Automated transcription software in qualitative research

Stephen Louw
King Mongkut's University of Technology Thonburi

Abstract

Qualitative research frequently requires transcription of spoken data for data analysis. The process of transcribing spoken text can be laborious and time consuming. Recent advances in natural language speech recognition software are making automated transcription of spoken texts an attractive possibility for researchers. To explore the value of automatic transcription as a viable tool for applied linguistics researchers, two popular automatic software programs were used to transcribe four audio recordings. The audio recordings included a mix of single and multiple speakers, native and non-native speakers of English, and recordings with low and high background noise. The output from each program for the recordings was compared with manual transcriptions for overall accuracy, as well as the programs' handling of turn taking, overlaps and interruptions. The findings are expected to show that while the transcriptions are often accurate, the lack of transcription detail means they are unlikely to serve the specific purposes of qualitative research.

1. Introduction

Qualitative research in applied linguistics frequently involves the collection of verbal data. In one issue of Applied Linguistics (issue 42/4), for instance, only two of the eight articles do not specify qualitative data involving transcription of interviews, spoken interaction, or observations. Through the collection and reporting of such data, qualitative research offers rich and thick descriptions of research sites, adds depth and credibility to a study's findings, minimizes confirmation bias, and informs the reader about transferability (Onwuegbuzie & Leech, 2006).

The benefits of rich, thick description of spoken data in qualitative research is contingent on the transfer of the spoken data into a format that facilitates its analysis and presentation to the reader. For this reason, qualitative researchers are faced with the need to transcribe this spoken data. Creating accurate transcriptions, however, is time consuming. Walford (2001) suggests a ratio of five hours of transcription to each hour of audio, though this depends on the nature of the text to be transcribed. Bird (2005), for instance, describes a transcription of two hours of audio that took 70 hours. The production of transcripts, then, poses a serious obstacle to researchers who wish to conduct qualitative research.

The problem of transcribing spoken language is not limited to researchers in applied linguistics. Transcriptions of spoken discourse are increasingly relied upon in business meetings and court trials. Developments in automated speech recognition technology (ASR), especially since 2010 (Swamy & Ramakrishnan, 2013), have found their way into much of our modern lives. One consequence of these developments is the increasing number of automated transcription services, both free and paid, which opens the possibility that researchers may be released from the laborious task of manual transcription. In this study, I investigate whether such ASR software serves as an appropriate aid to transcription of spoken data in qualitative research in applied linguistics. If so, developments in this technology represent a major affordance that will greatly ease the demands on qualitative researchers.

2. Transcriptions in qualitative research

A transcript aims to capture spoken language to facilitate the process of manipulation and analysis. In much published research, this process is described unproblematically: the data were transcribed 'verbatim' or 'in full' (Nascimento & Steinbruch, 2019). This assumes that a transcript can be a transparent representation of the spoken word, and that what was said (orally) can be represented ('in full') in the written medium. The assumption of transparency of transcriptions rests on a positivist epistemology in which there is a 'correct' or 'accurate' record of the speech event. There are, however, a number of problems with such an assumption.

Spoken language is highly performative, and includes elements not normally included in the written medium, such as intonation and volume. As an example, an ethnographer studying police interactions with the public may transcribe the following utterance:

PO4: Put your hands in the air.

This would capture the words spoken by a police officer to a criminal. Another ethnographer, this one studying the language use of a second-grade teacher with her EFL students, might capture the following piece of classroom discourse:

T: Put your hands in the air.

The transcription of these two utterances loses much of the difference in the way the instructions are given. We might presume, for instance, that the teacher's utterance is made with much less insistence, maybe even with a smile, followed by a pause while she waits to nominate a student, unlike that of the police officer. Since transcription of the spoken words misses a great deal of its pragmatic force, it cannot be said to be transparent.

Instead, transcription needs to be viewed as an interpretive act involving multipledecisions on the part of the researcher. Transcribing is a construction or re-presentation of communication from one medium to another. In the process of freezing the spoken language, information is necessarily lost. Denham & Onwuegbuzie (2013) list four elements of spoken language as likely lost in transcriptions: proxemics (the interpersonal space in the communication), chronemics (the speed of the delivery and the length of silences), kinesics (body language and posture), and paralinguistics (including volume, pitch and voice quality). A diligent transcriber may endeavor to include as much details in the transcript as possible, but Poland (1995) points out that additional detail impedes readability. The transcriber, then, is not simply converting speech to writing, but must constantly make careful decisions about what to include (and not include) in a transcript. For instance, there are questions of how to deal with non-verbal vocalizations such as laughing or crying, or feedback tokens such as 'yeah' or 'mmhmm' which add no content to the data but are representative of the listener's engagement in the discourse. For Bird's (2005) 70-hour transcription, she describes dealing with a recording of a dinner party in which it was necessary for her to consider whether the sounds of people enjoying their food and comments such as "Pass the sauce" count as data. Such decisions are representative of the constructed nature of the transcription, and are evidence of the subjective nature of transcription as the researcher re-constructs a written version of the spoken data.

Since the transcription process involves decisions about what can be included, it is necessarily selective and interpretive. The consequence of this is that there can be no single 'correct' transcript, only transcripts that are more or less useful to the researcher's purpose. This spotlights the issue of how to measure accuracy of transcripts. Poland (1995) recommends judging a transcript's quality on its trustworthiness based on two criteria: its faithfulness to the original speaker's intention, and its fit with the research project. On the issue of the fit between the transcript and the research theory, Lapadat (2000) argues that the transcript not only reflects theory, but can constrain theorizing. For this reason, transcription is better viewed as a process rather than as a product. This line of thinking challenges the notion of an automated transcript that can serve a researcher since algorithms are unable to make important decisions about meaning-making and its interpretation, nor about ways in which these meanings may best be represented.

In answer to this challenge of the value of automated transcription, Borkhove & Downey (2018) argue that the output from ASR programs is only part of the transcription process. Given the problems in automated transcription output, they propose the use of automated transcripts as a 'first draft' where the researcher revisits the data to edit and revise the transcripts. Borkhove & Downey's proposal is consistent with ten Have's (2007) description of approaching transcripts in 'rounds', where the transcriber revisits the audio multiple times, each time focusing on a specific feature of the talk until all the features of interest have been recorded. This places the automated transcription as a single part of a larger transcription process, eliminating the possibility of achieving a finished product from the ASR program. It is unclear, however, whether utilizing ASR programs in this way will actually lead to a saving of researchers' time.

3. Technology in qualitative research

Technology is an important part of qualitative research. An array of programs and software are available to qualitative researchers serving a multitude of important functions, including data storage, text search and retrieval, coding, data linking, memoing, and data display. Programs such as NVivo and NUD*IST have become widely used by qualitative researchers (Silver & Lewins, 2014), bringing a variety of benefits. Computers are more efficient and accurate at manual tasks like sorting and searching, freeing up a researcher's time to spend on higher-inference tasks such as theory building. Computer software is also highly consistent: within a desired set of parameters, computers do not suffer from fatigue (Weitzman, 2000).

In spite of these benefits, the introduction of software into research has brought with it controversy. Lee and Fielding (1991) warn that "the use of computers may tempt qualitative researchers into 'quick and dirty' research" (p. 8). The speed, efficiency, and apparent accuracy of the programs lull the researcher into a security that results in shortcuts or skipping processes. Another area of concern with researcher's reliance on technology is the loss of a closeness to the data (Weitzman, 2000). Arguments against the use of technology in qualitative research are tempered by calls for researchers to become aware of the theoretical assumptions underlying the program's design and ensure a match between the technology and the research objective. Failure to do so may lead to the misapplication of the technology, and therefore to bias in the findings.

Technology to identify and decode speech recognition dates as far back as the work of Davis, Biddulph & Balashek (1952). By 1990, consumer-ready products like Dragon Dictate were capable of interpreting spoken language to reroute telephone calls and were used in medical practice (Al-Aynati & Chorneyko, 2003). Speech recognition software has long been seen as a potential affordance in applied linguistics. For instance, Coniam (1999) explored the use of Dragon Software with second language speakers of English and concluded it had potential benefits for assessment and the teaching of pronunciation. More recent innovations in algorithms include neural networks, such as Attention-Based speech recognition (Chan et al, 2016), which can function without a language model. Despite the impressive progress made with this technology, automated speech recognition software is still prone to issues of accuracy which draws into question its usefulness as a tool for research. In this study, I aim to investigate ASR technology as a possible solution to transcription demands on researchers in applied linguistics. My focus was on the accuracy of two sample ASR programs with a view to identifying the value of the outputs for qualitative research in this field.

4. Methodology

Four kinds of transcription services are available: human transcribers (such astranscription HUB); manual transcription software (such as inqscribe); direct dictation software; and automated transcription from a recording. For this study, only programs in the final category were considered. I identified twelve of these, describing their transcription services as low-cost and high-quality (using language such as 'powerful tools', and 'fast and secure'). Some incorporated editing capabilities and mobile device compatibility.

I did not aim to review the programs, but to use them as exemplars of how ASR technology can be used by researchers. To identify two popular programs which would serve the purposes of this study, I consulted online reviews of the transcription providers (for example www.capterra.com/transcription-software/) and selected Sonix (www.sonix.ai) and Otter (www.otter.ai). The programs are based on 'deep machine learning', though I could not identify the difference between them since the algorithms are proprietary. In terms of offering, Otter was the more attractive, with 600 minutes of transcription for free per month, and functionality such as a keyword search, and the option to 'train' the program to identify speakers.

Previous studies (for example Kawahara, Nanjo, Shinozaki & Furui, 2003) have identified a variety of factors influencing accuracy of automated transcriptions: the number of speakers, speaker accent, and audio quality. These are characteristics of speech likely to be relevant in applied linguistics research. For this study, I selected five audio files which incorporate these variables (see Table 1). Four of these are sampled from data for a previous study (Louw, Watson Todd and Pattamawan, 2016). Audio 1, a monologue, was prepared specifically for this study. Each audio file was trimmed to three minutes.

The first criterion was the number of speakers. The separation of text from one speaker to another might be a source of difficulty for the AI programming, particularly when turn taking involves interruptions or overlaps. Therefore, the audio files were selected to compare the effects of this variable on the ASR programs. One file included a single speaker, one file had three speakers, and the rest were interviews between two speakers. To account for the possibility that the transcription programs could distinguish speakers based on gender, one two-speaker audio was a dialogue between two men, and a second was included between a woman and a man. The three-speaker audio included two men and one woman.

The second consideration was accent. My selection of the audio files included speakers from a variety of language backgrounds, including speakers from inner, outer and expanding circle countries (Kachru, 1990). If the ASR programming is based on a native-speaker model of English, non-native accents could compromise the quality of the transcriptions.

The final criterion for the data was audio quality. The AI websites specify the need for audio without background noise. However, audio recordings in applied linguistic research frequently take place in contexts where the researcher has little control over the ambient noise, such as classrooms. Audio 3 was included to investigate the effect of background noise. This file is taken from an interview conducted in a noisy coffee shop.

Table 1. Summary of the audio files.

	Number of speakers	Nationality of speakers	Gender	Background noise
Audio 1	1	Chinese	M	Low
Audio 2	2	South African, British	M, M	Low
Audio 3	2	South African, British	M, M	High
Audio 4	2	British, Chinese	M, F	Low
Audio 5	3	Australian, British, American	M, F, M	High

Each file was uploaded into both the Sonix and Otter platforms. Both platforms allow for editing of the transcription using the online text editor linked to the audio file. However, I opted not to use this function and downloaded the transcription output without further editing. Otter produced a .txt file output, and Sonix a word document. For analysis, the automatic time stamps in the Otter and Sonix transcriptions were removed.

In addition to the transcriptions produced by the automated transcribers, I prepared a manual transcription of the audio files using SoundScriber (http://www-personal.umich.edu/~ebreck/code/sscriber/). These transcriptions used the transcription conventions from Louw et al (2016). Pauses were not timed, but discourse features potentially relevant to a study of language were included, including response tokens, turn overlaps, repetitions, and interruptions. These manual transcripts served as a baseline for evaluating the accuracy of the automated transcripts. There is, of course, the possibility that the manual transcripts are flawed, but their use offers a relevant metric for the calculating the accuracy of the automated transcripts. For the purposes of the study, I follow Poland's (1995) definition of transcription accuracy as one that matches the audio. Although problematic in that this takes a positivist view that there is a 'correct' version, it allows for a comparison between transcripts.

With the manual transcription as the point of comparison, the text files were divided into turn constructional units (TCU) following suprasegmental speech patterns (Reed, 2009). These units of analysis were based on the natural pauses and breaks in the speech. Therefore, each speaker pause constituted the end of a phrase unit. Turn changes marked the end of a TCU. Where response tokens overlapped, the unit was divided at the point of the interruption.

To calculate the match between the transcriptions, the phrase units in each automated transcript were compared with the manual transcription using Text Compare (http://www.text-compare.com/). Differences were labeled as a match (M), close match (C), and no match (X). A close match (C) was defined as minor errors not affecting the intelligibility of the speaker, or where the meaning is easily recoverable. Following Bucholz's (2000) distinction between naturalized and denaturalized transcription, features associated with written language (punctuation, capitalization, spelling) were ignored unless the difference affected the interpretation of the text. Following the analysis, a side-by-side comparison of the three transcriptions for each audio file was constructed and the number of matches, mismatches and close matches calculated. Table 2 presents an example of this process (from audio 1).

Table 2. Sample of comparison of transcript outputs (from audio file 1)

Manual	Otter		Otter	
In the following three or four minutes I'm going to talk about my PhD study,	In the following three or four minutes, and we're going to talk about the mind PhD study	M C X	In the following three or four minutes we're going to talk about the mind putative study.	M C X

5. Findings

Table 3 summarizes the proportions of matches (M), close matches (C), and mismatches (X) of the automated transcriptions for all five audio files, using the manual transcription as a benchmark. The greater proportion of matches (M) indicates greater consistency between the manual and automated transcription.

Table 3. Proportion of matches (M), close matches (C), and mismatches of the automated transcriptions compared to the manual transcription.

	Otter			Sonix		
	M	C	X	M	C	X
Audio 1 (monologue)	67.50	12.50	20	60.00	11.25	28.75
Audio 2 (interview)	53.47	20.79	25.74	37.62	21.78	40.59
Audio 3 (interview - noisy)	38.82	15.29	45.88	28.24	17.65	54.12
Audio 4 (interview)	47.32	16.96	35.71	39.29	24.11	36.61
Audio 5 (3-way discussion)	25.69	13.19	61.11	24.31	11.11	64.58

Overall, the two programs were roughly comparable, though Otter outperformed Sonix, most noticeably with audio 2. Both programs were most accurate with audio 1 (one speaker), and least accurate with audio 5 (three speakers). The output from audio 1 shows the strengths of the ASR programs in transcribing single speaker audio with low background noise. For this transcript, there was an 80% match or close match between Otter and the manual transcription. For the audio files with multiple speakers, accuracy rates were lower. In audio 5 with three speakers, for instance, the proportion of matches and near matches was below 40% for both programs.

Predictably, the proportion of matches was noticeably lower in the audio with high background noise (audio 3). Accent, however, does not appear to have affected output. Files 1 and 4 included non-native English speakers with identifiable Chinese accents which the programs did not have difficulty transcribing. In fact, Sonix performed better with the non-native English speaker (audio 4) than with native speakers (audio 2).

An analysis of close matches in the transcripts identified three categories of errors:

1. errors with unstressed words: the ASR programs were sometimes not able to transcribe unstressed words, for example; 'this complex environment' was transcribed by Sonix as 'these complex environment',
2. missing words: words in the audio were not transcribed, for example the dysfluency 'towards to a more just and harmonious society' was transcribed by Sonix as 'towards a more just and harmonious society',
3. added words: the automated transcripts included words that were not in the audio, for example 'from my participants' was transcribed by Sonix as 'for all my participants'.

I would argue that these errors do not interfere in the meaning conveyed by the speaker and the correction can easily by recovered by the reader/researcher.

The sources of mismatches in the data are consistent with those from previous studies (Bokhove & Downey, 2018; Poland, 1995). These included homophones (for/four; Thai/tie) and near homophones (collars/colors). With some mismatches, the similarities to the original are easily identifiable, for example ‘one of the participants got fired’ was transcribed by Otter as ‘one of the participants caught fire’.

The ASR transcriptions of audio with multiple speakers had two further problems. First, large portions of the audio were missing. For instance, in audio 3 (NS), the manual transcript includes 567 words, whereas the Otter transcription totals only 471. Second, the ASR programs had difficulty identifying turn-changes between speakers, and instead presented interactional data as a single turn. The following extract from the opening turns of audio 3 shows how much of the interactional data is lost in the ASR output.

Extract 1. Comparison on manual ASR transcription, audio 3.

Manual	Sonix
S: Okay. R: Is it recording now? S: Yes. R: Okay right I will use my best voice, okay. S: Um the research is about…	Ok. Yes. I will use my best voice. The researchers about …

The high error rates for multi-speaker audio were particularly noticeable with audio 5, which included 3 speakers with multiple changes in turns and frequent feedback tokens. Table 4, an extract from this data, illustrates how the automated systems clearly found it difficult to untangle language in interaction.

Table 4. Comparison of transcriptions taken from audio 5.

Manual	Otter		Sonix	
C		X		X
I have one every week	Man every week.	X		X
T	Speaker	M		X
You have one every week	This week	X	you have one every week.	M
C		X		X
(laugh)		X		X
T		X		X
Okay so hey I can put in this week you will	he will and this	X	It’s like a wedding this week. He will,	X

6. Discussion

This study aimed to investigate automated transcription software in terms of its usefulness for the applied linguistics researcher. To do this, audio files taken from previous research in the field were run through two ASR programs and the outputs compared with a manual transcription of the same files.

The findings show that the two programs' accuracy rates were approximately equal, and suffered similar difficulties with low quality audio and audio involving multiple speakers. With regard to background noise, error rates in both ASR transcripts for audio 3, the interview which took place in a noisy coffee shop, were higher than for similar recordings with no background noise. This is unsurprising given that the ASR websites specify the importance of good quality audio with low background noise. However, this finding has consequences for researchers in applied linguistics who may need to collect audio data on site where background noise cannot be controlled, such as in classroom research.

A potentially useful finding from the data is that transcription quality was not affected by the speakers' accents, and in particular, the inclusion of non-native English speakers did not appear to affect ASR output. The strength of these programs to manage a variety of English accents is useful to researchers in applied linguistics, especially in studies on language learning.

The variable with the greatest impact on the quality of the transcriptions was the number of speakers. In the audio with a single speaker, the ASR outputs were remarkably accurate. Otter achieved 80% match or near match with the monologue. The large number of matches with the monologue file would appear to indicate that automated transcription offers a major affordance to qualitative researchers. This accuracy, however, can lead a busy researcher to complacency. With the accuracy of the transcripts, the possibility of exists that a researcher will accept the output unquestioningly, or that mismatches in the output are masked by plausible transcripts which do not demand editorial attention. In the data from this study, some mismatches are so close to the original that they appear, on casual inspection, to be correct. For example, in the following excerpt, the Sonix output is somewhat plausible when read in isolation, but the transcript in no way accurately conveys the speaker's meaning. The high quality ASR outputs, then, lead to the possibility of the 'quick and dirty' research that Lee and Fielding (1991) warned of with the introduction of technology into the research process.

Extract 2. Comparison of manual and ASR transcript, audio 1.

Manual	Sonix
In my study, four international queer or LGBT teachers told me about their experience as teachers in Bangkok. Drawing on their stories I investigate how the queer teachers negotiate their responsibilities as teachers	You must study for international queer or LGBT teachers told me about their experiences in Bangkok. Join in on their stories. I investigate how the queer teachers negotiate their responsibilities as teachers.

While the high accuracy with monologues is promising, the reality of applied linguistics research is that it frequently interests itself with interaction and dialogue. It is in this respect that the data highlight the biggest problem with automated transcriptions for qualitative researchers in the field. There was a significant drop in accuracy rates in audio with multiple speakers: in audio 5, with three speakers, the ASR transcripts show less than 25% match. Much of the problem the ASR programs had with multi-speaker audio was with the identification of changes of turn, and the loss of data with overlaps or interruptions.

The data included three audio files involving an interview. Mann (2011) argues that in interviews in applied linguistics, the interaction between interviewer and interviewee forms an important part of the research process. This findings from this study show that with such interaction, ASR was less accurate than with the monologue audio. During dialogic exchanges, much of the interaction between the speakers is missing in the ASR output, and the programs had difficulty differentiating speakers. For example, the following excerpt from audio 4 (between a male teacher and a female researcher) demonstrates the loss of much of the interactional data in the Otter transcript.

Extract 3. Comparison of manual and ASR transcript for audio 4.

Manual	Otter
S: …when you come to Thailand the infrastructures are absolutely shocking the pollution at the moment is terrible.	S1: …when you come to Thailand, the infrastructures are absolutely shocking. The pollution at the moment is terrible.
W: This moment is really	S2: At the moment it is really worse
S: Er yeah it's really bad.	
W: It could be the worst	
S: Yeah but the there's something about the people	S1: Yeah. But there's something about the people.
W: People here	

Notice that in the opening turn when one speaker holds the floor, Otter's transcription is accurate. With the turn changes, however, much of interactional data is lost, and although isolated words are identified, three entire turns are missing from the transcription. It can be argued that the transcript maintains an overview of the content of the interview and that these turns are expendable. However, for a researcher interested in discourse in interaction, this represents a significant loss of potential data.

Bokhove & Downey (2018) argue that, despite the errors produced by ASR tools, they form a useful 'first draft' for researchers dealing with audio data. For researchers who intend to revisit the transcripts to add information in rounds, as described by ten Have (2007), the automated transcript might be considered a useful first round to be edited and managed in subsequent revisions. Following Bokhove & Downey's suggestion, the ASR transcripts represent a significant contribution to qualitative researchers in preparing a first draft of the transcripts quickly and cheaply, thereby saving time and allowing the researcher to focus attention on specific aspects of the transcript.

The risk, however, is that the ASR outputs are identified by researchers as 'data' and the audio is not revisited for revision of the transcripts. In this scenario, the transcripts are seen as product rather than process (Lapadat, 2000), and no attempt is made to match the transcript with the theoretical position of the study. Lapadat (2000) warns of the danger of using transcripts without 'clarity of purpose' (p. 214), where the transcript output determines the nature of the study rather than the goals of the study determining the focus of the transcripts. To demonstrate this distinction, in one study focusing on authoritativeness in dialogic interaction between trainers and trainee teachers (Louw et al., 2016), it was necessary to investigate the ways in which turns were allocated, supported, or withheld. This theoretical position guided the transcript output and defined how the interaction between speakers was represented. The transcript (Extract 4), therefore, shows feedback tokens, latched turns (shown with '=') to indicate interruptions, and overlapping speech (shown with '{'). These features of the interaction between the speakers help identify the ways in which the trainees were or were not able to claim turns.

Extract 4. Manual transcript of audio 5

T	You might not.. so this year might not = that that
P	= Oh god yeah because Nop turned in turned that into making a baby rather than have a baby I was just I'm just leaving (laugh) {leaving
C	{It's probably just lucky that {Sam's not here
T	{That's that's fluency
P	(laugh) Yeah I know that one was just

By comparison, the ASR output (from Sonix) of this same stretch of speech presents much less information about the interaction, and is thus unsuitable for the specific purpose of this study.

Extract 5. ASR transcript of audio 5

you might not said this year might not project that much time in terms of managing a baby rather than have a baby also.

You an it was just lucky that said. So that's fluency

These findings draw into question the value of the ASR outputs as a useful source of input for qualitative researchers in applied linguistics. With such large gaps in the transcripts, and lost opportunities for identifying interaction following the theoretical position of the study, the output from these programs may be of little value. More problematic, however, is the possibility that through the use of the technology, the researcher becomes removed from the data. The transcription process, though time consuming, offers researchers an important opportunity to become familiar with the data and identify features of the data that may inform analysis. Weitzman (2001) warned of technology's effect of limiting a researcher's closeness to the data, and with the introduction of automated transcription, the possibility emerges that a researcher may consider it unnecessary to revisit the audio data with an analytic ear.

7. Conclusion

Transcription is an integral part of much qualitative research in our field. Done well, the output from transcriptions is a visual representation of the spoken data, facilitating further analysis. That transcription is such a time-consuming endeavor is why automated transcription services are an attractive option. This study shows that these ASR outputs are surprisingly good with certain data: single speakers with high quality audio. With audio involving multiple speakers or with background noise, the quality of the output is compromised and the trustworthiness of the transcriptions becomes questionable. These shortcomings may be resolved in future iterations of the technology. Nevertheless, even with highly accurate transcripts, it is important that researchers foreground their own theoretical purposes when utilizing transcripts, identifying features of the text which are important to their goals, rather than relying on what is produced by the ASR technology. At best, then, these automated transcripts represent a valid 'first draft' of a transcript rather than finalized output. As I pointed out in the introduction above, spoken data forms the backbone of a much research in our field, and transcripts serve to facilitate close analysis of this data. Considering the importance of transcripts for our field, we need some pause for consideration about the value of automation. Researcher using automated transcription services, then, need to approach the transcripts as drafts and act on them accordingly: revisit the audio to edit the output, make their own transcription decisions, and ensure that the final transcripts match the theoretical goals of the study.

References

Al-Aynati, M. M., & Chorneyko, K. A. (2003). Comparison of voice-automated transcription and human transcription in generating pathology reports. *Archives of pathology & laboratory medicine, 127*(6), 721-725.

Bird, C. M. (2005). How I stopped dreading and learned to love transcription. *Qualitative inquiry, 11*(2), 226-248.

Bokhove, C., & Downey, C. (2018). Automated generation of ‘good enough’transcripts as a first step to transcription of audio-recorded data. *Methodological innovations, 11*(2), 1-14.

Bucholtz, M. (2000). The politics of transcription. *Journal of pragmatics, 32*(10), 1439-1465.

Coniam, D. (1999). Voice recognition software accuracy with second language speakers of English. *System, 27*(1), 49-64.

Davis, K. H., Biddulph, R., & Balashek, S. (1952). Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America, 24*(6), 637-642.

Denham, M. A., & Onwuegbuzie, A. J. (2013). Beyond words: Using nonverbal communication data in research to enhance thick description and interpretation. *International Journal of Qualitative Methods, 12*(1), 670-696.

Kachru, B. B. (1990). World Englishes and applied linguistics. *World Englishes, 9*(1), 3-20.

Kawahara, T., Nanjo, H., Shinozaki, T., & Furui, S. (2003). Benchmark test for speech recognition using the Corpus of Spontaneous Japanese. In *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*.

Lapadat, J. C. (2000). Problematizing transcription: Purpose, paradigm and quality. *International journal of social research methodology, 3*(3), 203-219.

Lee, N. G. & Fielding R. M. (1991) *Using computers in qualitative research*. London, Sage.

Louw, S., Watson Todd, R., & Jimarkon, P. (2016). Teacher trainers’ beliefs about feedback on teaching practice: Negotiating the tensions between authoritativeness and dialogic space. *Applied Linguistics, 37*(6), 745-764.

Mann, S. (2011). A critical review of qualitative interviews in applied linguistics. *Applied linguistics, 32*(1), 6-24.

Nascimento, L. D. S., & Steinbruch, F. K. (2019). “The interviews were transcribed”, but how? Reflections on management research. *RAUSP Management Journal, 54*, 413-429.

Onwuegbuzie, A. J., & Leech, N. L. (2007). Validity and qualitative research: An oxymoron? *Quality & Quantity, 41*(2), 233-249.

Poland, B. D. (1995). Transcription quality as an aspect of rigor in qualitative research. *Qualitative inquiry, 1*(3), 290-310.

Reed, B. S. (2009). Units of interaction:“Intonation phrases” or “turn constructional phrases”. *Actes/Proceedings from IDP (Interface Discours & Prosodie)*, 351-363.

Silver, C., & Lewins, A. (2014). *Using software in qualitative research: A step-by-step guide*. London: Sage.

Swamy, S., & Ramakrishnan, K. V. (2013). An efficient speech recognition system. *Computer Science & Engineering, 3*(4), 21.

ten Have, P. (2007) *Doing conversation analysis: A practical guide (2nd ed)*. London: Sage.

Walford, G. (2001) *Doing qualitative educational research: A personal guide to the research process*. London: Continuum International Publishing.

Weitzman, E. A. (2000). Software and qualitative research. In Guba, E. G., Lincoln, Y. S., & Denzin, N. K. *Handbook of qualitative research (2nd edition)*, Thousand Oaks, Sage, 803-820.