including audio files, using tools such as ATLAS.ti, MAXqda, NVivo and Transana. Nevertheless, even proponents of bypassing transcription such as Evers (2011) report issues with alignment of codes to specific data segments in the audio (or video) files as this is much harder than coding segments on a typed transcript. She also relates issues with reduced opportunity for reflection that arises when coding directly onto the raw audio file compared with the multiple stages of transcription

and analysis. Conscious of a 'generation effect' in terms of exposure to technology, Evers (2011) asked her students, as well as colleagues to compare the experience of traditional transcription versus direct coding onto audio files. While the students were vocal about the benefits of time saved by not requiring a transcript, and also the closeness that they established with the participant's voice when coding directly, they complained about problems of losing contact with sections of data that they never listened to a second time and a perception of sloppy rephrasing of a participant's words. Some of the students even resorted to producing traditional transcripts for at least some sections of the data. Also noteworthy were indications that the visual nature of analysis of transcripts aids the researcher in tracking the analytical process, which is not (currently) replicated by the practice of direct coding onto audio file segments. The searchable nature of transcripts was frequently given as a key benefit over direct coding onto the audio file.

Ranchal et al. (2013) have studied the use of ASR tools in the context of lecture capture. In their study, they measured the accuracy of the ASR technologies using *word error rate* and *recognition accuracy* tools. Even after voice profile training on ViaScribe ASR software, they were only able to achieve accuracy rates close to 80%. Using a speaker independent post lecture transcription tool (IBM Hosted Transcription Service) that utilised a double pass approach, accuracy was increased to between 85% and 91% (Ranchal et al., 2013: 307). They then employed grad student teaching assistants to correct word errors in the resulting automated transcripts. An automated transcript with a word error rate of over 20% still required a teaching assistant unfamiliar with the course materials to spend up to 4 hours per hour of lecture audio to correct word errors (pp. 306–307).

Kawahara (2012) reports on a bespoke speaker-independent ASR system developed for the production of transcripts of plenary and committee meetings of the Japanese Parliament. The committee meetings are particularly challenging as they consist of multiple voices engaged in free speaking with interaction between speakers. Despite these challenges, the system is reported as consistently producing accuracy levels of at least 85% and, more commonly, approaching 90% for committee meetings and over 95% for plenary sessions. Parliamentary reporters further process the transcripts to correct errors in the automated transcripts using a post-editor tool that the reporters helped the development team to design. The resulting editing process produces a searchable archive file that consists of the transcribed text, together with the raw speech audio and video files that are aligned and hyperlinked by the speaker name and the words uttered (Kawahara, 2012: 2227). Kawahara indicates that the accuracy of the ASR tool is regularly monitored, and the lexical and language models used within the ASR are revised annually by the same parliamentary reporters who edit the automated transcripts.

The aim of the proof-of-concept described in this article is to consider whether we might utilise the functionality of freely available automated captioning services to save time in the research process, especially in the transcription process for audio recordings. To our knowledge, there has not yet been such a direct application of automated captioning to support the laborious and time-consuming transcription process in a research context.

Methodology

The methodology section tries to faithfully present the complete procedure of automatically transcribing three types of audio recording: two one-to-one interviews, a group interview and a recording captured as part of the observation of a lesson taught in a school. Note that, in this context, we use the terms 'audio' and 'video' interchangeably; it denotes how even when we only analyse audio, the described procedure requires uploading a video. However, this can be achieved by the addition of a single static image for the duration of the audio recording. This is a practice commonly used to enable audio files to be uploaded to YouTube.

Collecting the data

The three data sources used for this proof-of-concept were publicly available resources, for which we gained secondary data ethics approval from the University's ethics board (number 26617). Ethical aspects of the methods described in this study always need to be taken into account, as we discuss further towards the end of the article. We will refer to the three sources as T, C and I.

Source T consists of two one-to-one interview videos used in one of our methodology courses available on our department's YouTube channel.[3] The videos (with audio) contain interviews with two former teacher practitioners. The videos were downloaded in mp4 format and were captured using studio-quality radio lapel microphones with audio captured in Dolby Digital (AC3) format with data rates of 256 kbits per second.[4] Existing verbatim transcripts of each interview were also available, that had been produced manually by a teaching assistant. Figure 1 shows a fragment of the format of the existing transcript.

A: And how did you get into being a head teacher?

B: That's a good question, I did think about that and I am not sure how I into it really, I suppose I just I have always been really keen on making, doing the best job I can do and I suppose at the time which is a few years ago now being the best at what you could do meant that you got promoted and I suppose I sort of got there by default really.

Figure 1. fragment of the existing transcription (anonymised).

Source C consists of a classroom video from the TIMSS 1999 video study, downloaded from the TIMSS website.[5] The TIMSS study focused on grade eight mathematics and science teaching in seven countries, in which national samples of teachers were videotaped teaching an eighth-grade lesson in their regular classrooms. The website allows the download of both mp4 format videos, as well as transcripts in text format. The mp4 video of the US1 lesson was downloaded, consisting of audio and video. US1 is an USA eighth grade mathematics lesson which focuses on graphing linear equations, is 44 minutes in duration, and with 36 students enrolled in the class. The audio contains teacher talk, group dialogue and a fair amount of background sound. The transcripts were produced manually, based on protocols in a transcription manual.[6]

Source I consists of a video from the Chilcot Iraq Inquiry in the form of an interview with General the Lord Walker of Aldringham.[7] The website also contains a transcript of the interview in PDF format, see Figure 2, which was stored as a text file. The exact method of transcript creation for this source is unknown; the protocols of the Inquiry seem inconclusive.[8] However, as the final transcript on the website has been validated, we assume it is deemed a trustworthy account of the hearing.

(2.00 pm)

GENERAL THE LORD WALKER

THE CHAIRMAN: Good afternoon.

GENERAL THE LORD WALKER: Good afternoon, Sir John. Hello.

THE CHAIRMAN: Good afternoon ladies and gentlemen. Good afternoon, Lord Walker. In this session, we are hearing from General the Lord Walker, former Chief of the Defence Staff and, before that, Chief of the General Staff. The objective of the session is to examine the key discussions and decisions taken on Iraq and their

Figure 2. fragment of the PDF transcript of the interview with General the Lord Walker of Aldringham.

Applying automated captioning

The auto-captioning and caption-processing tools described below are all freely available, web-based tools. For the purpose of implementing the proof-of-concept videos T, C and I were uploaded to the private YouTube channel of the first author with the option of 'automatic captions' in English selected. Figure 3 shows an impression of the YouTube interface for uploading Source I.

After uploading the videos were left for a couple of hours to let the captioning engine create automated captions. Through websites that allow the downloading of these captions, like http://mo.dbxdb.com/Yang, YouTube themselves and www.diycaptions.com, the transcripts were downloaded. In most cases, captions could be downloaded in two formats: a 'text only' format, and a time-stamped subtitle file, often with the file extension .srt as shown in Figure 4.

To make the text files comparable to the existing manually produced transcripts, decisions had to be made as to what elements of the captioning to include. The caption file (a so-called .srt file) includes timestamps, for example, and the original transcripts include names or initials of the speakers. With a programme called 'Subtitle Edit 3.5.2'[9] the timestamps and durations were removed prior

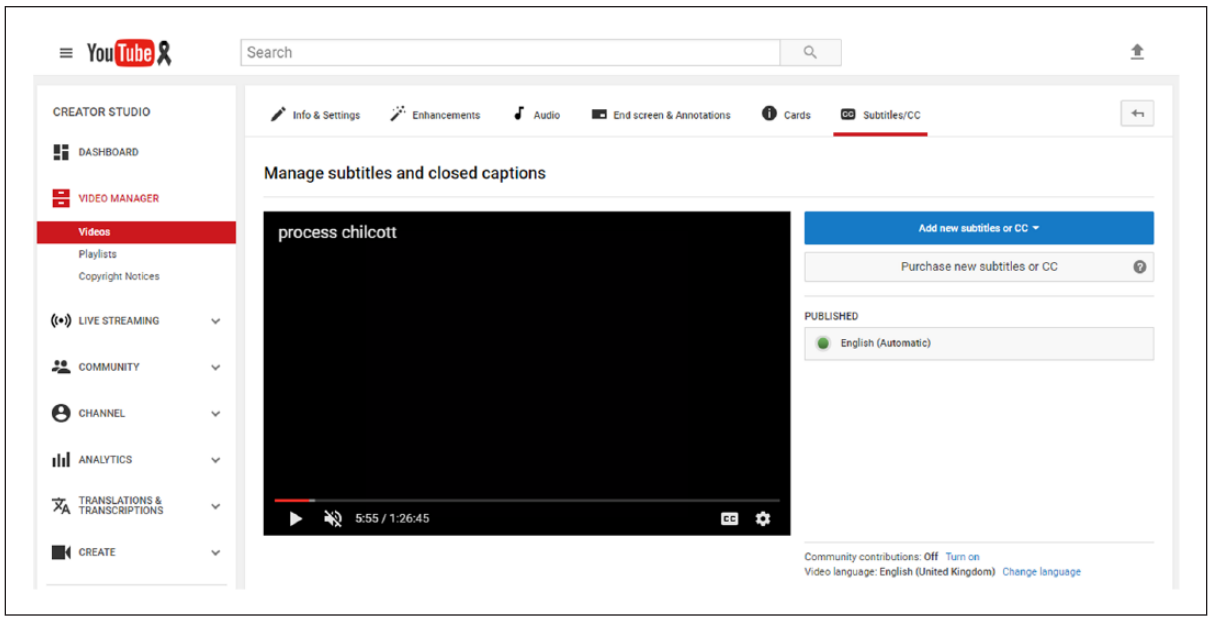


Figure 3. screenshot of the YouTube interface for captions.

to measuring the match between documents. In a real transcription setting, it might be a useful addition to the transcript to retain the timestamps to aid reconnection with the raw audio file. The inclusion of timestamps might facilitate connection between elements of the transcript and the corresponding section of raw audio file in the way that computer assisted qualitative data analysis software (CAQDAS) tools have increasingly made possible (Paulus et al., 2013: 99). This can be particularly useful when constructing specific forms of transcript that require multiple cycles of engagement with the raw audio file to capture the level of detail required in the transcript, or to determine what information might be omitted in a gisted form of transcript.

```
3
00:00:07,660 --> 00:00:13,270
interviewed today my first question is

4
00:00:10,270 --> 00:00:15,840
have you always been a teacher always I

5
00:00:13,840 --> 00:00:18,549
only ever wanted to be a teacher when I
```

Figure 4. fragment of a time-stamped sub-title file, obtained from YouTube.

For the core text from the automated transcripts, plain text-only versions were created, as demonstrated in Figure 5. We were aware that there most certainly would be some differences between automated and manually produced transcripts in formatting, but accepted this as 'less-than-perfect' outcomes that could simply arise from formatting issues. For each of the data sources, T, C and I we will present quantitative measures of the similarities between automated and manually produced transcripts, together with some qualitative description of the differences.

Results: comparing text similarity

There is a large variety of tools available to compare text similarity. For this pilot, we initially used the open source, windows-based WCopyfind 4.1.5 which is an open source windows-based programme that compares documents and reports similarities in their words and phrases.[10] One challenge with this software is that the settings should be tweaked to get a good match. To illustrate this, using default values,

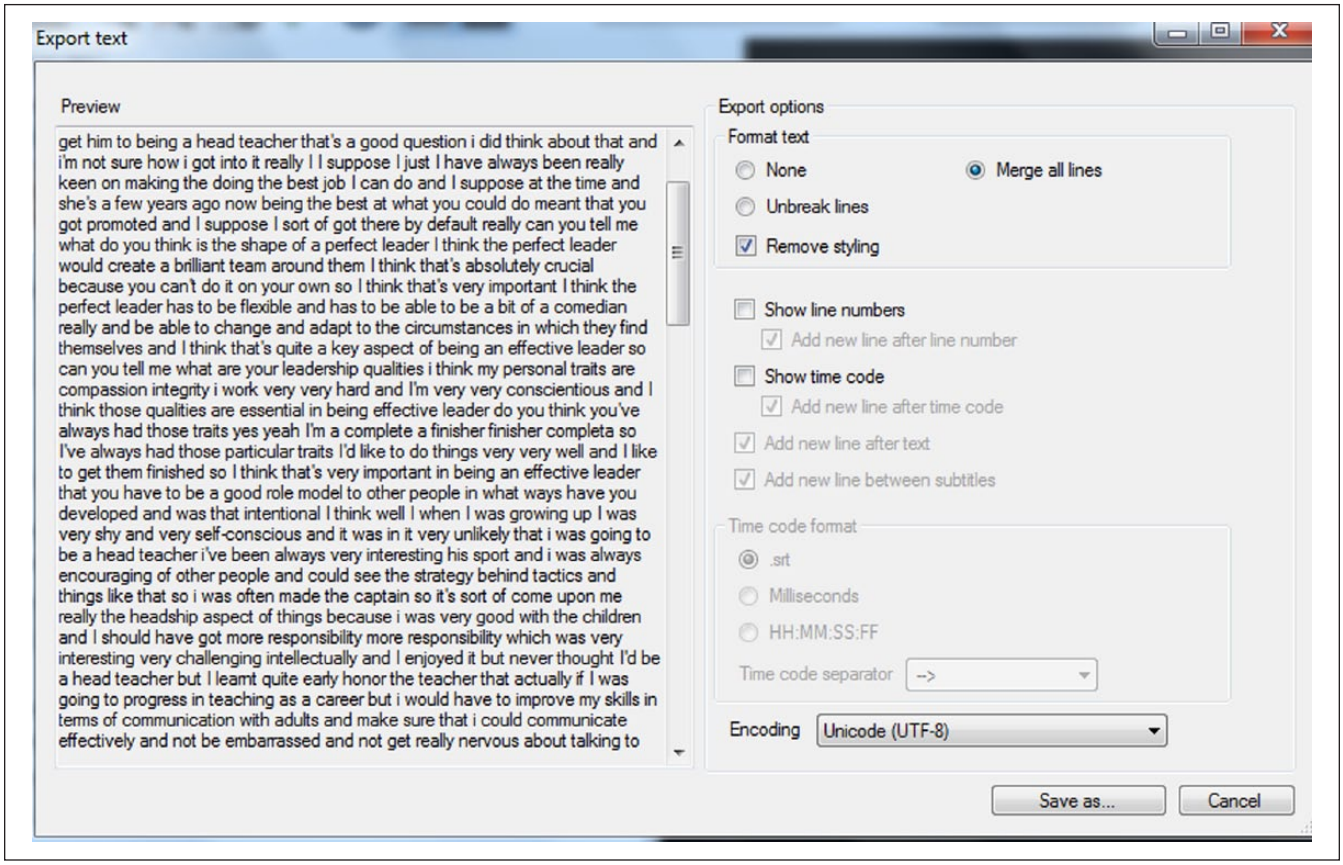
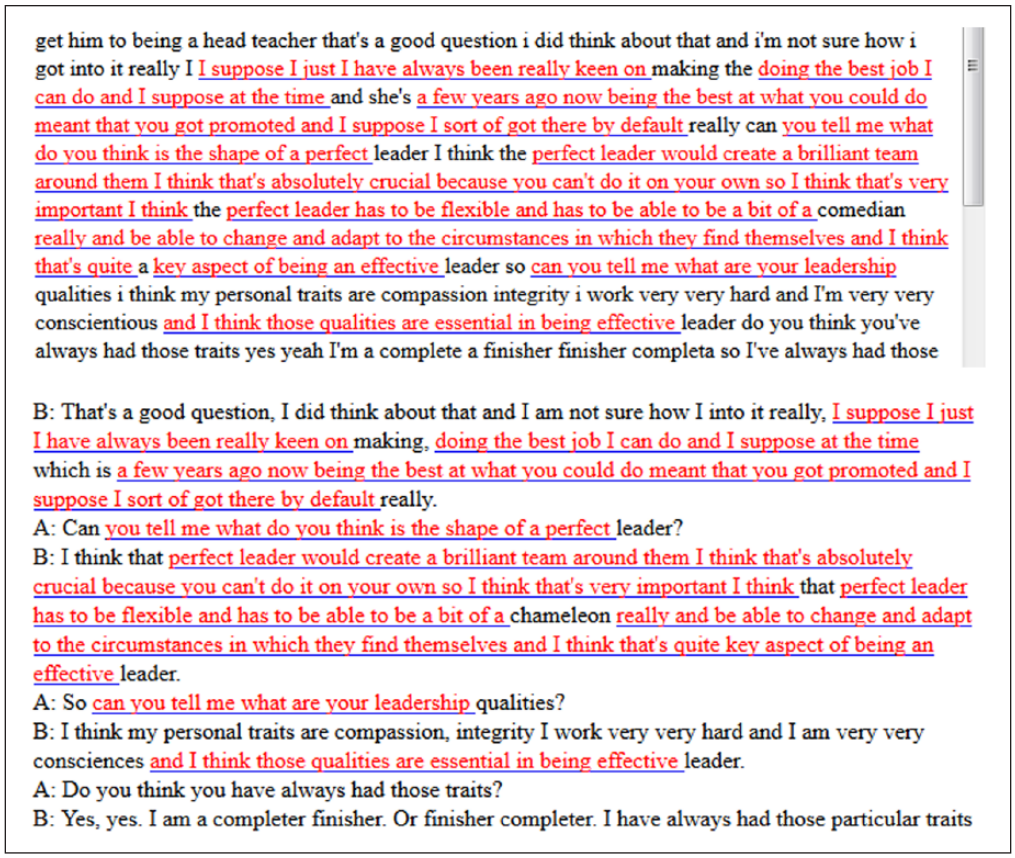


Figure 5. exporting the automated transcript in plain text format.


get him to being a head teacher that's a good question i did think about that and i'm not sure how i got into it really I I suppose I just I have always been really keen on making the doing the best job I can do and I suppose at the time and she's a few years ago now being the best at what you could do meant that you got promoted and I suppose I sort of got there by default really can you tell me what do you think is the shape of a perfect leader I think the perfect leader would create a brilliant team around them I think that's absolutely crucial because you can't do it on your own so I think that's very important I think the perfect leader has to be flexible and has to be able to be a bit of a comedian really and be able to change and adapt to the circumstances in which they find themselves and I think that's quite a key aspect of being an effective leader so can you tell me what are your leadership qualities i think my personal traits are compassion integrity i work very very hard and I'm very very conscientious and I think those qualities are essential in being effective leader do you think you've always had those traits yes yeah I'm a complete a finisher finisher completa so I've always had those

B: That's a good question, I did think about that and I am not sure how I into it really, I suppose I just I have always been really keen on making, doing the best job I can do and I suppose at the time which is a few years ago now being the best at what you could do meant that you got promoted and I suppose I sort of got there by default really.
A: Can you tell me what do you think is the shape of a perfect leader?
B: I think that perfect leader would create a brilliant team around them I think that's absolutely crucial because you can't do it on your own so I think that's very important I think that perfect leader has to be flexible and has to be able to be a bit of a chameleon really and be able to change and adapt to the circumstances in which they find themselves and I think that's quite key aspect of being an effective leader.
A: So can you tell me what are your leadership qualities?
B: I think my personal traits are compassion, integrity I work very very hard and I am very very consciences and I think those qualities are essential in being effective leader.
A: Do you think you have always had those traits?
B: Yes, yes. I am a completer finisher. Or finisher completer. I have always had those particular traits

Figure 6. Comparison of output of automated transcript (top) and the original transcript (bottom).

Figure 6 shows that there is more overlap between the automated transcript at the top and the original transcript at the bottom, than the red text indicates.

For example, at the top 'That's a good question' appears in both transcripts but is not flagged up as equal. The same applies to 'Do you think you have always had those traits?' The differences can be explained by the numerous options the software has to compare texts, and the various ways in which slight differences can come up as 'different'. Examples of what WCopyFind can take into account in determining what constitutes a difference are as follows: punctuation, numbers, cases, non-words, word length, and more. Despite this, comparison of the automated and manual transcripts showed 69% and 64% similarity, respectively. Nevertheless, for a better comparison, we turned to different software, as described in the next section.

The two interviews (source T)

To obtain a more sophisticated comparison we utilised the well-known plagiarism detection software from Turnitin (2017). First, the text from the automated transcript was uploaded, and after that the manual transcription. We made sure that no other sources were counted in the percentage match.

Table 1. Similarity between automated and manual transcripts for two interviews (pseudonyms used).

	Adams	Barnett
Word count automated	934	1816
Word count manual	947	1817
Turnitin % similarity	91%	92%

Table 1 shows that both automated and manual transcripts show a very high level of agreement. Further scrutiny of the comparison in Turnitin showed that many discrepancies were caused by relatively minor typos in the original, such as incorrect automated transcription of domain-specific words and names, as demonstrated in Table 2. This list is not meant as an exhaustive analysis of the transcripts but as demonstration that many of the errors actually are quite small and easily rectifiable. Small differences also occurred the other way, that is, that Turnitin would not flag them up while they were different from the recording.

It can be observed that the differences are minor, and mainly concern easily rectified issues or aspects that manual transcripts from human transcribers might not necessarily result in the optimal transcript.

The TIMSS video study (source C)

The results of the similarity check of transcripts derived from the TIMSS classroom study video are presented in Table 3. These are also quite favourable but with a similarity match of 68% the result is not as good as that achieved for the interview transcripts from Source T.

Table 2. Selection of qualitative differences between the manual and automated transcripts for (T). The 'correct' interpretation is indicated in italics.

Manual	Automated	Comment
Staid	*Stayed*	The automated transcript did not contain language typos or incorrectly interpreted words.
Carrier	*Career*	
Consciences	*conscientious*	
I am	*I'm*	This means the same but is picked up as 'different'.
Head teacher	*Headteacher*	The similarity check saw the two as different.
Completer	Complete a	The automated process confused some of the sounds. The interviewee used the word 'completer' in the context of 'being someone who completes things'.
	Text added by software for downloading the sub-titles	The similarity check saw these as textual differences.
This text had initials of the speakers		The similarity check saw these as textual differences.

Table 3. Similarity between automated and manual transcripts for the classroom video.

	US1 lesson
Word count automated	5659
Word count manual	5830
Turnitin % similarity	**68%**

Table 4 gives a selection of the types of differences in the two transcripts. It can be observed that the differences are more substantial than with the interview transcripts (Source T). Some sounds clearly have not been picked up correctly by the captioning algorithm. Another challenge lies in specific domain knowledge, for example, the mathematical content (numbers, especially fractions) that are not picked up. Surprisingly, the predictive algorithm did manage to correctly transcribe domain-specific terms such as 'y-intercept'.

The quality of the audio for Source C is inferior to that of source T, but given the fact that the audio recording is from a single microphone located in a busy classroom environment from a recording made using equipment and technology available in 1999, the resulting transcript seems decent. In these examples, the colloquial language sometimes adopted in classroom dialogue seems to be a challenge for ASR software.

Chilcot recording (source I)

Finally, the Chilcot recording, by far the longest of the recordings, also showed an almost two-third similarity match between the manual and automated transcripts, as indicated in Table 5.

However, the original manual transcript was by far the most contextualised data in that it followed a standard reporting convention for the inquiry proceedings. For example, the lead names indicating who said what were systematically included in the text of the manual transcript, as were line numbers. A qualitative comparison of the first pages indicated similar challenges for the ASR as those observed in the

Table 5. Similarity between automated and manual transcripts for the Chilcot interview video.

	Chilcot
Word count automated	14,187
Word count manual	16,587
Turnitin % similarity	**66%**

Table 4. Selection of qualitative differences between the manual and automated transcripts for (C). The 'correct' interpretation is indicated in italics.

Manual	Automated	Comment
Three page	free paste	The automated process confused some of the sounds.
You're	You	The automated process confused some of the sounds.
Every one of you	everyone you	The automated process confused some of the sounds.
There	Here	The automated process confused some of the sounds.
The –use the ruler, Robert. Use the ruler man. Make it neat. All right?	his easily make it mean when you put like a one in here what you're like	This is completely different. The predictive algorithm did not understand this at all.
Numbers e.g. 2 thirds or 2/3		The automated algorithm does not process numbers well.
try with this one	Tribalism	The automated process confused some of the sounds.
You've already forgotten	You void forgotten	The automated process confused some of the sounds.

Table 6. Selection of qualitative differences between the manual and automated transcript for (I). The 'correct' interpretation is indicated in italics.

Manual	Automated	Comment
Their	they're	The automated process confused some of the sounds.
In adjusting	into justing	
To imprint inary	Two preliminary	
Recognise	recognize	American-English spelling differences
What we hear	won't be here	This example was tabulated separately because errors can of course change the meaning completely.
Freedman	Friedman	Names are problematic, although it did pick up a name.
But of course, that was in 1991, so one might have hoped that, by 2003, those sorts of problems had been overcome.	because that was in nineteen one no one might have hoped that by 2003 the problems you know to come well	As with spoken numbers, years present problems here, specifically the tendency in speech to split a four digit year into a pair of two digit numbers.
SIR LAWRENCE FREEDMAN: Did you feel that it was creating risks for your forces? GENERAL THE LORD WALKER: It was obviously creating some risks, but we are used to dealing with risk.	*did you think this was creating risks for your forces well it was obviously creating some risks but I think I mean we 're used to dealing with risk*	The meaning of these two statements is similar but it demonstrates how the names in front of the statements result in a difference being flagged up.

comparison of the classroom study transcripts (Source C), as Table 6 demonstrates.

Although there were more differences between transcripts in this case than with those derived from the interview and classroom settings, here again many of the difference observed do not seem particularly problematic in terms of their impact on meaning, especially to an informed reader such as a researcher moving to the process of reviewing the transcript, or even to the process of coding the data. All three data sources seem to confirm that with minor effort one might be able to obtain reasonable 'first version' automated transcripts through the use of freely available web services.

Conclusion and discussion

This article set out to explore, as a proof-of-concept, whether it would be possible to use freely available features of web-based tools for automated captioning to produce automated transcripts of audio and video recordings. Our conclusion is that this indeed is possible and that this results in a reasonable first version of a transcript. If we take the conservative estimate that about two-thirds of the transcript, without any editing, can be obtained with a couple of minutes of uploading, a few hours of waiting time (that can, of course, be devoted to other tasks), and a minute of downloading, it is clear that the time savings should be substantial. For high-quality audio in optimal settings such as those used for one-to-one interviews the percentage match to manually produced transcripts can be even higher, surpassing 90%. Even with the possibility of such high accuracy rates, indicated by percentage of matching text discussed here, we are certainly not suggesting that auto-captioning would be the end of the transcription process; rather, it would facilitate it.

Saving time in the research process

In large-scale research projects generating a substantial volume of audio recording, outsourcing the process for the production of a first version transcript, is more commonplace. Even with outsourcing the process of transcription to highly experienced transcribers, it is by no means clear that the quality of the transcripts will be perfect. After all, audio is not always clear and transcribers recruited from outside the research team are likely to be unfamiliar with the contextual information of any audio recording, making transcription of some domain-specific sections of audio data particularly challenging. It is not unusual for members of the research team to review and edit externally produced transcripts as a quality check and also to aid in establishing closeness to the data. Indeed, the production of some forms of transcript (e.g. Jeffersonian) are considered to require several cycles of engagement with the raw audio data in order to capture the range of information required in the transcription (Paulus et al., 2013). One might argue that the editing of a first version automated transcript, produced using auto-captioning tools, would be analogous to the review that would normally be carried out on externally produced transcripts, and that even for smaller scale research projects the benefits of time saved in transcribing might be usefully invested in such a review process. Compared to outsourcing transcription, the automated process would also be very cheap, and is currently free, based on available tools. Some tools, like aforementioned www.diycaptions.com even provide an interface, not unlike existing CAQDAS tools, to manually correct any outstanding errors. Notwithstanding the evolution of the technology used, we think that considering the appropriateness of an automated process *first*, before going through the traditional, lengthy process, can be one of the enduring principles of our article.

Issues and advantages arising from automated transcription

This study highlights several challenges. First, as flagged up before, there are circumstances under which the automated transcription works less well. If the expectation is that one obtains perfect transcripts, then there will be disappointment. As the examples have shown, errors are more likely to occur with given names, differences in formatting and domain-specific terms. These errors all seem to relate to the complexity and quality of the audio recording. Factors that seem to influence this are as follows: the number of different speakers, the way turn-taking is organised (i.e. interruptions vs sequential speech), accents of speakers, how colloquial the language is, and even the audio quality of the recording. From a technical point of view, these challenges have been known for a while. For instance, Kawahara et al. (2003) acknowledged that the performance of ASR is affected by factors such as acoustic variation caused by fast speaking and imperfect articulation, linguistic variation such as colloquial expressions and disfluencies. Byrne et al. (2004) suggested that 'adequate accuracy … of spontaneous conversational speech' could be obtained (p. 433). In the last decade, as we have reported in the literature section, accuracy of the ASR has improved considerably, and we suggest that automated transcription is starting to become a feasible solution for research purposes: 'good enough' transcripts that can then be perfected with far less manual labour than before. A second challenge, in this article, might concern our method of determining the similarity between documents, which relies on Turnitin. The Turnitin algorithms are proprietary; therefore, we can't say on what basis the similarity is calculated. Nevertheless, as Turnitin is used widely to check similarity for the purpose of academic integrity checks, we feel it is reasonable to assume that the algorithms give a fair indication of the similarity. As indicated before, simpler methods to measure similarity between text extracts like those using the default values in WCopyFind, still yield percentage matches up to 70% for the best case interviews (Source T). In our view, it is reasonable to see this as a minimum, as we also established that this comparison method did not flag up some of those similarities. In the context of producing 'good enough' first transcripts, we think our approach sufficiently supports our conclusions regarding similarity. In our view, automated transcription, whatever technology used, should be a viable option for any researcher, with a manual data check still always in place.

Future developments for automated transcripts

As well as the auto-captioning tools described here, other freely available tools might offer possibilities for the generation of automated transcripts. Voice recognition software associated with the operating systems running on common mobile computing devices such as tablets and smartphones are increasingly able to render speech to text without any of the training to a specific voice that was required by older VRS systems. These systems would need to be coupled to playback of an audio recording, to generate automated transcripts as such technologies work by 'listening' to a file in real time as opposed to processing the complete digital file in the way that the auto-captioning tools process the audio data. The VRS tools built into common operating systems for mobile devices currently have to buffer after a set period of 'listening' to speech in order to process the acquired data, which would make the process awkward and time consuming, but it may be that the application programming interfaces (APIs) that provide the VRS functionality can provide real time processing in parallel with the process of data capture via 'listening' to the playback of a recorded audio file. As well as processing, the complete digital file rather than capturing data by 'listening' in real time, another advantage of the captioning software described here is that it can automatically add timestamps to a transcript, which might then be used to facilitate matching between the transcript and the raw data at specific points in the audio file, demonstrated by the utilisation of subtitling files.

Automated transcriptions might cover different languages; already YouTube's automated captions are available in English, Dutch, French, German, Italian, Japanese, Korean, Portuguese, Russian and Spanish. Not all of the languages are as accurately captured as the examples in this article, but with recognition techniques improving, different languages become available for that first 'good enough' transcription. Although the recognition of the lower quality TIMSS video (C) already was fairly high, it might be expected that the recognition quality might only increase and the higher quality audio derived from equipment available in recent years, exemplified by the audio recordings of the one-to-one interview sources used in this proof-of-concept, indicate that raw data using contemporary audio quality standards combined with automated transcription might yield very high-quality first transcripts in best case scenarios. Finally, another recent development is the integration of social networking tools into the captioning services for crowdsourcing captions that might yield higher quality outputs. Given the technological strides of the last decade, it is envisaged that the possibilities will only improve; another reason for us to suggest that it at the very least automated transcription should be considered as an option in the research process.

Ethical considerations

Finally, it is also important to consider ethical aspects, for example, regarding data protection and security issues. Captioning services, whether embedded in a tool like YouTube, or providing support for other tools, might store data on computers other than institutional ones. According to the Economic and Social Research Council (ESRC) these aspects form part of data protection and should be reviewed regularly (ESRC, 2017). Specifically for subtitles, a poignant example of a security issue, are the recent hacking incidents

that used so-called 'subtitle servers' to break into computers (Biggs, 2017). There also are ethical considerations regarding the location and safe storage of data on third parties servers: do research council rules allow storage of personal data on these? This has become particularly relevant following recent developments with Facebook and Cambridge Analytica (e.g. Greenfield, 2018) and the General Data Protection Regulation (GDPR), a regulation in EU law on data protection and privacy for all individuals within the European Union (EU) and the European Economic Area (EEA). One consequence of this is that data must not be available publicly without explicit, informed consent, and cannot be used to identify a subject without additional information stored separately. Of course, in the context of YouTube, even as unlisted or private videos, this means ensuring that these requirements are met. In general, it is important that researchers are fully cognisant of the ethical and security implications of the research and analytical methods they use. Whether using primary or secondary data with these tools, university ethics committee approval should firmly be in place, to ensure sufficient consideration of the ethical aspects. As stated before, we also did this for this study. A more sociological and ethical aspect might also be seen in the relation with the 'future of work' and the labour involved in transcription. Srnicek and Williams (2015), for example, argue that the crisis in capitalism's ability (and willingness) to employ all members of society, should lead to investment in labour-saving technologies. They envisage a positive feedback loop between a tighter supply of labour and technological advancement. In that light the wider implications of our 'work saving' propositions could be considered.

Notwithstanding these challenges, we suggest this study has shown the promise of automated generation of 'good enough' transcripts. In our view, it would be a 'common sense' approach to analyse qualitative data at scale.

Declaration of Conflicting Interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Note that some point out that the algorithm might not be biased but the programmer of that algorithm, for example, see O'Neill (2016).
2. https://googleblog.blogspot.co.uk/2009/11/automatic-captions-in-youtube.html
3. YouTube: www.youtube.com
4. Radio mics were Sennheiser EW 100-ENG G3 and the camera used to capture the video and audio was a Panasonic AVCHD.
5. www.timssvideo.com
6. https://timssvideo.squarespace.com/s/Transcription-Manual.pdf
7. http://www.iraqinquiry.org.uk/the-evidence/witnesses/w/general-the-lord-walker-of-aldringham/
8. http://www.iraqinquiry.org.uk/the-inquiry/protocols/witnesses-giving-evidence/
9. SubTitle Edit can be found at https://github.com/SubtitleEdit/
10. http://plagiarism.bloomfieldmedia.com/wordpress/software/wcopyfind/

ORCID iD

Christian Bokhove https://orcid.org/0000-0002-4860-8723

References

Anastasopoulos N and Baer AM (2013) MOOCs: When opening doors to education, institutions must ensure that people with disabilities have equal access. *New England Journal of Higher Education*. Available at: http://www.nebhe.org/thejournal/moocs-whenopening-the-door-to-education-institutions-must-ensure-that-participants-with-disabilitieshave-equal-access/ (accessed 3 June 2017).

Bain K, Basson S, Faisman A, et al. (2005) Accessibility, transcription, and access everywhere. *IBM Systems Journal* 44(3): 589–603.

Barton F, Bradbrook G and Broome G (2015) *Digital Accessibility: A Report for Citizens Online*. Edinburgh: Citizen Online.

Bennett C, Wheeler K, Wesstrick M, et al. (2015) Disabilities, opportunities, internetworking, and technology panel: Student perspectives. In:. *IT accessibility in higher education capacity building institute*, Seattle, WA, 4–6 February.

Biggs J (2017) Hackers are hiding malware in subtitle files. Available at: https://techcrunch.com/2017/05/24/hackers-are-hiding-malware-in-subtitle-files/ (accessed 3 June 2017).

Byrne W, Doermann D, Franz M, et al. (2004) Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Transactions on Speech and Audio Processing* 12(4): 420–435.

Chaudhuri S (2017) Adding sound effect information to YouTube captions. Available at: https://research.googleblog.com/2017/03/adding-sound-effect-information-to.html / (accessed 30 March 2018).

Clossen A (2014) Beyond the letter of the law: Accessibility, universal design, and human centered design in video tutorials. *Pennsylvania Libraries: Research & Practice* 2(1): 27–37.

Cohen L, Manion L and Morrison K (2007) *Research Methods in Education*. Abingdon: Routledge.

Collins RK (2013) Using captions to reduce barriers to Native American student success. *American Indian Culture and Research Journal* 37(3): 75–86.

Dahbi M (2004) English and Arabic after 9/11. *The Modern Language Journal* 88(4): 628–630.

Dempster PG and Woods DK (2011) The economic crisis though the eyes of Transana. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 12(1): 16.

Deng L and Yu D (2014) Deep learning: Methods and applications. *Foundations and Trends in Signal Processing* 7(3–4): 197–387.

Dresing T, Pehl T and Lombardo C (2008) Schnellere Transkription durch Spracherkennung? *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 9(2): 17.

Dubinsky A (2014) SyncWords: A platform for semi-automated closed captioning and subtitles. In: *INTERSPEECH*, Singapore, 14–18 September.

Economic and Social Research Council (ESRC) (2017) *Data Protection*. Available at: www.esrc.ac.uk/funding/guidance-for-applicants/research-ethics/frequently-raised-topics/data-requirements/data-protection/ (accessed 3 June 2017).

Evers JC (2011) From the past into the future. How technological developments change our ways of data collection, transcription and analysis. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 12(2): 38.

Federico M and Furini M (2012) Enhancing learning accessibility through fully automatic captioning. In: *Proceedings of the international cross-disciplinary conference on web accessibility*, Lyon, 16–17 April.

Fichten CS, Asuncion J and Scapin R (2014) Digital technology, learning and postsecondary students with disabilities: Where we've been and where we're going. *Journal of Postsecondary Education and Disability* 27(4): 369–379.

Fogel S (2017) IBM inches toward human-like accuracy for speech recognition. Available at: https://www.engadget.com/2017/03/10/ibm-speech-recognition-accuracy-record/ (accessed 30 March 2018).

Greenfield P (2018) The Cambridge analytica files: The story so far. Available at: https://www.theguardian.com/news/2018/mar/26/the-cambridge-analytica-files-the-story-so-far (accessed 30 March 2018).

Hunt G, Moloney M and Fazio A (2011) Embarking on large-scale qualitative research: Reaping the benefits of mixed methods in studying youth, clubs and drugs. *Nordisk Alkohol- & Narkotikatidskrift* 28(5-6): 433–452.

Jefferson G (2004) Glossary of transcript symbols with an introduction. In: Lerner GH (ed.) *Conversation Analysis: Studies from the First Generation*. Amsterdam: John Benjamins Publishing Company, pp. 13–31.

Johnson A (2014) *Video captioning policy and compliance at the University of Minnesota Duluth*. Master's Thesis, University of Minnesota Duluth, Duluth, MN.

Johnson BE (2011) The speed and accuracy of voice recognition software-assisted transcription versus the listen-and-type method: A research note. *Qualitative Research* 11(1): 91–97.

Juang BH and Rabiner LR (2004) *Automatic Speech Recognition – A Brief History of the Technology Development*. Available at: www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/354_LALI-ASRHistory-final-10-8.pdf (accessed 3 June 2017).

Kawahara T (2012) Transcription system using automatic speech recognition for the Japanese parliament (Diet). In: *Proceedings of the 24th innovative applications of artificial intelligence*, Toronto, ON, Canada, 22–26 July, pp. 2224–2228. Available at: https://pdfs.semanticscholar.org/ca66/f0725ee52026ca838c94bc5771818e801086.pdf

Kawahara T, Nanjo H, Shinozaki T, et al. (2003) Benchmark test for speech recognition using the Corpus of Spontaneous Japanese. In: *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo, Japan, 13–16 April.

Kvale S (1996) *Interviews*. London: SAGE.

Lee RM (1993) *Doing Research on Sensitive Topics*. London: SAGE.

Lewis M and Jackson D (2001) Television literacy: Comprehension of program content using closed captions for the deaf. *Journal of Deaf Studies and Deaf Education* 6(1): 43–53.

Liao H, McDermott E and Senior A (2013) Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription. In: *IEEE Workshop on Automatic Speech Recognition and Understanding*, Olomouc, Czech Republic, 8–13 December, pp. 368–373. Available at: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/41403.pdf

Matheson JL (2007) The voice transcription technique: Use of voice recognition software to transcribe digital interview data in qualitative research. *Qualitative Report* 12(4): 547–560.

O'Neill C (2016) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publishing Group.

Parton B (2016) Video captions for online courses: Do YouTube's auto-generated captions meet deaf students' needs? *Journal of Open, Flexible and Distance Learning* 20(1): 8–18.

Paulus T, Lester J and Dempster P (2013) *Digital Tools for Qualitative Research*. London: SAGE.

Punch K and Oancea A (2014) *Research Methods in Education* (2nd edn). London: SAGE.

Ramírez J and Górriz JM (eds) (2011) *Recent Advances in Robust Speech Recognition Technology*. Sharjah, United Arab Emirates: Bentham Science Publishers.

Ranchal R, Taber-Doughty T, Guo Y, et al. (2013) Using speech recognition for real-time captioning and lecture transcription in the classroom. *IEEE Transactions on Learning Technologies* 6(4): 299–311.

Roberts LC, Whittle CT, Cleland J, et al. (2013) Measuring verbal communication in initial physical therapy encounters. *Physical Therapy* 93(4): 479–491.

Ruan X (2015) The effect of caption modes on EFL students' video comprehension. *Journal of Language Teaching and Research* 6(2): 397–404.

Shiver BN and Wolfe RJ (2015) Evaluating alternatives for better deaf accessibility to selected web-based multimedia. In: *Proceedings of the 17th international ACM SIGACCESS conference on computers & accessibility*, Lisbon, 26–28 October, pp. 231–238. New York: ACM.

Srnicek N and Williams A (2015) *Inventing the Future: Postcapitalism and a World without Work*. London: Verso Books.

Suffridge D and Somjit P (2012) Bridging the distance with intro videos and weekly update videos. Presented at the 28th Annual Conference on Distance Teaching & Learning, Madison, WI, 8–10 August. Available at: www.uwex.edu/disted/conference/Resource_library/proceedings/62238_2012.pdf (accessed 3 June 2017).

Tarantola A (2016) Microsoft's speech recognition engine listens as well as a human. Available at: https://www.engadget.com/2016/10/18/microsofts-speech-recognition-engine-listens-as-well-as-a-human/ (accessed 30 March 2018).

Turnitin (2017) Originality checking. Available at: http://www.turnitinuk.com/en_gb/features/originalitycheck (accessed on 13 April 2017).

UK data archive (2017) *Create and Manage Data: Formatting Your Data – Transcription*. Available at: http://www.data-archive.ac.uk/create-manage/format/transcription (accessed 3 June 2017).

Wald M (2005) Enhancing accessibility through automatic speech recognition. In: Proceedings of Accessible Design in the Digital World Conference ADDW2005(ADDW 2005). Available at https://eprints.soton.ac.uk/262143/ (accessed 30 March 2018)

Wald M (2010) Synote: Accessible and assistive technology enhancing learning for all students ICCHP 2010. In: Miesenberger K, Klaus J, Zagler W, et al. (eds) *Computers Helping People with*

Special Needs: Lecture Notes in Computer Science. Berlin: Springer, pp. 177–184.

Wald M and Bain K (2007) Enhancing the usability of real-time speech recognition captioning through personalised displays and real-time multiple speaker editing and annotation. In: *International conference on universal access in human-computer interaction*, Vancouver, BC, Canada, 9–14 July, pp. 446–452. Berlin: Springer.

Walford G (2001) *Doing Qualitative Educational Research: A Personal Guide to the Research Process*. London: Continuum International Publishing.

Author biographies

Christian Bokhove is associate professor in Mathematics Education with research interests in everything regarding mathematics teaching in secondary schools. In addition, he likes to apply novel research methods to educational and classroom contexts.

Christopher Downey is associate professor of Education with research interests in educational effectiveness and improvement; particularly the use of data to inform expectations and decision-making, the implications of educational accountability and our understanding of social processes associated with learning and educational practice.