SSOAR

Open Access Repository
www.ssoar.info

Automatic Transcription of English and German Qualitative Interviews

Wollin-Giering, Susanne; Hoffmann, Markus; Höfting, Jonas; Ventzke, Carla

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Wollin-Giering, S., Hoffmann, M., Höfting, J., & Ventzke, C. (2024). Automatic Transcription of English and German Qualitative Interviews. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *25*(1). https://doi.org/10.17169/fqs-25.1.4129

Nutzungsbedingungen:
Dieser Text wird unter einer CC BY Lizenz (Namensnennung) zur Verfügung gestellt. Nähere Auskünfte zu den CC-Lizenzen finden Sie hier:
https://creativecommons.org/licenses/by/4.0/deed.de

Terms of use:
This document is made available under a CC BY Licence (Attribution). For more Information see:
https://creativecommons.org/licenses/by/4.0

gesis
Leibniz-Institut
für Sozialwissenschaften

Mitglied der
Leibniz-Gemeinschaft

FQS FORUM: QUALITATIVE SOCIAL RESEARCH SOZIALFORSCHUNG

Volume 25, No. 1, Art. 8
January 2024

Automatic Transcription of English and German Qualitative Interviews

Susanne Wollin-Giering, Markus Hoffmann, Jonas Höfting & Carla Ventzke

Key words: qualitative interviews; transcription; methods; automatisation; data protection; accuracy; automatic transcription

Abstract: Recording and transcribing interviews in qualitative social research is a vital but time-consuming and resource-intensive task. To tackle this challenge, researchers have explored various alternative approaches; automatic transcription utilising speech recognition algorithms has emerged as a promising solution. The question of whether automated transcripts can match the quality of transcripts produced by humans remains unanswered. In this paper we systematically compare multiple automatic transcription tools: *Amberscript*, *Dragon*, *F4x*, *Happy Scribe*, *NVivo*, *Sonix*, *Trint*, *Otter*, and *Whisper*. We evaluate aspects of data protection, accuracy, time efficiency, and costs for an English and a German interview. Based on the analysis, we conclude that *Whisper* performs best overall and that similar local-automatic transcription tools are likely to become more relevant. For any type of transcription, we recommend reviewing the text to ensure accuracy. We hope to shed light on the effectiveness of automatic transcription services and provide a comparative frame for others interested in automatic transcription.

Table of Contents

Licensed under a Creative Commons Attribution 4.0 International License, funded by the KOALA project

1. Introduction

The recording and subsequent transcription[1] of interviews is considered standard practice in qualitative interviewing, because transcribing is the most exhaustive and accurate way of reproducing spoken language in a written form. In conveying linguistic and content-related detail, through transcriptions researchers can be made aware of aspects that they might otherwise—intended or not intended—not remember, not write down, or not write down in enough detail. To obtain transparent and reliable research results, transcription is indispensable and we would recommend using it whenever possible and appropriate for the research goal.[2] [1]

However, creating a transcript is a time-consuming, resource-intensive task. To address this issue, various strategies have been proposed in the past, such as outsourcing it to external transcriptionists or trying to achieve high information density when note-taking. In recent years, however, automatic transcription based on speech recognition algorithms has increasingly come into focus as an alternative. Many companies have begun offering automated AI-supported transcription programs, either in cloud-based forms or as downloadable versions. By creating affordable options, these companies are trying to reach a broad spectrum of users, such as media professionals, businesspeople, journalists, and researchers. [2]

New technologies have been improved to such an extent in recent years, and continue to do so, that the question arises as to whether the quality of automatically generated transcripts is comparable to that of human-generated transcripts, at least for certain purposes. So far, this question has not been systematically answered. High-quality automated transcripts could replace manual ones, resulting in significant savings of time and resources for research. However, qualitative interviews often include confidential and intimate details,

1 We focus only on full transcriptions. Alternatives would be partial transcription and note-taking. In partial transcription, only parts of the interview are transcribed (MacLEAN, MEYER & ESTABLE, 2004). Dropping content while simultaneously being accurate might be suitable when large parts of the interview do not matter for the analysis, but usually this cannot be known beforehand. Note-taking includes everything from simply writing down memories, carrying out more complex variations such as "scripts" that serve as condensed interview summaries (RUTAKUMWA et al., 2020), cross-checking notes against the audio material (HALCOMB & DAVIDSON, 2006) to employing an active second interviewer with note-taking (LOUBERE, 2017) and note-taking with the help of a passive second interviewer (EASTON, LEXIER, LINDSTROM & YEO, 2019). Since it is intended to speed up data generation and analysis and simultaneously save time by omitting some content (ibid.), note-taking always represents a reduction of information in the form of a selective choice from a certain perspective. In this perspective the view is narrowed to certain aspects and steered away from other, possibly empirically interesting or essential aspects. We are not convinced by this approach. Note-taking only seems appropriate when people explicitly refuse to be recorded since they would otherwise be in danger, when the informality of the situation would be compromised, or when legal restrictions do not allow recording.

2 We write "appropriate" because sometimes, even a full transcript is already too much abstraction from the phenomenon of interest. For instance, when interactions or group dynamics are analysed, even the most detailed transcription system is insufficient. In such cases, working directly with audio-visual recordings is the better choice (MARKLE, WEST & RICH, 2011, §11-22). In other words, the choice of data to be analysed needs to align with the research question. In this paper we do not further consider direct analysis of audio or video because it is unnecessary for our type of research, and it is not a type of transcription.

making not only technical feasibility but also ethical implications surrounding data protection relevant. [3]

In this paper, we discuss the current capabilities of several automatic transcription services and develop an assessment framework for different transcription workflows. We then use this framework to make a systematic comparison of a German and an English qualitative interview, in contrast to manually generated human transcription. From the wide range of automatic transcription programs available, we selected *Amberscript*, *Dragon, F4x*, *Happy Scribe*, *NVivo*, *Sonix*, *Trint*, *Otter,* and *Whisper*. To make an informed choice when to use which option—manual or automatic transcription—we contrast issues concerning data protection, accuracy, time spent, and costs. [4]

To be more precise, in this paper we will compare the performance of automatic transcription for one type of transcript and research purpose: Our research in the field of science studies is based on semi-structured interviews conducted mostly with researchers with a specialised vocabulary. Our focus lies on the reconstruction of (research or research-related) processes and therefore on the analysis of the content of the conversation. Because we interview researchers who are easily identifiable through their research topics, we also must pay special attention to data protection and anonymity when it comes to sharing data with third parties—such as automated transcription services. [5]

To better understand the issue at hand, we begin by reviewing the literature on different levels of transcription detail that can be considered when creating a transcript. The choice of level depends on the theoretical approach being used, and should be made accordingly. We will outline the research traditions that correspond to each type of transcript (Section 2) as well as the challenges involved in representing speech (Section 3). By doing so, we pave the way for introducing a typology of transcription workflows (Section 4), which we assess based on four criteria: data protection, data quality (accuracy), time spent, and costs (Section 5). Our focus then turns to an empirical comparison of different automatic transcription services. We provide information about the interviews used for this comparison and introduce the automatic transcription tools, followed by a test of their accuracy and the necessary time (Section 6). Finally, we conclude that Whisper performs best based on all the criteria, although it cannot be used for detailed transcripts, and should always be accompanied by subsequent manual review (Section 7). [6]

2. The Level of Transcription Detail and Research Interest

The qualitative interview is a key practice in qualitative social research. Interviews are used as a means of transporting and generating information, such as experiences, know-how, attitudes, and interpretations. Furthermore, an interview can be described as a specific way of interaction that is different from everyday communication. This is because it serves a certain goal: obtaining the interviewee's descriptions regarding a particular issue. There is a wide spectrum of interview types[3], and, depending on the length and on the openness of the interviews conducted, researchers face large amounts of verbal and nonverbal information. To create a systematic basis for subsequent analysis, the researcher must make decisions about the extent and the way in which audiovisual data will be captured and transformed. [7]

Consequently, various ways of transcribing recorded talk are described in the literature. There are four main types (see Table 1). These can be placed on a spectrum ranging from the most detailed possible, taking into account linguistic features (phonetic and Jeffersonian transcription), to a mainly textual reproduction (verbatim and gisted transcription). Each of these four broad types has since been further differentiated and advanced over time as researchers began to locally adapt transcription rules to their specific research questions, methods, or interview types. Decisions are taken at various levels, such as when determining whether non-linguistic events, pauses, intonation, and dialects play a role. In practice, it may well be that one main transcription type is chosen, but elements from another type are used as well. For example, while verbatim is predominant, some language elements may be represented more in the style of Jeffersonian or gisted transcriptions. [8]

There is fairly widespread agreement on the statement that every transcription system reflects a certain methodological approach (KREUZ & RIORDAN, 2011, p.660; LAPADAT & LINDSAY, 1999, p.69) and that by "choosing not to transcribe a particular dimension, the researcher has implicitly decided that the dimension plays no role in the phenomenon in question" (KREUZ & RIORDAN, 2011, p.660). But apart from this claim, the literature lacks a comprehensive and systematic overview of matching transcription styles and research interests, leading DAVIDSON (2009, p.41) to conclude that

> "there is an absence of empirical studies that address how transcription is understood by researchers from within qualitative research, how qualitative researchers relate transcription to their theoretical approaches in specific research projects, or how

3 There are different axes along which interviews can be classified, such as the degree of openness of the questions and corresponding answers (structured, semi-structured, or unstructured/narrative interview), the focus (e.g. problem-centred interview, episodic interview), the medium (face-to-face, video interview, telephone interview, written interview by letter, chat, or email), the number of interviewers and interviewees (e.g. group discussion, group interview, focus group, individual interview), the level of an individual's knowledge (e.g. layperson vs. expert), and contexts of application (e.g. biographical interview, ethnographic interview, oral history) (see, e.g. BRINKMANN, 2014, p.285; EDWARDS & HOLLAND, 2013, p.29; FLICK, 2009, p.149; REICHERTZ, 2016, p.80).

> failure to relate transcription to theoretical assumptions impacts on the achievement of research goals in qualitative research."[4] [9]

And although HAMMERSLEY noted that "[o]ne sharp contrast here is between the very detailed transcripts used by some sociolinguists and conversation analysts, and the much less detailed ones employed by other sorts of qualitative researcher" (2010, p.556), it would be helpful to know more precisely what constitutes the dividing line. [10]

We therefore present in Table 1 the main transcription systems and their level of detail, and try to match them to the corresponding research interest. This overview is thus designed to serve as a basis for classifying what kind of research a person does, what kind of transcripts this kind of research requires, and subsequently whether this kind of transcript can and should be automated at all, at least at this point in time by means of the programs that we will compare in this paper:

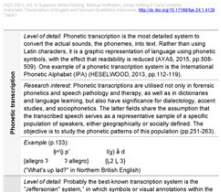


Table 1: Level of detail of transcription systems and corresponding research interests. Click here to open/download the PDF file. [11]

Also relevant in relation to research interests is the question of downward and upward compatibility. While it is theoretically possible to answer questions about content using a Jeffersonian transcript (provided the analysist knows how to read them), it is not possible to carry out full language analyses from a gisted or verbatim transcript. The decision in favour of one transcription system can therefore make certain research, especially secondary research, impossible. Having introduced the various transcription systems by their level of detail, we should also discuss some general points of language representation. [12]

3. The Constructive Nature of Transcription and its Implications

Authors of parts of the transcription literature make the constructivist argument that transcriptions should not be used in research or at least only with reservations, since the transfer of oral communication into a written record is associated with major difficulties. There is no transcription system, even the most detailed, where it is possible to fully transcribe oral communication (KOWAL & O'CONNELL, 2014, pp.65-66). Each system requires decision-making and is selective in choosing what to convey, e.g. breaks, gestures, or accents. [13]

4 This does not apply to conversation analysis, where there is an intensive debate on the relationship between the methods, theory, and research question used to investigate a phenomenon and the form of the transcript as a result (see, e.g. OCHS, 1979).

As a consequence, by using verbatim or gisted transcriptions and eliminating nonverbal communication, interviews can lose their contextual richness, and meanings may be restricted solely to the literal words spoken (LOUBERE, 2017, §8). In contrast, detailed transcripts might distract from the core message of the text (BUCHOLTZ, 2007, pp.786-788). Many researchers have pointed out that the impression created by transcripts can differ greatly from what is actually said (see e.g. COLLINS, LEONARD-CLARKE & O'MAHONEY, 2019, on filler words and their impact on impressions about the interviewee in spoken vs read language; BUCHOLTZ, 2007, on different types of variation inherent in all transcripts; or MISHLER, 2003, on different ways of structuring transcripts and their implications for speech representation). Moreover, there is also a constructive element when it comes to skills and cultural knowledge[5] regarding the interpretation of speech and paralinguistic behaviour on the transcriber's side (HAMMERSLEY, 2010, p.558). Transcription is therefore not a mere transformation of talk and interaction into written symbols, it is "a process that is theoretical, selective, interpretive, and representational" (DAVIDSON, 2009, p.37), meaning that "just by writing the interview down", it has already been analysed in a certain way and its content has been modified. Taken to the extreme, conversations are reproduced in a way that never took place. HAMMERSLEY challenged this notion, noting that if one were to embrace this radical but fundamental concept of a constructed world, it would no longer be possible to understand each other, because there would be no underlying framework to refer to. He concluded that transcripts "are indeed constructed, in an important sense, but they also rely upon what is given when we listen to or watch recordings" (2010, p.563). [14]

An important question emerges from the argument that many constructions are involved in the form of decision-making, influencing the results of research: To what extent can or should such decisions be removed from the control of the original researcher(s) and left to another party, be it another research team, a company providing manual transcription services, or automatic transcription software? One should be aware that all transcribers construct transcripts differently, according to conscious or unconscious standards (TILLEY & POWICK, 2002). Transcription is already an interpretative process, so not doing it yourself could already have a negative impact on your research in the sense that transcriptions can also improve your own interviewing ability in follow-up interviews, and could already set interpretations in motion through memories of the situation (KVALE & BRINKMANN, 2009, p.180). Speaking from an extreme perspective, REICHERTZ even argued that "transcriptions should be made within the research team and never be done by external parties" (2016, p.224, our translation). HALCOMB and DAVIDSON saw it less decisively, but nevertheless concluded:

> "Logically, it may be beneficial for researchers to transcribe their own interview data, given that they have first-hand knowledge from their involvement in the interview

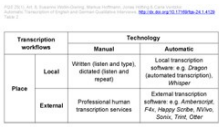
5 Cultural competence in understanding and transcribing language is more attributed to humans, less to AI. AI can probably only try to imitate this.

> process, expertise in the interview subject, and the advantage of having participated in both verbal and nonverbal exchanges with the participants" (2006, p.40). [15]

Awareness of such risks is, however, the first step towards eliminating them, in that the product of an automatic or external transcription must always be subject to reviews and corrections. We provide examples of the potential consequences in Table 5. Any analysis based on unchecked transcripts is at risk of building on potentially made-up statements. Even slight word variations can distort the meaning of an entire sentence. Because original transcripts and recordings usually cannot be shared between researchers due to data protection issues, qualitative research heavily relies on trust in others and the integrity of their data. Thus, any steps to ensure this integrity should be taken. The final act of any transcription, regardless of who or what produced the initial transcript, should therefore always be a review against the recording by someone from the original research team, both for proofreading and for information that cannot be transcribed with the chosen system. If the interviewer and the analyst are different people, we recommend that both of them review the transcript against the record. In the first case, the interviewer can reflect on experiential knowledge from the interview situation; in the second case, the analyst can get a feeling for any non-transcribed content that could be important for the interpretation. [16]

4. Transcription Workflows

In this section, we present a typology of transcription workflows according to the place of transcription and the technology used for it. This typology will help to make general points according to the criteria that we introduce in Section 5, and to position our test candidates for automated transcription in Section 6. We construct a typology of transcription based on two dimensions relevant for our comparison. The first is the *place* of transcription, which can either be *local* or *external*. "Local" refers to transcription processes where the recording stays with the researcher (or group) the whole time. "External" means that the recording is given to a third party for the purpose of transcribing. The second dimension is the *technology* of transcription. Here we distinguish between *manual* and *automatic*. We understand "manual" transcription as that which is produced through immediate action by a human transcriber. This is technologically supported to various degrees. "Automatic" means that the transcription process is delegated to a software algorithm. The combination of these two dimensions results in a 2x2 table, which we use to categorise possible transcription workflows. For each of the four fields in Table 2, we provide examples and discuss their general characteristics. The entries written in italics are part of our empirical comparison in Section 6, while *local-manual* written transcription provides the baseline for comparison:



Transcription workflows		Technology	
		Manual	Automatic
Place	Local	Written (listen and type), dictated (listen and repeat)	Local transcription software: e.g. *Dragon* (automated transcription), *Whisper*
	External	Professional human transcription services	External transcription software: e.g. *Amberscript, F4x, Happy Scribe, NVivo, Sonix, Trint, Otter*

Table 2: Four transcription workflows Click here to open/download the PDF file. [17]

Local-manual transcriptions are transcription procedures in which the recordings do not leave the research context and the transcription is done by a human transcriptionist. Manually typing interview transcripts, meaning either by the researcher him or herself or a colleague involved in the research process, is the most straightforward approach to transcription. Specifically, this means listening to the recording and writing it down. This process is usually supported by using software that facilitates transcription, such as a text editor with a user interface that allows control of the recording (time skips forward/backward, slowing down/speeding up, or the use of individual macros). Examples of common software in the German-speaking context are *F4* and free programs such as *Easytranscript*, *InqScribe*, and *Express Scribe*.[6] Additional tools such as foot pedals can be used to speed up transcription with such software. [18]

A different way to create local transcriptions manually is to use speech recognition software. The first dictation software used in qualitative research in the 2000s was *Nuance Dragon* (MacLEAN et al., 2004; PARK & ZEANAH, 2005). With this program, the transcriber listens to the recording and dictates it into a microphone. The repeated sentences are simultaneously converted into text by the pre-installed voice recognition software. The prerequisite for a good conversion of speech into text is the prior training of the software to the speaker's own voice (PARK & ZEANAH, 2005, pp.246-247), and it runs better on faster computers. We did not include dictation-to-speech-recognition software in our explicit comparison because our focus is on automatic transcription and how it compares to manual transcription in general. Although both dictation and automatic transcription use speech-recognition technologies, the active part of the dictation is attributed to the human. At the core of automatic transcription, no human action is required. Therefore, dictation can be understood as simply a technique to speed up the manual process. [19]

The second category, *local-automatic* transcription, includes ways of transcription where the recording remains with the researchers, but is entirely created by a software algorithm. Two of our test candidates fall into this category. Apart from the dictation function, *Dragon* offers the possibility to automatically transcribe recordings. While this function is still intended to transcribe single-speaker recorded audio files,[7] it is also possible to transcribe recordings of interviews with it. The output is a single-paragraph transcription of the whole interview. [20]

6 More examples of such software can be found in PAULUS et al. (2014, pp.101-108) or here: https://www.sosciso.de/en/software/datenumwandlung/transcription/ [Accessed: May 4, 2023].

7 https://www.nuance.com/dragon/transcription-solutions.html [Accessed: May 12, 2023].

A second example of this type is *Whisper,* a tool by *OpenAI* that was released at the end of 2022. *Whisper* is an automatic speech recognition system based on large language models, which transcribes and translates audio files. *Whisper*'s default function is "speech to text", while additional features can be programmed or are already offered by the user community as additional packages. Examples include a visual user interface and tools that make it possible to derive additional information from the audio data (time stamps, length of breaks between words, speaker assignment).[8] In its basic form, *Whisper* lacks a user-friendly installation process and needs some technical experience to deal with possible difficulties. We provide some basic help to install *Whisper* in Appendix 1: Installing Whisper. [21]

External-manual transcription, the third type, contains all the ways of transcribing that are done by third parties through the immediate action of a human transcriber. This usually means that professional human transcriptionists from external agencies (in most cases, private companies or self-employed one-person businesses) manually transcribe entire interviews. Nowadays, large transcription companies no longer have many permanent staff, but instead draw on a large network of freelance transcriptionists, with the advantage of being able to transcribe particularly quickly.[9] These transcription services may be aimed at a wide audience, such as journalists, researchers, as well as all other people who are concerned with speech-to-text work (for instance, office communication, medical practitioners, media, law) or can be more specific, targeting scholars conducting qualitative research. The latter in particular operate according to established scientific standards: those services apply transcription rules they have developed and which are commonly used in qualitative research. [22]

The last category, *external-automatic* transcription, consists of ways of transcription where the recording leaves the research context and is created by software algorithms. In recent years, providers of automatic speech-to-text generation based on deep learning have mushroomed. Basically, humans train AI-based transcription programs by developing machine-learning algorithms, which in turn are able to handle and semantically decompose natural language when fed with it. After repeatedly feeding these algorithms, the algorithms improve their ability to deconstruct sentences and understand content. When audio files are fed in, the algorithm looks for patterns and matches the audio with the corresponding text. When transcribing automatically, in most cases you upload an audio file to an online speech-to-text cloud software, and the software

8 One promising example of a community-improved version of *Whisper* is *noScribe*. It is easier to use and integrates good speaker assignment. However, it currently only runs on the user's local CPU, resulting in longer computation time: https://github.com/kaixxx/noScribe [Accessed: May 31, 2023].

9 Transcriptions can also be outsourced by sending small parts of the audio files to many people who transcribe them manually (micro-task crowdsourcing). In a later step, the parts are then combined into a single transcript (see, for instance, MARGE, BANERJEE and RUDNICKY, 2010 using *Amazon Mechanical Turk*). The fact that each transcriptionist only receives parts that are unrelated in terms of content seems to initially reduce the risk of data misuse. But since many people receive parts of personal data and some people can already be identified by single words or phrases, this procedure must also be scrutinised regarding data protection (see Section 5.1).

will create a transcript (usually rather quickly). Most of our test candidates fall into this category. [23]

An alternative but closely related version to this is real-time online transcription. One such service is offered by *Otter*, which can transcribe online conversations via Zoom[10] simultaneously. In addition, most smartphones are equipped with either built-in or third-party dictation apps that can perform this function. This means that while it is possible to produce real-time transcripts during an interview, these transcripts nevertheless may be questionable in terms of privacy and data protection. [24]

5. Criteria to Assess Transcriptions

Generally, to find out whether any kind of transcription has advantages or disadvantages compared to a *local-manual* transcription, criteria of comparison are needed. One of our major goals in this paper is to develop a framework that allows researchers to compare transcription workflows and specific programs. Building on that, we can use this framework to empirically compare different providers of automatic transcription with manual transcripts. Based on our experience with general social research practices, we propose four criteria of comparison:[11] *data protection*, *data quality (accuracy)*, *time spent*, and *costs*. These four criteria also structure our comparisons and will first be applied more generally to the transcription workflows introduced above. [25]

5.1 Data protection

In our opinion, the most important category—at least for research with private data—is the question of data protection. Data protection includes issues of maintaining confidentiality vis-à-vis interview partners, the responsible use of interview data and information related to it, and the conduct of research in line with standards of the field and more general frameworks, e.g. the General Data Protection Regulation (GDPR) in the European case. Data protection does not start and end with the interview and its transcription, but must be considered from the point of contacting possible interview partners (informed consent), to methods of analysis, to the storage or deletion of research data. [26]

The guiding principle for this criterion should always be to ensure that no harm befalls our interview partners in any way. What is the point of the most accurate, cheapest, and fastest transcription program if it no longer guarantees confidentiality and the personal rights of interviewees? DA SILVA (2021) noted

10 Zoom is an online video conferencing tool that became widely used during the COVID-19 pandemic and is increasingly used to conduct qualitative interviews. The whole issue about online interviews vs. face-to-face interviews is a discussion for another paper.

11 Given the rapidly changing landscape, this discussion usually happens online. For an academic text concerning this debate, see https://www.researchgate.net/post/What_is_the_best_software_for_transcription_of_interviews2 [Accessed: June 26, 2023]; more generally, see https://www.freedomtoascend.com/tools/best-transcription-software [Accessed: May 5, 2023], https://www.techradar.com/best/best-transcription-services [Accessed: May 5, 2023] and https://geekflare.com/best-transcription-software/ [Accessed: May 5, 2023].

that very little consideration is given to data protection when using and advertising new automatic transcription services for research purposes. To clear up any uncertainties in this respect, it can always be useful to contact the data protection office of your research institution before using automated transcription services. Data protection is therefore relevant for any type of transcription workflow, but creates different challenges for each of them. Especially for external transcriptions, some further points must be considered in addition to everything applying to the *local-manual* case:

- *Local-manual transcription*: Using *local-manual* transcription minimises data protection issues, since it is done by people immediately related to the research process. Nevertheless, measures still need to be taken to ensure the confidentiality of interview partners and that data from the interview cannot be accessed by others (for instance, by ensuring that only participating members of the research team have access to it and by avoiding the use of commercial cloud services for storage).
- *Local-automatic transcription*: Because of their similarity to the *local-manual* type, if recordings and transcripts do not leave the research context, no additional data protection issues arise.
- *External-manual transcription*: Data protection becomes a wider issue here, since personal and possibly sensitive information is transmitted to third parties. If this happens, interviewees must be informed beforehand. A contract guaranteeing confidentiality is usually concluded with the external party to resolve this. It is important to make sure that the contractor does not outsource the transcription to fourth parties with lacking data protection.
- *External-automatic transcription*: Especially in the case of real-time online transcriptions, there is no direct interaction with the external transcribing party in the sense of signing dedicated contracts and confidentiality clauses. In general, this *external-automatic* type is the most concerning of all types regarding data protection. With some providers, it is difficult to find out what happens with the uploaded files and the resulting transcripts. Others are more transparent, but may still be insufficiently safe because they store the data on US servers.[12] If either of these cases apply, *external-automatic* transcription should not be used for private data. [27]

12 In some cases, the EU grants non-EU countries what is known as an "adequacy decision", which means that the country can ensure an adequate level of protection for personal data through their own legislation and international agreements. In 2020, the European Court of Justice ruled that the previous adequacy decision between the USA and the EU (EU-US Privacy Shield) was invalid, because US data protection did not conform to European data protection standards and excessive interference with personal data by US authorities was possible. For this reason, the European Commission has strongly recommended against transferring data to servers in the US until a new agreement is made https://ec.europa.eu/commission/presscorner/detail/en/ip_22_7631 [Accessed: May 5, 2023]; this is likely to be the case soon. In contrast, the EU approved an adequacy decision for the UK in 2021 after the UK's withdrawal from the EU, thus ensuring the free flow of personal data between the EU and UK in compliance with the GDPR until 2025 https://commission.europa.eu/law/law-topic/data-protection/international-dimension-data-protection/brexit_en [Accessed: June 10, 2023].

5.2 Data quality (accuracy)

Data quality is the second most important category when deciding how to produce transcripts. This is simply because the worse a raw transcript is, the more time has to be spent on correction work afterwards. Too much correction work can ultimately lead to zero net advantage in either time or monetary cost. [28]

The first and more general dimension of data quality is the previously introduced *level of transcription detail* (Section 2). To estimate the fit of transcription workflows to different levels of detail, it is necessary to know whether additional (extra-)linguistic information can be included in the transcript. Examples include interaction (e.g. overlapping talk), pauses, intonation, phonetic representation (pronunciation particularities, preservation of dialects vs. conversion into colloquial language or even formal language), characteristics of interview participants (e.g. utterances, word interruptions, elisions, uncertainty), prosodic cues (stress, pitch, loudness, tempo, and/or elongation of words or syllables), punctuation (e.g. for direct speech, questions, fade-outs), extra-linguistic events (laughing, throat clearing, gestures), and speaker segmentation (FUß & KARBACH, 2019; MOORE, 2015). Depending on what kind of information is needed in the transcript, certain ways of or tools for producing transcripts may be excluded from further consideration. [29]

The second dimension of data quality concerns the accuracy of the intended transcription tool; assessing this obviously helps to estimate the amount of correction effort after the transcript has been generated. One way to do this is by looking at the word error rate (WER), a common measure in the literature on automatic speech recognition (VON NEUMANN, BOEDDEKER, KINOSHITA, DELCROIX & HAEB-UMBACH, 2022), to indicate the proportion of (in)correctly transcribed words in a text. Some authors have already addressed the accuracy of automated transcriptions by using the WER: BOKHOVE and DOWNEY (2018) compared manual with *YouTube* transcripts; MOORE (2015) looked at how well conversation-analysis transcripts could be created by using IBM's "Attila" speech recognition engine; and LIYANAGUNAWARDENA (2019) compared the performance of six online automatic transcription services with the same text read out by different speakers. Some automatic transcription services themselves claim to have accuracy rates of 85% (*Happy Scribe*, *Amberscript*), or even 90% for high-quality recordings (*NVivo*). Some vary their reported accuracy depending on the language, with 95.5% in English, 95.8% in Italian, and 96.5% in Spanish (*Whisper*); one service in our study simply mentioned its "top-notch accuracy" (*Sonix*).[13] We can expect that the higher the accuracy in a raw transcript is, the less time will be necessary for subsequent correction.

13 Happy Scribe: https://www.happyscribe.com/automatic-transcription-software [Accessed: May 4, 2023]; Amberscript: https://www.amberscript.com/en/products/transcription [Accessed: May 4, 2023]; NVivo: https://www.lumivero.com/products/nvivo-transcription [Accessed: May 4, 2023]; Whisper: https://www.assemblyai.com/blog/how-to-run-openais-whisper-speech-recognition-model [Accessed: May 4, 2023]; Sonix: https://www.sonix.ai/languages [Accessed: May 4, 2023].

- *Local-manual transcription*: *Local-manual* transcription is the most flexible type regarding level of detail, and it should be easier to produce a transcript exactly in the way it is needed than with the other three workflows. Accuracy depends on the transcriber, the audio/video quality, and the time spent on transcription.
- *Local-automatic transcription*: Currently, all available automatic tools only offer transcripts on a level of detail between verbatim and gisted, which means that they are not adequate for purposes of conversation analysis or similar approaches. The various tools perform differently in terms of accuracy in creating such transcripts; this point is the topic of the next section. The quality of the recording also plays a role in the output.
- *External-manual transcription*: Since the transcription is done manually, different levels of transcription detail and proofreading are available at corresponding price rates. With some experience over time, it should be possible to find a transcription provider that produces transcripts at the desired level of detail and accuracy.
- *External-automatic transcription*: All currently available automatic tools only offer transcripts at a level of detail between verbatim and gisted, which means that they are not adequate for purposes of conversation analysis or similar approaches. Accuracy depends on the specific tool and the recording. Since the level of detail is the same for all automatic transcription tools introduced later, we concentrate on accuracy as the only dimension of data quality in the comparison. [30]

5.3 Time spent

The main argument against transcribing interviews manually is the considerable length of time involved. GLÄSER and LAUDEL observed that in their own research projects, it took an average of six hours to transcribe one hour of interview material for verbatim transcription (2010, p.193). KVALE and BRINKMANN estimated five hours per hour of interview (2009, p.180). In more general terms, the ratio is somewhere between "4-60 hours per hour of audio or video recording, depending on the format used for transcription" (EVERS, 2011, §47). Obviously, detailed transcriptions take longer, especially if one needs to do more iterations, or mutual reviews and revisions. [31]

Depending on the necessary level of detail, outsourcing transcription tasks to other people or to technology can, at best, free up time that researchers would otherwise spend on manual transcription work. This allows researchers to focus on other aspects of their work, such as performing more comprehensive analyses or conducting more interviews than they might have been able to do otherwise (MOORE, 2015, p.269). In research projects where information from one interview needs to be analysed quickly to provide a basis for further interviews, a slow transcription procedure could either delay research or provide less information than desired. If there are time savings, automatic transcription could become a significant tool. However, any raw transcript requires additional correction, and the time needed for this must also be considered as part of the

time spent, and this amount of time also depends on the accuracy of the initial raw transcripts.

- *Local-manual transcription*: *Local-manual* transcription usually takes the longest, although it depends on the experience of the transcriber and possible acceleration via technical means. If the analyst is the person transcribing, the necessary reviews against the audio can be directly incorporated into the transcription process.
- *Local-automatic transcription*: The big advantage of automatic transcription and the motivation to use it is the possibility to outsource work to a computer. Nevertheless, a certain amount of additional time to check the accuracy of the initial text is always necessary. A closer look into these types of mistakes and the time to rectify them is provided in the next section. Depending on the program used and the local hardware the program is running on, producing the first raw transcript also takes processing time, but the researcher can also usually use that time doing something else.
- *External-manual transcription*: The speed at which you can get transcripts back usually depends on the agreement you have with your transcription service. For larger projects, it is not uncommon to collect several recordings in a batch instead of submitting them one-by-one after each interview; this also, however, creates bottlenecks. In the best case, the incoming transcript can be used for analysis immediately—after the necessary review by the analyst.
- *External-automatic transcription*: The time taken to produce the transcript is very short. But like *local-automatic* transcription, the accuracy of the raw transcript affects the amount of additional time needed for correction. [32]

5.4 Costs

Transcriptions require financial resources. Those costs usually come in the form of salaries for internal transcribers, payment of external transcription providers, or fees for services that offer automatic transcription. In addition, the researcher must account for costs of necessary hardware, although this should only apply once. These costs are not entirely independent of each other, since getting a particularly cheap deal for fast external transcription might come around in the form of longer correction times—and thus someone's salary or free time.

- *Local-manual transcription*: In our framework, we have included personnel costs attributable to the transcriber (the researchers themselves, or colleagues such as student assistants), as well as possible extra costs for software and hardware to facilitate manual transcription.
- *Local-automatic transcription*: This type of transcription usually requires good hardware to run properly. Different programs, however, have different requirements. In our particular cases, this meant, for instance, providing enough RAM for *Dragon*, while running *Whisper* needs a good graphics card with at least 8GB of dedicated RAM. Another alternative would be to have a

second computer running the transcription software for a longer time. Depending on the software, one-time licence fees might also apply.

- *External-manual transcription*: Agencies usually charge per interview minute. Typical rates for interviews in German are between €1.20/min and €1.60/min. But this would only apply for a gisted to verbatim level of detail, near-perfect recording quality, one or two speakers, and standard language.[14] The general rule is: the longer the interview and the higher the necessary level of detail, the more expensive it will be.
- *External-automatic transcription*: Prices for this type also depend on the particular tool, but since the programs run on third-party computers, no local powerful hardware is necessary. Different payment models exist; the three most common are a rate according to interview minute, a rate per hour, or monthly/annual plans with a set contingent of "free" interviews for transcription. Examples are provided in Table 3. [33]

6. Empirical Comparison

In this section, we use parts of an English and a German interview to compare several automatic transcription tools with *local-manual* written transcription. We introduce the tools and our method for comparison, present our results for accuracy and time spent, and discuss those together with data protection and costs. [34]

6.1 Properties of the tools under review

In Table 3 we introduce the candidates for our empirical comparison. The baseline for the comparison is *local-manual* written transcription, supported by a standard freeware transcription software and done by an experienced transcriber. Seven of our nine automatic transcription candidates (*Amberscript*, *F4x*, *Happy Scribe*, *NVivo*, *Sonix*, *Trint*, and *Otter*) fall into the *external-automatic* category, and are thus relevant for additional data protection issues. For those instances, we noted down where the servers are located, if they promise to comply with GDPR standards, and what happens with the uploaded audio data and transcripts. Some online services do not use their own servers but rely on cloud computing services such as Amazon Web Services (AWS). [35]

The different tools provide support for various languages and sometimes come with additional functionality, either online, as an extra download, or with automatic transcription as part of a wider package of other software (for instance, in *Dragon* and *NVivo*). All these automatic transcription tools, except *Whisper* and *Dragon,* charge fees for the transcription of entire interviews; the free trials are only useful for testing the programs. Recall that most automatic transcription can currently only produce a level of detail between verbatim and gisted. Transcripts with more demanding levels of detail need to be created manually, although attempts are

14 For instance, see https://www.transkripto.de/transkriptionsservice [Accessed: April 10, 2023] or https://www.meintranskript.de [Accessed: April 10, 2023].

also being made to achieve a Jeffersonian level of detail automatically (MOORE, 2015).

	Data protection issues	**Languages**	**Additional functions beyond automatic transcription**	**Requirements**
Written (Local-manual)	Location: - Server: - GDPR & data: depends on local practice	Depends on the transcriber	(No automatic transcription) formal review can be integrated	Costs: salary, transcription tools Setup: any transcription software
Amberscript (External-automatic)	Location: Netherlands Server: EU GDPR: yes Data: used to improve algorithms, opt-out possible	German, English, 37 other languages	Manual transcription; subtitling services (manual & automatic); transcript editor	Costs: €11-20/hour, depending on the plan; €2.55/minute (manual); 10 min free trial Setup: browser registration; no installation needed
Dragon 15 (Local-automatic)	Location: USA Server: - GDPR & data: depends on local practice	German, English, 9 other languages	Dictation; transcript editor	Costs: full version €699; certain minimum hardware requirements before installation can proceed;[15] Setup: download & installation
F4x 2019 (External-automatic)	Location: Germany Server: DE GDPR: yes Data: deleted immediately after transcription	German and English	Transcript editor; tool for qualitative analysis	Costs: €5-12/hour, with lower-cost options for students; extra cost for editor & analysis tool; 30 min free trial Setup: browser registration; no installation for automated transcription only

15 The minimum hardware requirements *for dictation* can be found here. Especially for running the automated transcription, we recommend higher hardware requirements, https://nuance.custhelp.com/app/answers/detail/a_id/27849/~/system-requirements-for-dragon-15-home [Accessed: May 23, 2023].

	Data protection issues	**Languages**	**Additional functions beyond automatic transcription**	**Requirements**
Happy Scribe (External-automatic)	Location: Ireland Server: EU GDPR: yes Data: not specified	German, English, 60 other languages	Manual transcription; subtitling services (manual & automatic); translation; transcript editor	Costs: €0.20/minute (automatic), €3.50/minute (manual); 10 min free trial Setup: browser registration; no installation needed
Nvivo[16] (External-automatic)	Location: USA Server: EU (for Africa, Europe & Middle East) GDPR: yes Data: deleted 90 days after transcription	German, English, 26 other languages	Transcript editor; tools for qualitative analysis and collaboration	Costs: €27-33/hour, yearly plan; 15 min free trial Setup: browser registration, cumbersome account setup; no installation required for automated transcription only
Sonix (External-automatic)	Location: USA Server: Amazon Web Services (AWS) cloud platform region West (USA) GDPR: no Data: deleted if user deletes it	German, English, more than 33 other languages	Automatic subtitling; tool for collaboration; transcript editor; real-time transcription forthcoming	Costs: $10/hour, $5/hour + $22/month; 30 min free trial Setup: browser registration; no installation needed

16 *NVivo* is primarily a computer-assisted qualitative data analysis software (CAQDAS) that also includes transcription in its package. Providers of other CAQDAS programs, such as *Transana*, https://www.transana.com/blog/2023/03/25/automated_transcription/ [Accessed: October 12, 2023] have recently begun to offer transcription services as well. This appears to be a growing trend. While we cannot cover all current and future candidates, the criteria proposed here can assist in considering their potential use.

	Data protection issues	**Languages**	**Additional functions beyond automatic transcription**	**Requirements**
Trint (External-automatic)	Location: UK Server: Amazon Web Services (AWS) cloud platform (not specified) GDPR: yes Data: kept for recovery if user deletes it	German, English, 29 other languages	Transcript editor	Costs: €44-65/month depending on the plan; 7-day free trial Setup: browser registration; no installation needed
Otter (External-automatic)	Location: USA Server: USA GDPR: yes Data: used to improve algorithms	English	Tool for collaboration; live transcription; transcript editor	Costs: $8.33-20/month; free monthly plan with limits Setup: browser registration; no installation needed
Whisper (Local-automatic)	Location: USA Server: - GDPR & data: depends on local practice	German, English, many other languages	Translation into English	Costs: free software; good hardware recommended (min 8GB graphics card) Setup: no registration; complex installation; community applications available

Table 3: Features of the transcription tools under review [36]

6.2 Measures for comparison

So far, there are only fragmentary statements provided in the literature on the extent to which automatic transcriptions are useful, for instance regarding their use for conversation analysis (MOORE, 2015) or the provision of accessibility (LIYANAGUNAWARDENA, 2019).[17] On the one hand, this is because we are in a field of permanent development: new services with different features and platforms are emerging all the time. On the other hand, automated transcription services are reluctant to make statements about transferability: what produces

17 Comparisons of transcription services mainly take place on internet blogs, see https://www.medium.com/descript/comparing-the-accuracy-of-automatic-transcription-services-519fec134465 [Accessed May 4, 2023] or https://www.theopennotebook.com/2019/12/17/say-what-a-non-scientific-comparison-of-automated-transcription-services/ [Accessed: May 4, 2023].

good transcription results for one person's specific research needs is not necessarily transferable to other types of research or other subjects. [37]

For the following comparison, we use our own research as an example. It is based on qualitative semi-structured (online and face-to-face) expert interviews conducted with researchers, (research-related) policymakers, or research managers, with a focus on practices and processes of the recent past. In semi-structured interviews, talk is structured in a certain way. For example, the conversation generally alternates, but with the number of mutual interruptions and overlapping speech somewhat lower than casual conversations. Usually, these interviews include one interviewer and one interviewee. Since the transcripts function more as a source of information than a source of meaning, a verbatim to slightly gisted level of detail is enough, but, as analysts, we would also note content-relevant verbal or nonverbal incidents within the interview situation. We conduct interviews in German and English and have both native and non-native speakers with their accompanying accents, although standard language without dialects and without slang is the normal case. Our interviews often include sophisticated and technical language when it comes to explaining research processes and this also creates an interesting area to test automatic transcription. Interviewing researchers requires special care regarding data protection and anonymity, since individuals can be easily identified due to the small circles of researchers working in these fields, their specific job structures, and especially individual or group-specific research topics (LAUDEL & BIELICK, 2019). [38]

Two interviews serve as the basis for our comparison. The first interview (in English, between a native English speaker and an English-speaking German) was conducted by a colleague of ours and is a supervision talk with another researcher. It was done during the COVID-19 pandemic and thus conducted via Zoom and recorded with the explicit permission to use it for this purpose. The second interview (in German, both German natives) was conducted by one of the authors of this paper and is part of an ongoing project. We received permission from the interviewee to use the interview for this additional purpose. We employ English and German interviews for two reasons: first, these are the languages of our research; second, automatic transcription programs are usually trained in English, with German being a secondary, often less accurate but frequently included language. While our candidates might perform better in other languages than German, we can only expect worse results compared to English. Automatic transcription tools developed for specific languages might work better in those languages, but our language skills and European context limits us to these two. [39]

There are other differences besides the language in the two interviews. Those include the gender of the speakers, the speed and style of speaking, and the sound quality, as well as the content of what is said. This should be kept in mind for the following analysis. To become more familiar with the interviews, we have included the parts that are provided in the comparison as audio snippets for the reader in Appendix 2: Audio Files. [40]

We conducted our manual transcription with the help of *F4*transkript, at a slightly gisted level of detail, and with minimal annotations, time stamps, and speaker names. The manual transcript serves as the basis for the comparison. It has the format, speaker separation, correct spelling, and punctuation we would usually need for further extractive qualitative content analysis (GLÄSER & LAUDEL, 2019). We then used the Microsoft Word text comparison function to compare each software transcript with the manual transcript, and produced a text containing markups for each comparison (Table 4). These texts contain words that do not appear in the manual transcript but that do appear in the automatic transcript of the respective software (substituted or inserted words: crossed out and red). Also included are words that occur in the manual transcript but not in the automatically created transcript (deleted words: underlined and red). Black words represent correct transcription. We used these markup texts to measure proxies of accuracy, create error profiles, and track the time spent aligning the automatic transcript with the written transcript:

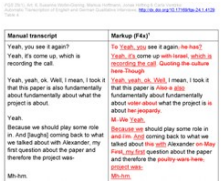


Table 4: Comparison of manual transcript with an automatically generated transcript. Click here to open/download the PDF file. [41]

6.2.1 Accuracy proxies

As introduced before, we mainly use the word error rate (WER), which is a standard measure in linguistic analysis. The WER is the quotient of all words that were incorrectly inserted, deleted, or substituted, divided by the total number of words in the manually created reference transcript. The word accuracy measure is the inverse of this rate (1-WER). [42]

To measure the accuracy of the different automatic transcriptions, we counted the three different types of words in the markup document for each transcript (substituted/inserted, deleted, correct). Punctuation marks were ignored in the comparison. We also normalised the transcripts insofar as we decided that whenever transcription services transcribed words in a validly alternative way in comparison to the manual reference, we would not count this as a transcription error (e.g. "we'll" instead of "we will", "going to" instead of "gonna", "2" instead of "two", "acknowledgments" instead of "acknowledgements", "Okay" instead of "OK", etc.). Moreover, alternative sentence division that does not affect the sentence content was also not included in the error rate (e.g. "you go to the bottom, you make a convincing argument" and "you go to the bottom. You make a convincing argument"). [43]

Based on this approach, we came up with four accuracy proxies for each program:

1. the percentage of words not found by the transcription program (deletions);
2. the percentage of words the transcription program "made up" or misunderstood (insertions and substitutions);
3. the word accuracy measure (1-WER);
4. the average length of the longest correct sequence per paragraph. [44]

6.2.2 Error profiles

In addition, it is necessary to understand the nature of common mistakes. Since the word error rate is not sensitive to word meanings (FAVRE et al., 2013), the absence of meaningful words such as nouns is valued the same as the absence of more meaningless words such as articles. That is why specifically looking at the mistakes and categorising them helps to understand which transcription tools make which kind of errors. We inductively found the following types:

- similar-sounding words or word groups;
- misunderstood proper nouns;
- missing single words (meaningless);
- missing single words (meaningful);
- missing sub-sentences;
- made-up words (not similar);
- word endings wrong / same word stem;
- spelling mistakes;
- wrong speaker assignment. [45]

Similar-sounding word groups can change the whole meaning of a sentence, whereas incorrect spelling usually does not lead to major misunderstandings. Missing sub-sentences make understanding the context hard in comparison to missing single, meaningless words. Of course, automatic transcription services with high rates of missing words and sub-sentences have in turn low rates of similar-sounding words, since the possible number of misunderstood terms is lower. Thus, we can see that a comparison between the programs based on only one type of error does not seem to make much sense. Rather, we must compare the overall error profile between programs. [46]

Determining which kinds of errors are more important or less disturbing in a transcript ultimately depends on the intended analysis and corresponding research question. For instance, for most verbatim or gisted transcripts (where the main interest is the content of the interview), missing meaningless words or incorrect spelling often do not matter much. Ultimately, however, we can only provide an idea of which programs produce which errors. It is up to the user to decide how much those errors matter. For our research, we can say, for instance, that misunderstanding proper nouns or the occurrence of similar-sounding words

or word groups would be a problem, because interviewing researchers involves a specific vocabulary where correct understanding of these terms is crucial for the subsequent analysis. [47]

6.2.3 Time spent

The last thing we systematically tested is the time needed to produce a sufficient (meaning equal to *local-manual*) transcript. This is composed of two components: the time needed to produce the transcript, and the time taken to rectify errors from automatic transcription. This gives an indication of how much time can be saved by using automatic transcription, which is ultimately one of the main motivations to use it. [48]

We might assume that automatic transcription does not take any time at all, precisely because it is automated. Nevertheless, automatic transcription services also need a certain amount of time to create a transcript. With online services, this is a combination of the time taken to upload the source file, which depends on your internet access, and the actual automatic transcription conducted on the servers of the online service. For offline programs, this is the processing time of the local computer, which is affected by computer performance. For both types, some time is also necessary to sign up for or install the program. However, because transcription is a repetitive task, we think it is more important to measure the time for recurring work instead of the initial time spent installing the program. [49]

To this end, we requested two trained transcriptionists (student assistants who had already transcribed about 20 1.5-hour qualitative interviews each) to correct all the transcripts of the services compared here.[18] For our research situation, we can accurately assess whether it makes sense to use automatic transcription in combination with our experienced transcriptionists to save time, or whether manual transcription with transcriptionists is faster.[19] [50]

18 The corrections were done by loading the text file of the transcript into *Easytranscript* instead of using the online editors of each service to create a common basis for the comparison.

19 In a study with over 200 measurements from different interviews, *F4x* tested the time saved with its own automatic transcription service, including correction. The study distinguished between students and experienced transcriptionists: For manual transcription, the students needed on average 6.3 times the interview duration, whereas for correcting the automatic transcriptions, they needed on average 5.1 times the time. This saved 70 minutes of correction time per hour of interview material (advantage of 19%). The experienced transcriptionists transcribed almost as quickly as they corrected. To sum up: the slower one transcribes manually, the more one benefits from *F4x*: http://web.archive.org/web/20220725003200/https://www.audiotranskription.de/f4x/ [Accessed: June 1, 2023].

6.3 Results

Our initial results are based on accuracy proxies and error profiles for our English and German interviews. To give an idea of the transcripts, we show parts of the transcriptions of each interview, which can also be found as .mp3 files in Appendix 2. We address the issue of the time needed in the "Discussion" section below. We begin with the English interview. [51]

6.3.1 Accuracy of the English interview

The English interview consists of 23 paragraphs in which the interviewer and interviewee take turns to speak. The average number of words per paragraph is 34, the shortest paragraphs are only one or two words long (mostly comments like "mhm"), and the longest paragraph contains 284 words. The 5-minute interview clip contains a total of 787 words. In comparison to the German interview, there are fewer speaker changes and longer monologue-like passages, and it also seems to be spoken more slowly. [52]

In these text snippets from the English interview (Table 5), we can already see that for most automatic services (except *Dragon* and *F4x*), there is a good possibility to understand the content. Overall, although there are some mistakes, all English transcripts were consistently better than the German ones. This is because speech models are usually based on English training files, and there are also differences in our audio material such as the gender and the speed of the speaker, the topic, style, or dialects.

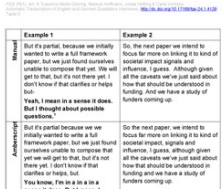


Table 5: Examples from the English transcripts. Click here to open/download the PDF file. [53]

In terms of *accuracy* (Figure 1), the automatic transcription programs seem to be rather accurate. Most services achieve a word accuracy measure of around 85%. Overall, *Otter*, *Trint*, *Sonix*, *NVivo*, *Happy Scribe,* and *Amberscript* do not seem to vary too much. As already seen in the small text examples, this high level of accuracy does not, however, extend to *F4x*, at 59%, and *Dragon*, at 54%.[20] *Dragon* in particular has a high rate of deletions, meaning that many words are not included in the transcript. With a high number of substitutions or insertions, you not only have to add many words when correcting transcripts, but also need to replace wrong words or word groups with correct ones. By far the highest word accuracy measure was achieved by *Whisper*, at 93%. It also has the lowest rate

20 The programs under study are also constantly being developed and improved. We ran most of our test in 2022, meaning that some of our candidates already had newer (and supposedly better) versions available. For instance, *F4x* has had a new engine since October 2022, and promises 50% improved recognition, https://www.audiotranskription.de/en/f4x/ [Accessed: June 4, 2023] and *Dragon* is also already available in Version 16.

of deletions compared to the other programs. Looking at the average length of the longest sequence of correct words per paragraph shows something about the distribution of errors. The higher this number, the longer you can expect to find correctly transcribed parts of the interview. In Figure 1, while the blue percentages relate to the performance of the programs in general, the orange dots are an interview-specific number, and can show us only something about how the tools performed in relation to each other. Across all four proxies for accuracy, *Whisper* performed best, meaning that it had the lowest rate of deletions, substitutions, and insertions, the highest word accuracy measure, and the longest correct strings of words in the English interview. All others apart from *Dragon* and *F4x* performed relatively well.

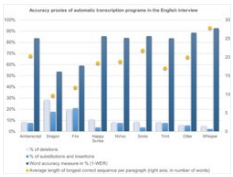


Figure 1: Accuracy proxies of automatic transcription programs in the English interview. Please click here for an enlarged version of Figure 1. [54]

Looking at the *error profiles* in Figure 2, we see that one error occurs quite frequently in the transcripts of *Dragon* and *F4x* and to some extent also in *Otter*, *NVivo*, and *Trint*: similar-sounding words or word groups are created by the algorithms. This often changes the content of the text and is time-consuming to correct, as it requires listening to the audio carefully. One example from Table 5 is "Kenneth societal impact" or "Canada societal impact" instead of "kind of societal impact". Other examples include "is a shocking news" instead of "as we've shown, you lose"; "poultry" or "poetry" instead of "project"; "condos" or "hunters" instead of "funders"; "embellishments" or "admonishments" instead of "acknowledgements"; or "they mix" instead of "they are mixed". [55]

While the previously mentioned programs had problems with similar-sounding word groups, the transcripts of *Amberscript*, *Dragon*, *Happy Scribe,* and to a certain extent *Sonix* simply lacked many words. Around half of the missing words were neglectable and had no significant meaning in the sentence. Some of the programs just leave out filler words, such as "you know" or stuttering. Missing significant words were usually verbs or subjects. Missing sub-sentences, except for *Dragon*, and to a lesser extent in *F4x* and *Happy Scribe*, are only a minor problem of these automatic transcription services and mostly consisted of a maximum of three words. In comparison to the German interview, missing sub-sentences generally occurred less frequently. [56]

Made-up words, i.e. words that are simply created by the algorithm but whose presence cannot be explained by similar-sounding words in the recording, are especially common in *Dragon* and *F4x,* and still to a high degree in *Amberscript*. All programs apart from *Whisper* were unable to recognise proper nouns correctly. Out of the five names in the transcript, *Whisper* was able to recognise

every proper noun, although some were rather special or unusual. At least one name was consistently understood by all programs. Also, spelling was only an issue for *Happy Scribe*. *Happy Scribe* twice wrote "Funders" instead of "funders" in the middle of a sentence. This is only a minor problem for corrections and for our purposes does not make a difference for the analysis.

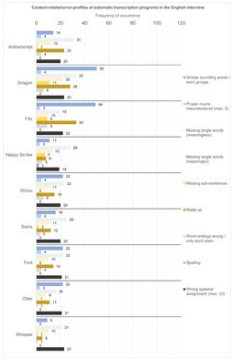


Figure 2: Content-related error profiles of automatic transcription programs in the English interview. Please click here for an enlarged version of Figure 2. [57]

All transcriptions suffered from poor separation between paragraphs. *Amberscript*, *Happy Scribe*, *NVivo*, *Sonix,* and *Trint* reliably separate the transcript into Speakers 1 and 2, but come up with only a few—five to six (instead of 23)—paragraphs in total, because they simply ignore or filter out intermediate remarks such as "Yeah" or "Mhm". In the worst case, this can result in loss of meaning; in less severe cases, they make the transcript less readable (though sometimes also more readable). *Otter* demonstrated the same problem of very few separated paragraphs, but instead of separating speakers, titled each paragraph with "unknown speaker". *F4x* created a large number of paragraphs without a titled speaker assignment by splitting sentences halfway through, while *Whisper* did not separate between speakers and created paragraphs after sentences and sub-sentences. Intermediate remarks such as "Yeah" or "mhm" were left out. In *Dragon,* there was no speaker or paragraph subdivision, and only a single paragraph in total. For programs that cannot assign speakers, separating the transcript into the right paragraphs will take more time. [58]

6.3.2 Accuracy of the German interview

The German interview consists of 37 paragraphs in which the interviewer and interviewee take turns to speak. The average number of words per paragraph is 27, the shortest paragraphs are only one or two words long (mostly comments like "ja" or "ok"), and the longest paragraph contains 174 words. The 5-minute interview clip contains 984 words. In comparison to the English interview, there are more speaker changes, shorter passages, a faster speed of speaking, and more sub-sentences. Many of the sentences were not finished or were interrupted, and there were frequent changes in grammar between the beginning and the end of a sentence. [59]

In this example, we can see that for most automatic services (apart from *Dragon*), it is at least possible to understand what is being spoken about. Nevertheless, it is necessary to check the transcripts against the audio in all cases. The results of automatic transcriptions of the German interview (Table 6) range from incomprehensible to unintentionally funny to fairly okay:

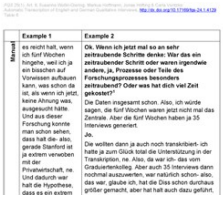


Table 6: Examples from the German transcripts. Click here to open/download the PDF file. [60]

Figure 3 shows our *proxies for accuracy* in the German interview. *Dragon's* word accuracy measure is below 5%, whereas *Whisper* reached above 70%. *Sonix* surpassed 60%, *Trint*, *NVivo*, *Happy Scribe,* and *Amberscript* were around 50%, and *F4x* was slightly above 30%. A more complete picture of the actual correction effort can be obtained by adding the percentages of missing and wrong words together. Although the effort to delete made-up words in *Dragon* seems to be small with approximately 10%, substitutes must still be found for the large number of missing words (deletions). Here, we can see that a low percentage of wrongly substituted/inserted words is only relevant together with a low rate of deletions. With the exception of *Dragon*, most tools have a deletion rate of around 30%, with a wrong word rate of 10-20%. *F4x* performs worse in both cases, *Sonix* is better with deletions (making wrong words more relevant), and *Whisper* performs best in all regards. When looking at the number of correct words per paragraph, the picture does not really change, but shows that *Sonix* performs better than the other online transcription services. We also found much shorter correct word strings than in the English interview. Just as with the English interview, *Whisper* performed best in our comparison, *Dragon* made the most mistakes, and *F4x* came in second to last.

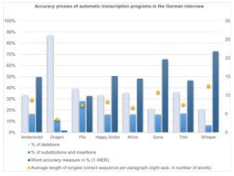


Figure 3: Accuracy proxies of automatic transcription programs in the German interview. Please click here for an enlarged version of Figure 3. [61]

Looking at the *types of errors* individually (see Figure 4), we can see that *Dragon* again stands out. It has so many missing sub-sentences that few other errors are left. For all other tools, missing meaningless words and wrongly understood similar words were the most common reasons for mistakes. Of the *external-automatic* tools, *Sonix* again performed best, while having a generally similar

error profile to the others. *Whisper* arguably came out on top again, although *Sonix* performed better in four of the error categories.

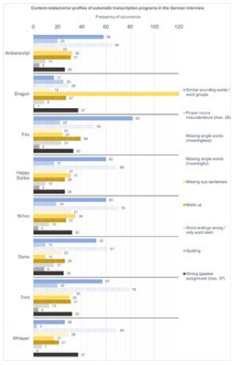


Figure 4: Content-related error profiles of automatic transcription programs in the German interview. Please click here for an enlarged version of Figure 4. [62]

Once again, all transcriptions suffer from the problem that paragraphs and speakers are not correctly assigned. Like in the English interview, the programs ignore or filter out intermediate remarks such as *jetzt* [now], *glaub ich* [I think] or "mhm", which results in the large number of mistakes in the "meaningless single word" category, and a far lower number of total paragraphs. We also find that all programs apart from *Whisper* were unable to recognise proper nouns correctly. Altogether, there were 28 proper nouns (some reoccurring) in the original transcript (three German surnames, seven names of places, 16 words connected to the research object, two words related to methodology, and two to university structures). At the two poles were *Whisper* at the top, with only three mistakes, and *Dragon* with the worst result, with 25 mistakes out of the 28. Although spelling was again not much of a problem (though still more common than in the English transcripts), *Sonix* performed particularly poorly; there were nine instances of spelling mistakes compared to a maximum of five in the other programs. Examples include lower-case spelling of nouns (e.g. "leute", "hervorhebung", "das zentrale", "förder context") or the separation of actually connected compound words (e.g. "förder context", "extrem trennung"). While these spelling errors would require correction, they can be understood and, in some cases, left in for the analysis. Because of the overall larger number of errors, we would expect the German interview to need more correction time. This is discussed in the next section. [63]

6.3.3 Accuracy compared

While the two interviews are not entirely comparable (due to speaker speed, native/non-native speakers, longer paragraphs), the results show that, across the board, all programs performed better in the English interview. It is very unlikely that this can all be explained by the differences just mentioned. While we had word accuracy measures of around 80-90% for the English interview (with the exception of *Dragon* and *F4x*), the transcripts of the German interview were around 50%, with only *Sonix* and *Whisper* performing better. The same trend can

be seen in the average length of the longest sequence of correct words per paragraph. While the paragraphs are generally shorter in the German interview, the English transcripts provide much longer parts of the interview without errors. *Dragon* is the least reliable automatic transcription tool in both cases, but we must remember that this functionality is intended to transcribe single-speaker recordings for trained voices. This exercise has shown that it should also only be used for that purpose. [64]

This makes *F4x* the second most poorly performing program, especially in the English context, where all other tools performed much better. Of the *external-automatic* tools, *Sonix* returned slightly better results than the others, but all in all they performed reasonably well in the English interview, and with a great deal of room to improve in the German interview. *Whisper* was consistently, in almost all accuracy proxies and error categories, in the top spot. The only thing it really could not do is correct speaker assignment. The output is always one sentence or sub-sentence per paragraph. This is partially solved through community plugins.[21] [65]

6.3.4 Time spent

We measured two times for our interviews and transcription tools. The first was the time it took to create the initial transcript, while the second was the time needed to correct the transcript (Table 7):

Time spent for a 3-minute interview part[22]	**Transcription (English interview)**	**Transcription (German interview)**	**Correction (English interview)**	**Correction (German interview)**
Local-manual (experienced transcriptionist)	10 min, 30 sec	11 min	3 min, 10 sec	3 min, 10 sec
Amberscript	40 sec	2 min, 35 sec	4 min, 35 sec	8 min, 10 sec
Dragon	1 min	1 min	7 min, 30 sec	11 min, 15 sec
F4x	1 min, 40 sec	2 min, 15 sec	6 min, 50 sec	10 min, 45 sec
Happy Scribe	1 min, 10 sec	1 min, 25 sec	3 min, 40 sec	8 min, 25 sec
NVivo	2 min	2 min	3 min, 40 sec	8 min, 5 sec
Sonix	40 sec	3 min	3 min, 50 sec	7 min, 40 sec

21 For instance, *noScribe*, https://github.com/kaixxx/noScribe [Accessed: June 10, 2023], *WhisperX*, https://github.com/m-bain/whisperX [Accessed: June 10, 2023] or *aTrain*, https://business-analytics.uni-graz.at/en/research/atrain/ [Accessed: November 26, 2023].

22 The transcription times change with interview duration in a nonlinear fashion, meaning that the transcription of a 30-minute interview segment will not necessarily take 10 times the duration of a 3-minute interview excerpt. For example, for a 15-minute excerpt it took *Amberscript* 5 minutes (English & German); *Dragon* 5 minutes (English & German); *F4x* 9 (English) and 10 minutes (German); *Happy Scribe* 3 (English) and 9 minutes (German); *NVivo* 6 (English) and 8 minutes (German); *Sonix* 13 (English) and 15 minutes (German); *Trint* 8 minutes (English & German); *Otter* 6 minutes (English); and *Whisper* 284 (English, CPU) and 364 minutes (German, CPU), 13 (English, GPU 8GB) and 18 minutes (German, GPU 8GB), or 3 (English, GPU 10GB) and 4 minutes (German, GPU 10GB).

Time spent for a 3-minute interview part	**Transcription (English interview)**	**Transcription (German interview)**	**Correction (English interview)**	**Correction (German interview)**
Trint	2 min, 55 sec	3 min, 10 sec	3 min, 45 sec	8 min, 55 sec
Otter	55 sec	-	4 min, 5 sec	-
Whisper (model large-v2)	30-61 min (CPU), 2 min, 45 sec (8GB GPU), 50 sec (10GB GPU)	40-84 min (CPU), 3 min, 35 sec (8GB GPU), 1 min (10 GB GPU)	3 min, 30 sec	6 min, 30 sec

Table 7: Time spent transcribing and correcting a 3-minute sound file [66]

In several correction passes, which were necessary to control for the learning effect that occurred during the repeated correction of the same interviews,[23] two of us, who are experienced manual transcribers, took the following times to correct: The English interview took between 3.5 and 7.5 minutes of correction time. The programs did not differ significantly in this respect. *Whisper* was corrected slightly more quickly, in only 3 minutes and 30 seconds. This is nearly as low as the amount of time needed for a review of the manual transcript. [67]

The German interview roughly took between 8 and 11 minutes to correct, with the exception of *Whisper,* which took 6.5 minutes. The longer correction time was mainly caused by more errors in the German transcripts. The quality of the audio file also played a role, both indirectly through the number of errors in the transcript and directly for easier understanding during the correction process. [68]

In both the English and German cases, we found that using the programs indeed saved time in comparison to manual transcription. And for the English interview it became clear that when we add together the time for setup, transcription, and correction, automatic transcription is always faster than manual transcription. In the case of *Whisper*, we can really expect considerable time savings for transcription in both German and especially English, although the local hardware setup makes a big difference for the computation time. [69]

23 Correction times will be slightly higher in real-world research scenarios since transcribers are not expected to work with the same interview repeatedly.

6.4 Discussion

We can now see what kind of errors our different candidate tools produce and how long it takes to create a proper transcript with them. Together with the information on data protection issues, requirements, and functionality of the tools (Table 3), we can assess how useful and appropriate they are. If you need to transcribe sensitive data, i.e. most interview data that are collected in qualitative social research, *data protection* should be your priority. Both *Dragon* and *Whisper* —as *local-automatic* transcription tools—can be used without further consideration. To be on the safe side, if it is either not clear 1. where servers are located, 2. if GDPR compliance is assured, or 3. what happens to the data, it is better to not use external services with non-public data. If servers are in the US, or if data is shared or reused, it is also not advisable to use the program. This leaves only *Amberscript* (with opting out) and *F4x* as viable options. This perspective is, of course, European. While our colleagues in the US and elsewhere work under different regulatory frameworks, this does not imply that they care less about data protection. *Local-automatic* approaches to transcription are always a safer choice, and some external services offer sufficient data protection. We encourage everybody to take privacy concerns seriously and uphold research standards of our field over legal requirements. [70]

We already described each program's *accuracy* and the *time* needed. *Whisper* and *Sonix* performed best with accuracy proxies and error profiles. These differences were more strongly pronounced in the German transcript, while the English one worked rather well for all tools (with the consistent exception of *Dragon* and *F4x*). The time needed is directly related to the accuracy. Time savings can be observed by almost all tools, with more time needed to transcribe and correct the German interview. If no high-performance hardware is available, the transcription time for *Whisper* is much longer, but can still be outsourced to a second computer running in parallel. [71]

Finally, considering *costs* and requirements of the different tools, we have a range from free software and good hardware (*Whisper*), to expensive, single payment (*Dragon*), to different pay-as-you-go or monthly subscription models, which become more expensive with the more interviews you have (all others). If we take a qualitative research project with 60x1.5h interviews as an example, this would result in total costs ranging from €450 (*F4x*) to €2,430 (*NVivo*), with the other providers in between and some monthly plans where the payment depends on the spread of the workload (*Trint* and *Otter*). *Whisper* costs only as much as the hardware it will run on and the accompanying electricity costs, which might be kept in mind if a second computer needs to work on transcription for days. Time and personnel costs for checking the transcript apply for all tools. [72]

These four elements are not equally relevant. If our interview participants must worry about their own sensitive data and information, then the trust necessary for an interview can break down. In this respect, the highest priority is data protection. Every type of external transcription must address this element before it can be used. The second priority concerns accuracy. The more accurate the

transcription, the less time is lost for corrections, the faster the transcript can be reintegrated into the research process and contribute to it. A quickly generated transcript can, for instance, be more effectively used in an iterative process of questionnaire adaptation to check whether all relevant information was collected in the interview. The time taken for transcription is connected to accuracy and is also a relevant decision criterion. Finally, the cost should be taken into consideration. The question of cost plays a major role—especially for students and early career researchers, where funding is often precarious—but ethical considerations still must not be ignored. [73]

Taken together, we found that *Whisper* proved to be the best candidate for automatic transcription for our purposes. To see if this net gain of time holds up for whole interviews, we started to use *Whisper* in our own project. For the first seven interviews, the total time of transcription work per interview hour ranged from 2 hours 15 minutes to 17 hours. This *included* the time *Whisper* needed to create the initial transcript with various setups. If we assume that the computing time can be used for something different, we arrive at 1 hour 30 minutes to 5 hours 5 minutes per interview hour, which is still below the average *local-manual* transcription time of 6 hours (GLÄSER & LAUDEL, 2010, p.193), or at least the same with 5 hours (KVALE & BRINKMANN, 2009, p.180). This particular interview was in German, with a fast-speaking researcher who used extensive technical vocabulary paired with a tendency to not complete sentences correctly. With more favourable interview conditions, the accuracy of *Whisper* is better in both English and German, resulting in much faster times. Our average for the seven interviews was 3 hours 30 minutes of correction time per interview hour. [74]

While manual transcription is also faster with easily understandable interview speakers, there is a certain limit to how quickly people can transcribe. How easily the audio can be understood appears to make a bigger difference for the accuracy and subsequent correction time of automatic transcription. This means that if you use a high-quality audio file (meaning both the recording and the clarity of the spoken text in it), automatic transcription has the potential to be much faster. [75]

Table 8 shows the summarised results of all transcription tools based on our four criteria for transcription. A green arrow pointing upwards means that the tool performed well in the respective criterion; an orange arrow pointing to the right means a moderate result with room for improvement; and a red arrow pointing downwards refers to significant insufficiencies. For data protection, this means how well the tool/method can be used while ensuring that no harm comes to the interview partner. Accuracy refers to the quality of the raw transcript. Time spent and costs should be understood relatively between the nine tools and written *local-manual* transcription:

Table 8: Summarised results. Click here to open/download the PDF file. [76]

7. Conclusion

Our goals in this paper were to systematically compare different automated transcription tools with manual transcription, to provide a framework to do so, to argue why some tools can be used for certain kinds of research, and what researchers should keep in mind while doing so. No automated transcription result should be used in the original form it is produced in. Some kind of manual review is always necessary afterwards, the duration and effort of which depends on the reliability of the initial output. We introduced four criteria to assess the adequacy of approaches to transcription: 1. data protection and privacy issues; 2. the quality of transcripts, including level of detail and accuracy; 3. the time needed; and 4. the costs and requirements. These four points are not equally important. If personal data of interviewees is shared with unknown third parties, even the best automated transcription is not worth the possible irreparable damage. If a free tool provides poor results, it also does not help much. In the end, the relevance of each criterion depends on the specific research process and researcher. [77]

Our results and experience indicated that *Whisper* performed best. First and foremost, it runs locally, and there are therefore no issues regarding data protection beyond those arising from having the audio file or transcript on your computer. Second, it was the most accurate of our candidates. In almost every category for both interviews, *Whisper* produced the best results. Third, this high degree of accuracy also translates into a relatively short time to review the transcript. With English interviews, the transcripts were often almost comparable to the final read through of a manual (verbatim to gisted) transcript. The runtime for the actual transcription is where *Whisper* is both slower than the online tools but still much faster than manual transcription—provided it can run on a good graphics card. If appropriate hardware is available, you can have *Whisper* running in the background while doing other things. This brings us to the fourth point, which is the only apparent disadvantage of *Whisper*. While the software packages themselves are all free, you need to get a good graphics card (more specifically, one with at least 8GB RAM) and a corresponding PC to run it on, especially if you hope to work on that computer at the same time. In contrast to the other candidates, this hardware would be a one-time investment and thus should pay off quite quickly. The only other tool where a single payment is enough is *Dragon*—which, at least in its automatic transcription mode, produced disappointing—but often funny—results. [78]

The *external-automatic* transcripts of third-party services meanwhile also delivered good results and will certainly continue to be improved. This rapidly changing landscape makes the specific results of this exercise prone to be outdated quickly. But we can still conclude with several points that should remain relevant for the longer term: [79]

First, the approach we proposed to evaluate transcription tools can be useful to compare new tools in the future. How important *data protection*, *accuracy*, *time needed*, and *cost requirements* are usually depends on the individual situation of the researcher. For qualitative social science research working with non-public data, data protection should come first. To better operationalise our comparisons of the different programs, we also introduced different proxies for accuracy and error profiles. [80]

Second, while *Whisper* might be overtaken in terms of performance and requirements in the future, the general workflow of *local-automatic* transcription seems to be the future trend.[24] Because it can combine accurate transcripts, (in the best case) low costs, and no extra issues for data protection, programs in this category are and will continue to be good candidates. [81]

For the third point, as of now, automatic transcription only works at a *level of detail* that is *between verbatim* and *gisted* transcripts. This means that nonverbal information about the interview (breaks, laughs) and disfluencies in speech are often not transcribed. If the research question is more focused on the ways something is said, rather than its content, or meaning-making in the interview situation, automatic transcription of the kind tested in this paper does not provide a high enough level of detail. [82]

Finally, despite the powerful large-language models we can now use to automate transcription, they will always need a review against the recording by a human before analysis. The implementation of automatic transcription programs always has certain biases towards finishing sentences or deleting filler words, which might or might not change the content. However, to be able to determine that, a person with enough knowledge about the epistemic purpose of the transcripts needs to listen to the audio at least once and adjust the transcript accordingly. [83]

In our own research, we have recently begun to use *Whisper* for the transcription of our interviews and can report considerably faster high-quality transcription of interviews. With this paper, we hope to give other practitioners a useful overview of what might be worth a try, and what might not. [84]

24 We can already observe the integration of *Whisper* in other software packages, such as *Quirkos*, https://www.quirkos.com/learn-qualitative/qualitative-automated-transcription.html [Accessed: October 12, 2023], a tool for qualitative analysis. They offer *external-automatic* transcription by locally using *Whisper* in combination with encrypted transfer of audio files and transcripts.

Acknowledgements

We thank Grit LAUDEL and Jochen GLÄSER for feedback and the members of our regular research group meeting for fruitful discussions. Furthermore, we thank the two interviewees for allowing us to use parts of the recordings for this purpose, Elaheh AHMADI for help with the preceding discussion paper, Teresa GEHRS and Matt REES for excellent proofreading made possible through TU Berlin, and the editor and reviewers for helpful comments.

Appendix 1: Installing Whisper

Since *Whisper* has performed best in our comparison, but is not as easy to set up as other automatic transcription tools, we want to provide some leads to do so. Click here to download the PDF file with tips for the installation.

Appendix 2: Audio files

Because this paper is about the quality of automatically generated transcripts, we want to include the recordings for reference for the interested reader. We asked our interviewees for explicit permission to include not only the transcriptions but also the audio snippets in the publication. The recordings correspond to the transcripts in Table 5 and Table 6 and can be downloaded here.

References

Ayaß, Ruth (2015). Doing data: The status of transcripts in conversation analysis. *Discourse Studies*, *17*(5), 505-528.

Bokhove, Christian & Downey, Christopher (2018). Automated generation of "good enough" transcripts as a first step to transcription of audio-recorded data. *Methodological Innovations*, *11*(2), https://doi.org/10.1177/2059799118790743 [Accessed: July 17, 2023].

Brinkmann, Svend (2014). Unstructured and semi-structured interviewing. In Patricia Leavy (Ed.), *The Oxford handbook of qualitative research* (pp.277-299). Oxford: Oxford University Press.

Bucholtz, Mary (2007). Variation in transcription. *Discourse Studies*, *9*(6), 784-808.

Collins, Harry; Leonard-Clarke, Willow & O'Mahoney, Hannah (2019). "Um, Er", How meaning varies between speech and its typed transcript. *Qualitative Research*, *19*(6), 653-668.

Da Silva, Joseph (2021). Producing "good enough" automated transcripts securely: Extending Bokhove and Downey (2018) to address security concerns. *Methodological Innovations*, *14*(1), https://doi.org/10.1177/2059799120987766 [Accessed: July 17, 2023].

Davidson, Christina (2009). Transcription: Imperatives for qualitative research. *International Journal of Qualitative Methods*, *8*(2), 35-52, https://doi.org/10.1177/160940690900800206 [Accessed: July 17, 2023].

Easton, Lee; Lexier, Roberta; Lindstrom, Gabrielle & Yeo, Michelle (2019). Uncovering the complicit: The disrupting interview as a decolonising practice. In Lynn Quinn (Ed.), *Re-imagining curriculum: Spaces for disruption* (pp.149-169). Stellenbosch: African Sun Media.

Edwards, Rosalind & Holland, Janet (2013). *What is qualitative interviewing?* London: Bloomsbury.

Evers, Jeanine C. (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*(1), Art. 38, https://doi.org/10.17169/fqs-12.1.1636 [Accessed: July 17, 2023].

Favre, Benoit; Cheung, Kyla; Kazemian, Siavash; Lee, Adam; Liu, Yang; Munteanu, Cosmin, Nenkova, Ani; Ochei, Dennis; Penn, Gerald; Tratz, Stephen; Voss, Clare & Zeller, Frauke (2013).

Automatic human utility evaluation of ASR systems: Does WER really predict performance? *Proceedings Interspeech*, 3463-3467.

Flick, Uwe (2009). *An introduction to qualitative research* (4th ed.). Los Angeles, CA: Sage.

Fuß, Susanne & Karbach, Ute (2019). *Grundlagen der Transkription, Eine praktische Einführung* (2nd ed.). Opladen: Barbara Budrich.

Gläser, Jochen & Laudel, Grit (2010). *Experteninterviews und qualitative Inhaltsanalyse als Instrumente rekonstruierender Untersuchungen* (4th ed.). Wiesbaden: VS Verlag für Sozialwissenschaften.

Gläser, Jochen & Laudel, Grit (2019). The discovery of causal mechanisms: Extractive qualitative content analysis as a tool for process tracing. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *20*(3), Art. 29, https://doi.org/10.17169/fqs-20.3.3386 [Accessed: November 26, 2023].

Halcomb, Elizabeth J. & Davidson, Patricia M (2006). Is verbatim transcription of interview data always necessary? *Applied Nursing Research*, *19*(1), 38-42.

Hammersley, Martyn (2010). Reproducing or constructing? Some questions about transcription in social research. *Qualitative Research*, *10*(5), 553-569.

Heselwood, Barry (2013). *Phonetic transcription in theory and practice*. Edinburgh: Edinburgh University Press.

Jefferson, Gail (1985a). An exercise in the transcription and analysis of laughter. In Teun A. van Dijk (Ed.), *Handbook of discourse analysis. Volume 3: Discourse and dialog* (pp.25-34). London: Academic Press.

Jefferson, Gail (1985b). On the interactional unpackaging of a "gloss". *Language in Society*, *14*(4), 435-466.

Jefferson, Gail (2004). Glossary of transcript symbols with an introduction. In Gene H. Lerner (Ed.), *Conversation analysis: Studies from the first generation* (pp.13-31). Amsterdam: John Benjamins Publishing Company.

Korobov, Neill (2001). Reconciling theory with method: From conversation analysis and critical discourse analysis to positioning analysis. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *2*(3), Art. 11, https://doi.org/10.17169/fqs-2.3.906 [Accessed: July 17, 2023].

Kowal, Sabine & O'Connell, Daniel C. (2014). Transcription as a crucial step of data analysis. In Uwe Flick (Ed.), *The Sage handbook of qualitative data analysis* (pp.64-78). London: Sage.

Kreuz, Roger J. & Riordan, Monica A. (2011). The transcription of face-to-face interaction. In Wolfram Bublitz & Neal R. Norrick (Eds.), *Foundations of pragmatics* (pp.657-679). Berlin: De Gryuter.

Kvale, Steinar & Brinkmann, Svend (2009). *InterViews: Learning the craft of qualitative research interviewing* (2nd ed.). Los Angeles, CA: Sage.

Lapadat, Judith C. & Lindsay, Anne C. (1999). Transcription in research and practice: From standardization of technique to interpretive positionings. *Qualitative Inquiry*, *5*(1), 64-86.

Laudel, Grit & Bielick, Jana (2019). Forschungspraktische Probleme bei der Archivierung von leitfadengestützten Interviews. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *20*(2), Art. 10, https://doi.org/10.17169/fqs-20.2.3077 [Accessed: November 26, 2023].

Liyanagunawardena, Tharindu (2019). Automatic transcription software: Good enough for accessibility? A case study from built environment education. *EDEN Conference Proceedings*, *Volume*(1), 388-396.

Loubere, Nicholas (2017). Questioning transcription: The case for the systematic and reflexive interviewing and reporting (SRIR) method. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 18(2), Art. 15, https://doi.org/10.17169/fqs-18.2.2739 [Accessed: July 17, 2023].

MacLean, Lynne M.; Meyer, Mechthild & Estable, Alma (2004). Improving accuracy of transcripts in qualitative research. *Qualitative Health Research*, *14*(1), 113-23.

Marge, Matthew; Banerjee, Satanjeev & Rudnicky, Alexander I. (2010). Using the Amazon mechanical turk for transcription of spoken language. In Institute of Electrical and Electronics Engineers (IEEE) (Eds.), *2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2010)* (pp.5270-5273). Red Hook, NY: Curran Associates, Inc., https://doi.org/10.1109/ICASSP.2010.5494979 [Accessed: November 26, 2023].

Markle, D. Thomas; West, Richard Edward & Rich, Peter J. (2011). Beyond transcription: Technology, change, and refinement of method. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *12*(3), Art. 21, https://doi.org/10.17169/fqs-12.3.1564 [Accessed: October 12, 2023].

Mishler, Elliot G. (2003). Representing discourse: The rhetoric of transcription. In Yvonna S. Lincoln & Norman K. Denzin (Eds.), *Turning points in qualitative research: Tying knots in a handkerchief* (pp.297-326). Lanham: AltaMira Press.

Moore, Robert J. (2015). Automated transcription and conversation analysis. *Research on Language and Social Interaction*, *48*(3), 253-270.

Ochs, Elinor (1979). Transcription as theory. In Elinor Ochs & Bambi B. Schieffelin (Eds.), *Developmental pragmatics* (pp.43-73). New York, NY: Academic Press.

Oliver, Daniel G.; Serovich, Julianne M. & Mason, Tina L. (2005). Constraints and opportunities with interview transcription: Towards reflection in qualitative research. *Social Forces*, *84*(2), 1273-1289.

Park, Julia & Zeanah, A. Echo (2005). An evaluation of voice recognition software for use in interview-based research: A research note. *Qualitative Research*, *5*(2), 245-251.

Paulus, Trena; Lester, Jessica & Dempster, Paul (2014). *Digital tools for qualitative research*. London: Sage.

Psathas, George & Anderson, Timothy (1990). The "practices" of transcription in conversation analysis. *Semiotica*, *78*(1-2), 75-99.

Reichertz, Jo (2016). *Qualitative und interpretative Sozialforschung: Eine Einladung*. Wiesbaden: Springer VS.

Rutakumwa, Rwamahe; Mugisha, Joseph Okello; Bernays, Sarah; Kabunga, Elizabeth; Tumwekwase, Grace; Mbonye, Martin & Seeley, Janet (2020). Conducting in-depth interviews with and without voice recorders: A comparative analysis. *Qualitative Research*, *20*(5), 565-581. https://doi.org/10.1177/1468794119884806 [Accessed: July 17, 2023].

Sacks, Harvey; Schegloff, Emanuel A. & Jefferson, Gail (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, *50*(4), 696-735.

Tilley, Susan A. & Powick, Kelly D. (2002). Distanced data: Transcribing other people's research tapes. *Canadian Journal of Education / Revue canadienne de l'éducation*, *27*(2/3), 291-310.

von Neumann, Thilo; Boeddeker, Christoph; Kinoshita, Keisuke; Delcroix, Marc & Haeb-Umbach, Reinhold (2022). On word error rate definitions and their efficient computation for multi-speaker speech recognition systems. *arXiv*, https://doi.org/10.48550/arXiv.2211.16112 [Accessed: July 17, 2023].

Authors

Susanne WOLLIN-GIERING is currently working at Technical University Berlin in the projects "FUFAF – Functions and Consequences of Unemployment in Researchers Careers" and "EPAC – Effects of Pandemic-related Disruptions on Academic Careers". In her work she focuses on the intersection between the sociology of work and the sociology of science, particularly in the area of academic careers.

Contact:

Susanne Wollin-Giering

Sozialwissenschaftliche Wissenschafts- und Technikforschung
Technische Universität Berlin
Straße des 17. Juni 135, 10623 Berlin, Germany

Tel.: +49 30 314 27377

E-Mail: susanne.wollin-giering@tu-berlin.de
URL: https://www.tu.berlin/en/sos/about/team/susanne-wollin-giering-ma

Markus HOFFMANN is a research associate and PhD candidate at the Technical University Berlin in the research group for social studies of science and technology. He is currently working on two projects to investigate the field-specific effects and ways of handling unemployment and restrictions induced by the COVID-19 pandemic respectively. His research interests include field-comparative science studies and qualitative approaches of analysing content.

Contact:

Markus Hoffmann

Sozialwissenschaftliche Wissenschafts- und Technikforschung
Technische Universität Berlin
Straße des 17. Juni 135, 10623 Berlin, Germany

Tel.: +49 30 314 26596

E-Mail: markus.hoffmann@tu-berlin.de
URL: https://www.tu.berlin/en/sos/about/team/markus-hoffmann-ma-ma

Jonas HÖFTING is currently a Master's student in Information Science at Humboldt University Berlin. He supports the projects "FUFAF – Functions and Consequences of Unemployment in Researchers Careers" and "EPAC – Effects of Pandemic-related Disruptions on Academic Careers" as a student assistant at Technical University Berlin.

Contact:

Jonas Höfting

Sozialwissenschaftliche Wissenschafts- und Technikforschung
Technische Universität Berlin
Straße des 17. Juni 135, 10623 Berlin, Germany

E-Mail: j.hoefting@tu-berlin.de

Carla VENTZKE supported the projects "FUFAF – Functions and Consequences of Unemployment in Researchers Careers" and "EPAC – Effects of Pandemic-related Disruptions on Academic Careers" as a student assistant at Technical University Berlin.

Contact:

Carla Maria Ventzke

Senatsverwaltung für Inneres und Sport
Klosterstr. 47, 10179 Berlin, Germany

E-Mail: c.ventzke@posteo.de

Citation

Wollin-Giering, Susanne; Hoffmann, Markus; Höfting, Jonas & Ventzke, Carla (2024). Automatic transcription of English and German qualitative interviews [84 paragraphs]. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *25*(1), Art. 8, https://doi.org/10.17169/fqs-25.1.4129.