frontiers | Research Topics

Capturing talk: The institutional practices surrounding the transcription of spoken language

Edited by
Felicity Deamer, Helen Fraser, Kate Haworth, Martha Komter, Debbie Loakes and Emma Richardson

Published in
Frontiers in Communication
Frontiers in Psychology

ISSN 1664-8714
ISBN 978-2-8325-4933-9
DOI 10.3389/978-2-8325-4933-9

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Capturing talk: The institutional practices surrounding the transcription of spoken language

Topic editors
Felicity Deamer — Aston University, United Kingdom
Helen Fraser — The University of Melbourne, Australia
Kate Haworth — Aston University, United Kingdom
Martha Komter — Netherlands Institute for the Study of Crime and Law Enforcement (NSCR), Netherlands
Debbie Loakes — The University of Melbourne, Australia
Emma Richardson — Loughborough University, United Kingdom

Citation
Deamer, F., Fraser, H., Haworth, K., Komter, M., Loakes, D., Richardson, E., eds. (2024). *Capturing talk: The institutional practices surrounding the transcription of spoken language*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4933-9

Table of contents

TYPE Editorial
PUBLISHED 08 May 2024
DOI 10.3389/fcomm.2024.1417465

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Mila Vulchanova,
NTNU, Norway

*CORRESPONDENCE
Helen Fraser
✉ helen.fraser@unimelb.edu.au

RECEIVED 15 April 2024
ACCEPTED 22 April 2024
PUBLISHED 08 May 2024

CITATION
Fraser H, Haworth K, Deamer F, Loakes D, Richardson E and Komter M (2024) Editorial: Capturing talk: the institutional practices surrounding the transcription of spoken language. *Front. Commun.* 9:1417465.
doi: 10.3389/fcomm.2024.1417465

Editorial: Capturing talk: the institutional practices surrounding the transcription of spoken language

Helen Fraser[1]*, Kate Haworth[2], Felicity Deamer[2], Debbie Loakes[1], Emma Richardson[2,3] and Martha Komter[4]

[1]Research Hub for Language in Forensic Evidence, The University of Melbourne, Parkville, VIC, Australia, [2]Aston Institute for Forensic Linguistics, Aston University, Birmingham, United Kingdom, [3]Department of Communication and Media, School of Social Science and Humanities, Loughborough University, Loughborough, United Kingdom, [4]Netherlands Institute for the Study of Crime and Law Enforcement (NSCR), Amsterdam, Netherlands

KEYWORDS

transcription, misconceptions about language and linguistics, language ideologies, forensic linguistics, forensic transcription, police interviews and interrogations, entextualization

Editorial on the Research Topic
Capturing talk: the institutional practices surrounding the transcription of spoken language

Transcripts are a ubiquitous feature of virtually all modern institutions, many of which would be unable to function without them. Nevertheless, transcription remains an under-researched subject—a situation that *Capturing talk: the institutional practices surrounding the transcription of spoken language* seeks to remedy.

The initial aim of this Research Topic was to expose and examine under-appreciated features of "entextualization" (the process of representing spoken language as written text). One of these features is the fact that a transcript can only ever be a representation of speech, not a copy—and thus can never represent speech exactly. Another feature, well-articulated by Sarangi (1998), is the unequal power over the process of transcription exercised by, on the one hand, the speakers whose voices are represented, and, on the other, by those controlling the transcription process.

Where Sarangi's interest was mainly in health and social services institutions, the present Research Topic has a leaning toward legal institutions, where, arguably, these power inequalities are even more starkly contrasted—as demonstrated by the territory-defining volume (Heffer et al., 2013).

Four of the papers in this Research Topic deal with police interviews, providing insight into differing practices across jurisdictions and type of interview (e.g., whether with witnesses or suspects). Several papers examine the practice of converting an interview into a "statement," written up by the officers who conduct the interviews. Beginning with interviews with witnesses in England and Wales (E&W), Milne et al. analyze a sample of such statements against transcripts produced by the researchers from an audio recording. The omissions, additions, distortions, and other errors in the police versions give cause for deep concern.

An extended study analyzing the creation of records of interviews with suspects in the Netherlands is recounted by Komter, which, again, contrasts transcripts prepared by police interviewers, with the author's transcripts prepared from audio recordings. Again, many concerning limitations on the police transcripts are observed and analyzed. However, while her own transcripts are far more detailed, Komter acknowledges that she too is necessarily selective in what she chooses to represent, guided by the evolving research questions she seeks to investigate.

One practice Komter discusses is that of police records presenting an interview as a monolog, in the voice of the interviewee, rather than as the question-and-answer dialogue it actually was. This practice is also investigated by Eerland and van Charldorp, again focusing on the Dutch context. These authors study how readers of the statements were influenced by three different styles of reporting (monolog, dialogue and narrative), with the troubling finding that the style of reporting affected perceptions of the statements' accuracy and comprehensibility.

In many jurisdictions, police interviews with suspects are routinely audio- or video-recorded. However, this does not signal the end of problems with the representation of these high-stakes interactions. The last of our interview papers is Haworth et al., which summarizes the key findings to date of an ongoing study of the transcription of electronic records of interviews with suspects in E&W. It demonstrates a range of problems with official police transcripts even when these ostensibly capture the dialogue "verbatim," and proposes that consistency, accuracy, and neutrality are the foundational features that should underpin any police interview transcript.

A second group of papers studies transcription in non-legal institutional settings. Holder et al. delves into two very large and highly structured organizations with serious security needs: NASA and the US Military. Both make extensive use of audio and video recordings capturing employees as they work—with transcripts produced either routinely, or on demand. The authors look into the two organizations' use of these transcripts, again comparing the official transcripts with their own transcripts of selected sections, using conversation analysis (CA) conventions.

Park and Hepburn also examine CA-style transcripts. Taking as an example Rachel Mitchell's interview of US Supreme Court nominee Brett Kavanaugh about his alleged historical sexual misconduct, these authors compare the information retrievable from a richly detailed Jeffersonian transcript with an orthographic transcript that "wipes out" or "skates over" crucial aspects of speech used by speakers and listeners in constructing the message expressed by the speech.

Another institutional use of transcripts covered in *Capturing Talk* concerns workers on the assembly line of a small factory in Sweden. Carlsson and Harari report an observation-and-interview study of the instruction manuals created by the workers. While they find much to commend in the retention of power by the creators and users of the manuals, the authors observe room for improvement in the "information design" of the texts, recommending that consultation of linguistics experts could offer benefits.

Voutilainen showcases the high quality of transcripts produced as an official record of the complex and challenging multicultural discussions of wide-ranging Research Topics covered by the parliament in Finland. His account demonstrates how much thought, research and work goes into managing all the factors that need to be considered to create transcripts of this standard.

In a return to the legal setting, a further group of papers examines transcripts of forensic audio, i.e., recordings of speech used as evidence in criminal trials. These are often of very poor quality, meaning that the transcript is intended not as a record of what was said, but as assistance to the court in determining what was said. Internationally, it is common for such transcripts to be provided by police investigating the case. While the courts recognize that police transcripts might contain errors, they rely on judges and/or juries being able to check the transcript against the audio. This ignores well-established research findings that the very act of checking a transcript can cause the listener to hear in line with the transcript, even if it is demonstrably false. For this reason, linguists sometimes recommend that, to ensure accuracy, transcripts should be produced by independent experts in transcription.

However, mere independence may not be enough, and Love and Wright point out some important caveats around this recommendation. They had eight trained transcribers produce transcripts of poor-quality forensic-like audio—finding huge divergences in the content of the transcripts (<3% of conversational turns were transcribed consistently by all eight participants). This demonstrates that transcribing poor-quality forensic audio needs not just expertise in linguistics, but a managed, evidence-based method.

Recently, a common response to any discussion of the difficulty of transcribing poor-quality audio has been: "Why not let AI do it?" Loakes investigates this suggestion, finding that, while modern automatic speech recognition (ASR) systems are extremely efficient at transcribing good-quality audio, their performance on poor-quality forensic-like audio is low. Even the best-performing system, Whisper, scored only around 50% accuracy, with others far lower.

Harrington also observed low scores for ASR transcripts of poor-quality forensic-like audio. Bridging two of the main areas considered in this Research Topic, she also trialed ASR on recordings of police interviews. The resulting transcripts, though not problem-free, score far higher than those of covert recordings, with errors easier to identify. Harrington makes innovative recommendations for how ASR could be used as a "first draft" interview transcript, to be refined via human transcribers.

Two papers consider the transcription and translation of forensic audio featuring languages other than English. Gilbert and Heydon look at translated transcripts of Vietnamese recordings used as evidence in a drug-related trial. They point out significant errors in the translations, but note that, unless the defense goes to the expense of hiring their own translator/interpreter, such errors are unlikely to be detected—and suggest that audio in languages other than English is often admitted with inadequately tested translations.

Lai presents results of a large national survey of the practices and concerns of translators and interpreters who undertake forensic casework across a wide range of languages. Here, too, results indicate a number of important deficiencies in current practice for translating forensic audio featuring languages other than English—and Lai makes valuable recommendations for improvement.

Finally, taking an authoritative overview of the key issues relevant to this Research Topic, Fraser provides a systematic review of interdisciplinary research on transcripts and transcription, and sets out a series of interacting factors that are known to affect a transcript's reliability. Using examples from a range of legal and academic situations, Fraser argues that, to ensure a transcript is suitable for its intended purpose, it is essential that all the factors be appropriately managed.

Taken as a whole, *Capturing Talk* amplifies two observations made in both Sarangi (1998) and Heffer et al. (2013), which, though not the exclusive focus of any individual paper, are highlighted throughout the Research Topic. First, the strong role that context inevitably plays in the interpretation of a transcript implies that "recontextualization" (using a transcript in a context other than the one it was created in) is likely to change its interpretation. Second, even the most expert linguistic analysis of transcripts produced by others is not itself a neutral or "objective" activity. However, this does not mean that such analysis must be "subjective" in any limiting sense. Rather it indicates a need for transcripts to be produced and analyzed by independent, context-aware experts able to devote appropriate attention to all relevant factors.

Most importantly, all contributions to *Capturing Talk* emphasize that transcription is far from the simple transduction of "sounds" into letters that it is often assumed to be by those who have not studied its intricacies. It is a highly complex and fascinating Research Topic worthy of taking its place as a dedicated field of research in its own right, particularly in view of the widespread misconceptions and unhelpful language ideologies that still beset the institutional practices surrounding the transcription of spoken language.

Author contributions

HF: Writing – original draft, Writing – review & editing. KH: Writing – review & editing. FD: Writing – review & editing. DL: Writing – review & editing. ER: Writing – review & editing. MK: Writing – review & editing.

Funding

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

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Heffer, C., Rock, F., and Conley, J. (2013). *Legal-Lay Communication: Textual Travels in the Law*. Oxford: Oxford University Press.

Sarangi, S. (1998). Rethinking recontextualization in professional discourse studies: an epilogue. *Text Talk* 18, 301–318. doi: 10.1515/text.1.1998.18.2.301

frontiers in Communication

ORIGINAL RESEARCH
published: 03 December 2021
doi: 10.3389/fcomm.2021.779227

Translated Transcripts From Covert Recordings Used for Evidence in Court: Issues of Reliability

David Gilbert and Georgina Heydon *

Social and Global Studies Centre, School of Global, Urban and Social Studies, RMIT University, Melbourne, VIC, Australia

Nation states increasingly apply electronic surveillance techniques to combat serious and organised crime after broadening and deepening their national security agendas. Covertly obtained recordings from telephone interception and listening devices of conversations related to suspected criminal activity in Languages Other Than English (LOTE) frequently contain jargon and/or code words. Community translators and interpreters are routinely called upon to transcribe intercepted conversations into English for evidentiary purposes. This paper examines the language capabilities of community translators and interpreters undertaking this work for law enforcement agencies in the Australian state of Victoria. Using data collected during the observation of public court trials, this paper presents a detailed analysis of Vietnamese-to-English translated transcripts submitted as evidence by the Prosecution in drug-related criminal cases. The data analysis reveals that translated transcripts presented for use as evidence in drug-related trials contain frequent and significant errors. However, these discrepancies are difficult to detect in the complex environment of a court trial without the expert skills of an independent discourse analyst fluent in both languages involved. As a result, trials tend to proceed without the reliability of the translated transcript being adequately tested.

Keywords: translation, transcription, covert recordings, drug investigations, forensic linguistics, language policy, evidence, interpreting

OPEN ACCESS

Edited by:
Emma Richardson,
Aston University, United Kingdom

Reviewed by:
Chuanyou Yuan,
Guangdong University of Foreign Studies, China
Veronica Bonsignori,
Foro Italico University of Rome, Italy

****Correspondence:***
Georgina Heydon
georgina.heydon@rmit.edu.au

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *18 September 2021*
Accepted: *03 November 2021*
Published: *03 December 2021*

Citation:
Gilbert D and Heydon G (2021) Translated Transcripts From Covert Recordings Used for Evidence in Court: Issues of Reliability. Front. Commun. 6:779227. doi: 10.3389/fcomm.2021.779227

1 INTRODUCTION

Electronic surveillance technology is an effective means of collecting evidence used to prosecute serious and organised crime. Evidence presented in drug-related trials is often in the form of audio recordings of conversations held in LOTE. The recordings are usually obtained through telephone interception or covertly placed listening devices. The audio recordings are presented as primary evidence in the form of an audio file. To make sense of the evidence, the audio files are accompanied by transcripts in English having been translated from languages other than English (LOTE). These translated transcripts often contain drug-related code words and jargon.

Research conducted at RMIT University, Melbourne, Australia aimed to determine the reliability of translated transcripts presented as evidence in court in drug-related trials. The research focused on determining:

1) What evidence, if any, points to systemic deficiencies in language capability relied upon to combat illicit drug-related crime?
2) How do identified deficiencies affect the judicial process?
3) What causal factors contribute to these deficiencies?

2 CONTEXT OF THE RESEARCH

The context of this enquiry is framed within two key areas as follows: at the micro level involving linguistic analysis of translated transcripts from electronic surveillance related to serious and organised crime revealing evidence of language capability deficiency; and at the macro level where analysis findings reveal causal factors leading to the distortion of evidence in court trials.

Evidence of deficiencies at the micro level show that translated transcripts of intercepted telephone calls presented as evidence in court used to prosecute serious and organised crime contained significant errors, many of which were not detected by the court. At the macro level, the research revealed significant deficiencies in interpreter and translator training, workplace practices, and the process of skills recognition of professional interpreters and translators. Collectively, the data provide evidence of systemic deficiencies in language capability and, when viewed through the lens of criminal justice, the findings reveal significant and systemic distortions of evidence presented in criminal trials presenting a clear risk to the integrity of the judicial process.

3 LITERATURE REVIEW

Review of the literature reveals a gap in knowledge relating to the accuracy, or perceived accuracy, of translated transcripts used for evidentiary purposes, particularly relating to drug-related code words and jargon from languages other than English (LOTE) used as evidence in court. This is most likely due to the unique specialist skills and experience required to conduct this type of research. [Moreno (2004), 34] noted a lack of empirical research concerning the accuracy of alleged drug-related codewords presented as evidence in court, stating that "There is no indication in any related literature that there has ever been a real effort to study or test the reliability of any drug jargon definitions." A review of the literature at the time of writing reveals that the empirical research discussed in this paper is unique and fills the gap in knowledge Moreno had previously identified. Importantly, [Moreno (2004), 35] states that "[t]he problem with the lack of objective data is that it prevents judges from measuring the reliability of this evidence pretrial and, once admitted, prevents jurors from gauging its weight," adding that "In the context of drug jargon interpretation, judges and juries cannot measure the probability that expert testimony is reliable by comparison to a professional standard or empirical evidence." More recently, [Capus and Griebel (2021), 74] researched the visibility of translators responsible for producing translated transcripts, and state that research in this area is lacking. A review of the literature reveals that this research may be the first to contain objective empirical data that sheds light on deficiencies in translated transcripts that often remain undetected during drug-related trials.

3.1 Transcription: A Specialised Skill

Transcribing LOTE directly into written English is a specialised skill not normally practiced by community interpreters and translators. Highly developed listening skills are required of the translator or interpreter to capture important elements of evidentiary value when producing translated transcripts. National skills recognition of interpreters and translators is the responsibility of the National Accreditation Authority for Translators and Interpreters (NAATI). NAATI, a private business owned by the state and federal governments of Australia, conducts testing for interpreters and translators and issues certification for successful candidates. It also recertifies those interpreters and translators who successfully revalidate their skills. The transcription of spoken LOTE into written English is a specialised skill that is not tested nor certified by NAATI. Law enforcement agencies rely upon professional certification of interpreters and translators issued by NAATI as a minimum level of proficiency for producing translated transcripts for evidence in court. NAATI testing does not specifically address transcription skills and NAATI does not provide formal skills recognition for this form of specialised skill.

3.2 Transcription Approaches

Translators and interpreters who participated in the research claimed that they had not been given specific training on transcription methodology prior to being tasked with producing translated transcripts for evidentiary purposes. Court interpreters agreed at interview that producing translated transcripts is a specialised skill requiring high-order listening skills above those required for community interpreting. In Australia, it is common practice for the law enforcement translator or interpreter to transcribe the intercepted language other than English (LOTE) directly into written English for evidence purposes. Courts are not provided with a transcript of the intercepted spoken LOTE in the source language (the LOTE). Australian courts are provided with the audio recording of intercepted communications and a translated transcript in English. Therefore, the transcription process is not transparent to the court, the Prosecution or the Defence.

Australia has yet to establish nationally recognised guidelines to produce translated transcripts for court purposes. The research revealed that there was a high level of inter-dependance relied upon by interpreters and translators tasked with producing translated transcripts. Participants stated that they learn from each other in the absence of formal transcription training and skills recognition. The reported ad-hoc nature of acquiring transcription skills presents an unacceptable risk that systemic deficiencies in approaches to the transcription task will remain embedded within the law enforcement transcription environment.

The National Association for Judiciary Interpreters and Translators (NAJIT) in the US published a position paper providing "general guidelines and minimum requirements for transcript translation in any legal setting" (NAJIT, 2009). In the US, it is mandatory that transcription is conducted by transcribing the spoken LOTE into written LOTE by one person, then the written LOTE is translated into English by a certified translator. The Home Office in the United Kingdom has produced guidelines for the engagement of interpreters in criminal investigations where transcription is required. These guidelines are not made available to the public and carry a privacy marking of "Official Sensitive" [Home Office (2021), 13].

The US approach provides a clear audit trail of how the intercepted speech was transcribed and translated when presented as evidence in court. However, this is a process not normally practiced in Australia. Transcribing from spoken LOTE directly into written English, as practiced in Australia, is likely to result in evidence presented in the form of disjointed and mostly non-sensical English purported to have been said in the intercepted LOTE. The reason for using this method in Australia is assessed as being driven by financial resource constraints and the absence of policy guidance in relation to the methodology to be used when producing translated transcripts for evidentiary purposes.

Problems have been identified with evidence in the form of recordings, transcripts, and translations presented in US courts. Fishman (2006) states that juries may find recordings played in English difficult to understand as they often contain nuances, codewords, jargon and/or idioms. Written translated transcripts are often distributed to jurors as an aid to understanding the content of recordings. This is particularly the case where the translated transcript contains jargon, codewords or slang translated from a LOTE. Jurors cannot make use of evidence in the form of audio recordings of conversations held in LOTE unless they are assisted with an English translation. The level of subjectivity is significantly increased when a LOTE has been translated into English, as the translating and interpreting process is "much more an art than a science, let alone a mechanical process" [Fishman (2006), 476]. Laster and Taylor (1995) and Nakane (2009) share this viewpoint.

3.3 Transcription Accuracy

Transcribing covertly obtained recordings in LOTE that contain code words and/or jargon is complex. The process involves an approach requiring the translator to adopt translation strategies that seek to preserve notions of translation accuracy to preserve the integrity of the evidence. Translators are required to exercise critical decision making when producing translated transcripts for evidentiary purposes. Without a systematic approach to transcribing from LOTE into English, the resultant product is likely to contain errors bringing into question the key attribute of reliability and may be subject for unjustified interference by the translator. It is often the case that the translator struggles to transfer exact meaning into English due to distinct differences between languages. Exact meaning is elusive and the distance between an utterance in a LOTE and how it has been translated largely depends upon context. [Baker (2011), 60–61], states that:

> Accuracy is no doubt an important aim in translation, but it is also important to bear in mind that the use of common target-language patterns which are familiar to the target reader plays an important role in keeping the communication channels open.

In reference to the field of forensic translation, [Darwish (2012), 75], states that it is important that "evidentiary clues are not sacrificed for the sake of naturalness." The author concedes that it is inevitable that compromises will have to be made, although the preservation of meaning should be maintained being careful to avoid unjustifiable intervention or interference by the translator. A sound approach to the translation process will lower the risk of evidence being intentionally or inadvertently distorted. Specialised skills training and knowledge is required to produce covertly obtained translated transcripts. Darwish proposes that "in most situations" translated documentation presented as evidence is translated by those who have significant biases or are "simply incompetent." The author states that this adversely affects forensic analysis and may contribute to miscarriages of justice (2012, 19). The concerning issue of transcript translation not being adequately assessed for reliability is not peculiar to the Australian context. [NAJIT (2009), 6] states that "transcript translation remains an area that is not uniformly regulated in courts nationwide."

Translators working for law enforcement are required to transfer equivalent meaning at word and, where possible, sentence level as closely as possible while also conveying sense. NAJIT guidelines (2009, 6) state that translations should contain attributes of accuracy and completeness and, where appropriate, be natural and idiomatic while faithfully reflecting register, style, and tone of the original text. However, NAJIT has not provided a definition of accuracy. The idea of accuracy is an ambiguous concept in terms of the translation process and when faced with evaluating a translated text. First, the translator conducts an analysis of the original text and then interprets what the text means within the context it is placed. The translator is also required to consider the assumed meaning intended by the originator of the source utterance in LOTE. Once the translator has formed an impression of context, the process of transcribing the original text into English can begin. Therefore, it is important that the translator has access to information about context extrinsic of the original text as part of the analysis and decision-making process. Only then is the translator suitably equipped to transfer intended meaning from LOTE into English while preserving the evidentiary value of the original text. [NAJIT (2009), 6] notes that contextual information may assist the translator in "comprehending distorted sound" or clarifying "ambiguous utterances" but with an emphasis that any final translation should contain "only what he or she actually hears in the source recording."

Fraser et al. (2011) researched the potential influence on the hearer of recorded conversations from "priming" their senses by providing them with background information. The research revealed that it is likely that people will hear what they expect to hear based on extrinsic information provided prior to listening to the recording. It follows that law enforcement translators may also be influenced by background or intelligence information when transcribing intercepted communications. This creates a dilemma for the law enforcement translator where they either produce a translated transcript in a vacuum without information relating to context, or they have access to background information that may influence what they hear in the source recording. Whichever approach is taken to the transcription process, the law enforcement translator still needs to document what was heard and, as closely as possible, convey the communicative function of the intercepted utterances in a format acceptable as evidence. Ideally, the final product is a translated transcript in English that makes sense. Hence, the importance of transcribing the spoken LOTE into written LOTE prior to it being translated into English so that transparency of the transcription process is achieved. This way, any influence of priming or unjustified intervention by the translator can be more easily detected during a quality control process.

It is highly probable that the translator will interfere during the translation process. Translators should declare where they have interfered during the translation process to convey sense. [House (2009), 42] proposes that "a translated text can never be identical to its original, it can only be equivalent to it in certain aspects." This raises a dilemma when it comes to the quality control of translated transcripts. When assessing translations for accuracy, equivalence, and objectivity, one may arrive at alternative acceptable translated versions of the original LOTE text. Internal consistency, linguistic integrity and translation integrity are dependent upon the strategies applied by the translator. Attempting to maintain a balance between readability and accuracy is an inherent part of the transcription process (cf. Tilley, 2003).

Importance of contextual information to the translation process cannot be underestimated. Consensus in relation to translation accuracy is dependent upon a mutually agreed perspective of what is known and what is expected. In criminal trials, the Prosecution and the Defence are required to agree that the translated transcripts are accurate prior to a trial commencing. However, the notion of accuracy is often determined at word level. The contextual meaning of words contained in translated transcripts is determined by the jury through the adversarial process. Prior to hearing arguments put by the Prosecution and the Defence, the jury is provided with a copy of the translated transcripts in English. The jury will then hear what the Prosecution and Defence allege those utterances mean within the alleged context of the evidence presented. The jury, being the trier of fact, is charged with determining the accuracy and reliability of the evidence presented at trial.

3.4 Translated Transcripts and the Expert Witness

The practice of calling police officers as expert witnesses in relation to the translation of code words and jargon is a significant area of investigation in this research. Police officers often provide expert witness testimony to explain the meanings of terms and phrases contained in translated transcripts. This is to assist the jury to understand the alleged context in which the intercepted conversations in LOTE took place. Expert witness testimony in these circumstances is often delivered by monolingual police officers who further interpret the meaning of alleged drug-related code words contained in translated transcripts.

Police officers in the United States routinely testify on the modus operandi of drug traffickers and dealers and how drug jargon is to be translated. Moreno (2005) states that they are called upon to testify by the Prosecution on the basis that:

1) Illicit-drug offenders routinely use drug-related codewords and jargon.
2) Jurors are unlikely to understand drug-related terminology without expert assistance.
3) Police officers are proficient in the identification and translation of drug-related jargon

It has been shown that Judges are reluctant to question the expertise of police officers who testify as expert witnesses called to explain the meaning of drug-related code words and jargon (Moreno, 2005). In United States v. Boissoneault 926 F.2d 230, 23 (2d Cir. 1991) the court of appeal held that "experienced narcotics agents may explain the use and meaning of codes and jargon developed by drug dealers to camouflage their activities." However, jurors may become confused when hearing testimony proffered by a police officer who is both the police investigator and the expert witness. The confusion arises from the question of whether the testimony is based on the police officer's general experience or whether the testimony is drawn from the officer's role as investigator. Moreno (2005) asserts that the court will usually accept expert evidence proffered by police officers as credible and accurate.

Research has been conducted in relation to potential systemic biases in the judicial system, but few studies have been carried out on the reliability of translated transcripts (Nunn, 2010). Nunn's research revealed that transcripts are subject to distortion to add weight to the evidence in favour of the Prosecution in criminal trials and estimates that 81 per cent of "wiretaps" relate to targeting the illicit-drug trade. Importantly, Nunn (2010) found systemic police bias influenced the transcription process. A police officer with relevant experience, training and knowledge may give evidence as an expert witness in relation to drug-related code words as they appear in a translated transcript into English, however, this testimony is based upon the assumption that the translated transcript is accurate. This research reveals that trials commence without the accuracy of translated transcripts having been challenged due to resource constraints. This therefore increases the potential risk of accused persons not receiving a fair trial.

4 METHODS

Identifying potential or actual deficiencies in foreign language capability relied upon by law enforcement agencies requires access to reliable and credible sources of data that is not subject to publication restrictions due to the sensitive nature of law enforcement or national security related operations. The public court system provides an opportunity to observe and collect qualitative and quantitative data relating to serious and organised crime available in the public domain. The triangulation of four data collection methods ensured validity and reliability of the research findings. The first method involved observation of three drug-related trials held in the County Court of Victoria from 2012 to 2014. These trials provided direct access to audio recordings in Vietnamese and associated translated transcripts relied upon by law enforcement agencies used as evidence to prosecute persons accused of carrying out acts of serious and organised crime. The translated transcripts were produced by community interpreters and/or translators employed by law enforcement agencies, many of whom are contractors also working for one or multiple government and private agencies. Extracts from translated transcripts contained in this paper provide a detailed analysis of evidence revealing Australia's deficiencies in the forensic translation process. The second approach was to interview County Court judges, Prosecution and Defence barristers, Court Interpreters, and interpreters/translators who had experience in relation to producing translated transcripts presented as evidence in court. Third, transcripts from court proceedings were analysed. The fourth method involved quantifying data retrieved from the AUSTLII database.[1]

This article draws mainly on the first tier of data collection conducted during the observation of three criminal trials heard in the Victorian County Court between May 2012 and March 2014 where translated transcripts were used as evidence. The field researcher[2] recorded courtroom activity through extensive note taking to document details and events as electronic recording of trials is not permitted in the Victorian County Court.

Observation of the three trials enabled the development of a data collection strategy to answer the previously mentioned research questions. The field researcher is a professional Vietnamese translator with experience in transcribing intercepted communications for law enforcement and military purposes. Therefore, data collection efforts were focused on trials where translated transcripts from Vietnamese to English were presented as evidence in drug-related cases. The translated transcripts were of conversations held in Vietnamese that had been covertly recorded by telephone interception or listening device. The field researcher directly observed more than 100 h of trial proceedings across the three separate trials comprising the three case studies in addition to a further three trials on an opportunity basis. Participant-observation methods were not applied. The researcher did not attempt to influence the conduct of the three trials or the court environment during the observation period. The researcher listened to the covertly obtained telephone intercept and listening device recordings containing conversations in Vietnamese played to the court. Using detailed notes, the researcher then compared what he had heard and documented with the corresponding translated transcript in English which was read aloud to the court by an appointed court official.

The researcher documented examples of errors detected in the translated transcripts which are presented in the Results (cf. **Section 5**). The findings from Tier 1 established a platform from which to design other methods of data collection which were applied in Tiers 2, 3, and 4. As the findings from Tiers 2, 3, and 4 also contribute to the Discussion (cf. **Section 6**), a brief description of each method follows (cf. Gilbert, 2014).

Evidence of significant errors contained in translated transcripts was detected during observation of trials at Tier 1. Examples of discourse analysis conducted during Tier 1 Case Study 1 are provided (cf. **Section 5**). The data collection method used in Tier 2 was in the form of questionnaires and interviews. Key stakeholders provided valuable information concerning the preparation of translated transcripts. The sample populations engaged for data collection at this level included judicial officers of the Victorian County Court, barristers, court interpreters, community interpreters/translators who had previously been engaged by law enforcement agencies to conduct transcription tasks for evidentiary purposes. Participants with appropriate skills, knowledge and experience relating to the production and use of translated transcripts from LOTE were selected based on their ability to provide relevant information. A focused and targeted approach was necessary due to the small number of suitable participants who were able to provide information about the specialist areas of transcription for law enforcement, legal processes, intelligence, court interpreting and transcription for military purposes. Participants were issued with written information explaining how the information they provided would be analysed and presented. Closed, multi-choice questions were used in the questionnaires (cf. Gilbert, 2014, Appendices F to I). Questionnaires preceded a second level of data collection in the form of in-depth interviews. Participants were advised that they could withdraw from the process at any point if they wished to do so. The interviews and questionnaires were designed to collect information relating to 1) evidence of language capability deficiencies in the non-traditional security sector of law enforcement; and 2) how language capability relied upon in the military environment for transcription tasks compares with the principles and methods applied in the law enforcement working environment.

Tier 3 involved the collection of court transcripts and discourse analysis of the collected data from three criminal trials involving serious and organised crime specifically related to illicit-drug activity. Each trial was categorised as a separate case study. The Victorian Government Registration Service provided access to court transcripts that were used to triangulate the data collected in Tiers 1 and 2. The Australasian Legal Information Institute (AUSTLII at austlii.edu.au) provided information for Tier 4. Four appealed cases were analysed and a keyword search on "code words" was conducted. The four case reports recorded details of drug-related trials. Translated transcripts had been admitted as evidence in the four trials containing alleged drug-related code words. The cases selected were heard in the Victorian Supreme Court of Appeal and the New South Wales Criminal Court of Appeal. This data collection method and subsequent analysis revealed the approach the courts take to allowing or disallowing evidence proffered by expert witnesses relating to the content of translated transcripts. The four cases reflected contention in relation to the alleged meaning of drug-related code words.

A systematic method of triangulating the data was used to process the data provided by participants. Data saturation was

[1] The collection of data in this research was approved by the College Human Ethics Advisory Network, RMIT University Approval number CHEAN A 0000015703-09/13 dated 7th November 2013

[2] Dr David Gilbert (First author)

achieved to the point where no new categories were identified. Commenting on discourse analysis, [Wood and Kroger (2000), 81] emphasise the importance of having sufficient data to arrive at a reliable and well-grounded conclusion regardless of whether data saturation has been achieved. The authors find that when considering the data collected for discourse analysts "bigger is not necessarily better." Due to the specialised areas under investigation in this study, enough data were collected to establish evidence of deficiencies in language capability relied upon by law enforcement for evidentiary purposes, specifically in relation to alleged illicit-drug activity.

5 RESULTS

Significant distortions of meaning were detected in translated transcripts across three separate trials. Each trial represents a case study for the purposes of this research. Translated transcripts from intercepted telephone conversations and listening device recordings were proffered as evidence in the three Vietnamese drug-related trials. Court transcripts containing expert opinion evidence proffered by police officers concerning the alleged meaning of drug-related code words were also analysed.

Discourse analysis of the recorded Vietnamese conversations revealed that the word "thingy" had been incorrectly used in the English transcripts and was not used by the accused as a code word for drugs. Rather, the word "thingy" appeared in the English transcripts instead of using optimally appropriate anaphoric and exophoric reference words such as "it," "that" and "there." Further evidence confirmed that the word "thingy" was misused as a code word for drugs among Vietnamese interpreters and translators working for law enforcement agencies. During one of the case studies, a translator responsible for producing a translated transcript containing numerous references to the word "thingy" was called to give evidence in court. The translator giving evidence stated that interpreters and translators working on law enforcement drug-related operations routinely use the word "thingy" when they were unsure of what was being referred to in intercepted conversations. The use of "thingy" and other phenomena identified in the analysis of translated transcriptions from the case studies in this research are presented below.

5.1 Case Study 1

The trial was held in the County Court of Victoria. The accused person was being tried for allegedly having imported a commercial quantity of heroin and had been charged with drug-trafficking offences. Translated transcripts of conversations held in Vietnamese between the accused and other persons were presented as evidence. The translated transcripts were produced by a community translator under contract to a law enforcement agency. Police used methods of telephone interception and covertly placed listening devices to obtain the audio recordings. The brief of evidence presented by the Prosecution in this case also included other forms of evidence such as expert witness testimony, documents, witness statements, and various items. Vietnamese court interpreters assisted the court and interpreted for the accused when the accused gave evidence as a witness in his own defence.

The court played the intercepted audio recordings of conversations in Vietnamese aloud during the trial. This was necessary because the accused was giving evidence. It was therefore necessary that a translated transcript in English of the intercepted recordings in Vietnamese was read to the court so that the jury and court officials could understand what was allegedly contained in the recordings. The format established by the court for examination and cross-examination of the witness was implemented as follows:

- Counsel draws reference to an audio recording of utterances related to the line of questioning during examination or cross-examination of the accused.
- A court interpreter advises the accused that an audio recording is about to commence.
- The audio recording in Vietnamese is played to the court.
- The translated transcript in English is then read to the court by an independent court official.
- Counsel continues with the line of questioning with reference to the recorded conversations.
- The court interpreter interprets Counsels' questions from English to Vietnamese for the witness.
- The witness replies in Vietnamese.
- The court interpreter interprets the witness' response from Vietnamese into English for the court.

This method was implemented to enable all present in the court to understand the evidence and legal proceedings in English and Vietnamese.

Problems concerning the translated transcripts were observed on the first day of the trial. The field researcher compared the audio recordings of Vietnamese conversations played to the court with the translated transcript read to the court in English. Significant errors of distortion, omission and unjustified additions were identified in the translated transcripts that contained numerous serious English grammatical errors. In relation to "correctness" of translating evidentiary documents, [Darwish (2012), 66], states that "a grammatical mistake that disguises itself as another correct grammatical form may not be detected as such and may cause interference with the original intents of the message." It became apparent that this type of translator interference was evident in the translated transcripts presented at the trial.

A Victoria police officer gave evidence as an expert witness during the trial. The officer proffered expert opinion evidence explaining the meaning of alleged code words and jargon as they appeared in the translated transcripts. The researcher observed that poor lexical choices and misinterpretations contained in the translated transcript were further interpreted by the police officer for the court. The police officer gave evidence that the words "thingy," "gear" and other words as they appeared in some segments of the translated transcript were references to heroin. Words alleged to be either code words or jargon became the focus of the study. It was noted that the word "thingy" as it appeared with the context of the translated transcript does not have a direct lexical equivalent in Vietnamese. The Vietnamese word “ấy” had been frequently translated as “thingy.” The Vietnamese word “ấy” is an exophoric or anaphoric reference word meaning this, it, or that. In some contexts, it may also be used to refer to a third person. However, the word had been poorly translated as the English word “thingy.” This mistranslation resulted in large sections of the translated transcript containing non-sensical English.

Further errors and inconsistencies were identified in the translated transcripts. These errors adversely affected communication between the witness and court officials causing significant delays and confusion. The only persons who were aware of significant errors contained in the translated transcript were the court interpreters present in the courtroom and the field researcher. This is because they were proficient in both languages. It is assessed that most errors contained in the translated transcripts remained undetected by the judge, jury and the rest of the court. Some areas of the translated transcript were challenged by the Defence.

Observations of trial proceedings indicated that:

- significant errors appeared throughout the translated transcript
- a further level of interpretation of what was contained in the translated transcript was applied when the court interpreter remedied mistakes contained in extracts of the translated transcripts cited by counsel during examination of the witness
- court interpreters did not voluntarily draw the court’s attention to errors contained in the translated transcript
- legal argument transpired about the meaning of utterances contained in the translated transcripts.

The researcher recorded notes during observation of this trial for subsequent discourse analysis of selected utterances heard in Vietnamese when audio recordings were played to the court. The audio recordings in Vietnamese were transcribed into written Vietnamese and then translated into English. The examples below contain grammar analysis of the translated transcript extracts revealing significant errors of translator interference. The following five utterances are part of a conversation intercepted by a covertly placed listening device. The translated transcript was presented as evidence in court. The Prosecution alleged that the intercepted conversation was held between two persons in a room engaged in the act of dividing heroin for subsequent distribution. The audio recording has been transcribed from the intercepted Vietnamese speech and is labelled “Source text.” A word-for-word literal translation from Vietnamese to English is then provided. This is followed by the corresponding translated transcript that was read to the court so that the judge and jury may make sense of the intercepted conversation in Vietnamese. Finally, a proposed alternative translation produced by the field researcher in consultation with a professional Vietnamese court interpreter of 25 years’ experience is provided. A critical analysis of the selected utterances is then provided to help the reader understand where the distortion of meaning and/or omission occurs. It is noted that a transcript of the original audio recording in Vietnamese is not provided to the Court. Only the original audio recording and a translated transcript in English are made available to the Court as evidence.

The following data (utterances one through five) are reproduced from Gilbert, 2014 (cf. Gilbert 2017).

1) Utterance One

Source text	Đụ-mẹ. Tôi-không-biết-chia. Tôi-chia-ra-tôi-mất-thấy-mẹ. Chia-nó-chút-chút-lần-nào-cũng-mất.
Literal translation	Fuck-mother. I-not-know-divide. I-divide-out-I-lose-father-mother. Divide-it-little-little-time-each-also-lose.
Translated transcript	Mother fucker! I don’t know how to divide it. Divide it and I would lose damn it.
Proposed alternative translation	Mother fucker! I don’t know how to divide it. I lose (some) when I divide it, damn it. Each time I divide into small portions I lose (some).

There is an omission in the translated transcript. The final statement “Each time I divide into small portions I lose (some)” does not appear in the translated transcript. This is assessed to be a serious error as it adversely affects the element of textual cohesion when considered within context of the utterances that follow.

2) Utterance Two

Source text	Chia-là-mất,-chết.
Literal translation	Divide-is-lose,-dead.
Translated transcript	Lose it, God oh God, is it dead?
Proposed alternative translation	Dividing (it) means losing (some), damn it!

A statement and an idiomatic exclamation appear in the source text. The audio recording did not contain any question related to something or someone being dead as it appears in the translated transcript. The literal meaning of the Vietnamese idiomatic expression “chết” is “dead” in English. However, in the above context, the expression “chết” is used to denote frustration and can be optimally translated idiomatically as “damn it!” as shown above. During the trial, the Prosecutor asked a non-English speaking witness to clarify, through a court interpreter, what or who was “dead.” This resulted in significant confusion and delay during the trial. The issue was not satisfactorily resolved, and the line of questioning was dropped after the issue was eventually clarified by a Vietnamese court interpreter. The translated transcript also contains the expression “God oh God.” This is assessed to be an unjustified addition. The audio recording did not contain an idiomatic expression that justifies the insertion of “God oh God.” This is assessed to be an example of interpreter interference.

3) Utterance Three

Source text	Cái đó đó, có ấy chút xíu à, tại thằng kia lấy thử chút xíu.
Literal translation	Classifier-that-that,-have-it-little-(particle),-because-guy-that-take-try-little-bit.
Translated transcript	That one, only thingy a little bit, because the guy thingy, tested a little bit.
Proposed alternative translation	That one; it's smaller because that guy took a little bit to try it.

The extract of the translated transcript above contains the word "thingy." At this trial a police officer proffered expert opinion evidence that the word "thingy" was a reference to heroin. Appearance of the word "thingy" renders this segment of the translated transcript nonsensical. The choice to use the word is assessed as unjustified translator interference and renders the translation awkward, ambiguous and lacking coherence. It is assessed that the jury would increasingly rely upon the expert evidence provided by the police officer in this trial to understand this part of the transcript that contains the word "thingy." This has significant implications for the defendant due to the inherent bias of the word being described by the police officer as code word for drugs. Significant distortion is contained in this part of the translated transcript. It is assessed that the word "thingy" in this context is highly unlikely to be a reference to heroin. Rather, it is an example of translator interference resulting in a poor translation.

4) Utterance Four

Source text	Không-có-mấy-đâu,-xíu-xíu-à,-nó-cạo-chút-xíu-à.
Literal translation	Not-have-much-at-all,-little-little-(particle),-he/she-scrape-little-bit-(particle).
Translated transcript	No thingy, he scratched a little bit.
Proposed alternative translation	Not much at all, just a bit, he/she scraped a bit (off).

The word "thingy" appears again in the above extract. However, there is no Vietnamese word in the audio recording that be attributed to the word "thingy." The use of the word "thingy" as it appears in this segment is assessed to be a significant mistranslation.

5) Utterance Five

Source text	Nói-anh-vậy-đó,-mấy-cái-này-chắc-tôi-cân-dư.-Dư-chút-xíu. . .-mệt quá,-mẹ.
Literal translation	Speak-you-like-that,-few-these-probably-I-weight-excess.-Excess-little-bit. . .-tired-too,-mum.
Translated transcript	To tell you that bro, these I weighted and they may have been weighted with extra. A little bit extra but (mumbles) I was so tired, damn it.
Proposed alternative translation	Well, having said that, perhaps I'll add extra to the weight of these ones. Just a little extra. . .God, I'm so tired!

The translated transcript contains an incorrect translation of the phrase "Nói anh vậy đó" which has been translated into English as "To tell you that bro." The way this has been translated causes a break in textual cohesion with the previous utterance. An alternative and arguably more appropriate translation is "Well, having said that . . ." as shown in the proposed alternative translation above. Unjustifiable intervention by the translator is evidenced by using the word "bro" which is assessed to have originated from the translator having had access to extra-linguistic knowledge of the assumed context (in this case drug-related activity). This appears to have influenced the translator's choice of register. The word "you" instead of the word "bro" is assessed to be more appropriate in this context noting its evidentiary value.

5.2 Additional Data Collection

In addition to the trial discussed in Case Study 1, two further drug-related trials were also observed in the County Court of Victoria. Telephone intercept and listening device recordings in the Vietnamese language formed part of the brief of evidence in both trials. The alleged accuracy of the contents of the telephone intercept transcript was challenged by the Defence. The translated transcripts were not read aloud to the court in these trials. Important to this research, the law enforcement translator who had transcribed the recorded conversations from Vietnamese into English was called to give evidence as an expert witness. The translator was questioned by counsel in relation to the alleged accuracy of the translated transcripts. The translator giving evidence admitted to making several errors contained in the translated transcripts of which were subsequently amended as appropriate. Notably, the person who gave evidence of having transcribed the audio recordings gave evidence that the person did not hold professional qualifications as a translator but held qualifications as a professional interpreter.

Errors contained in the translated transcripts resulted in significant delays. References to Vietnamese names throughout the trial caused confusion for the jury. The word "thingy" was frequently heard when extracts of the translated transcripts were referred to by counsel. In both trials the Prosecution alleged that the English word "thingy" as it appeared in the translated transcripts meant drugs.

In addition to qualitative interviews and quantitative analysis of court transcripts, the data collection strategy included a keyword search of "thingy" at the Australasian Legal Information Institute (AUSTLII) website. The database returned a range of trials where twenty-five references to the word "thingy" had been identified. Three references were associated with cases outside Australia. A breakdown of the types of cases where twenty-two references to "thingy" appears is as follows: sex offences (14), theft (2), drugs (3), and other (3). Two of the three drug-related cases contained references to the word "thingy" drawn from translated transcripts. The likelihood that the use of the word "thingy" by law enforcement translators is cross-jurisdictional was established. One of the drug-related cases was heard in the New South Wales Court of Criminal Appeal in 2010 and the other in the Victorian Supreme Court of

Appeal 2011. In both trials the word "thingy" was contained in translated transcripts from Vietnamese derived from electronic surveillance. The results of the keyword search reveal that the word "thingy" forms part of a genre of language unique to the specialist field of producing Vietnamese drug-related translated transcripts.

5.3 Summary

Extracts from the translated transcript in Case Study 1 contain significant errors of translation due to unjustifiable translator intervention and poor word choices. The utterances lack coherence across the five samples as demonstrated when compared with the proposed alternative translations. An inconsistent approach seems to have been adopted by the translator. Forensic translation requires the translator to apply a consistent approach to ensure logical coherence at all levels of text. The sampled translations demonstrate a failure of cohesion at lexical, sentence and text levels. Nevertheless, prior to commencement of the trial at Case Study 1, the Prosecution and Defence counsels agreed that the translated transcript containing the above extracts was accurate.

6 DISCUSSION

6.1 Implications of Language Deficiencies for the Judicial System

Significant errors contained in translated transcripts are compounded when monolingual police officers provide expert witness testimony concerning the meaning of alleged drug-related terms appearing in translated transcripts from LOTE. Evidence from a police officer to the effect that the word "thingy" is a reference to drugs increases the risk of the accused not receiving a fair trial.

The data samples provide evidence that the information relied upon by the jury is confusing and cannot stand alone without further interpretation being applied by another source of information. As the translated transcripts are assessed to be potentially misleading and confusing, it can be argued that the evidence might have been excluded in accordance with Sections 135 and 137 of the Evidence Act 2008 (Vic) ("the Act") had the translated transcripts been properly assessed for reliability in terms of accuracy prior to commencement of the trial. At Section 137 of the Act, it is stated that "in a criminal proceeding, the court must refuse to admit evidence adduced by the prosecutor if its probative value is outweighed by the danger of unfair prejudice to the accused." The court relies heavily upon the Defence and the Prosecution agreeing to the translated transcripts being accurate when balancing the "probative value" of the evidence against the "danger" of unfair prejudice.

Further evidence was obtained revealing that significant errors contained in translated transcripts are likely to go undetected at trial. A Defence barrister commented during interview that nobody really knows whether the translated transcripts are accurate or not. Translated transcripts are rarely assessed to determine accuracy due to either the unavailability of funding from Legal Aid or a failure to have them evaluated by an independent certified translator. During one of the case studies the judge described the difference between translation accuracy at word level and the meaning of utterances contained in the translated transcript. The rationale behind this reasoning provided an avenue by which the translated transcript could be admitted as "accurate" but only so far as it is accuracy applies to words on a page. The judge explained that, while the words on their own may be accurate translations, it is for the jury in its capacity as the trier of fact to decide what the words in the translated transcript mean within an alleged context. There is usually only one version of the translated transcript presented to the court to assist the jury. No alternative versions are considered other than those interpretations arising from legal argument over the content of the translated transcript. The important attribute of context within which a conversation takes place is critical in the translator's decision-making process when deriving sense from intercepted utterances. Translators and interpreters with experience in producing translated transcripts stated during interviews that background and intelligence information about the covertly recorded conversations was not made available to them to assist with making sense of the intercepted utterances. They stated that this information was withheld from them for reasons of impartiality and to preserve the integrity of the evidence. Therefore, the translator producing the transcript applies a further level of interpretation when producing the transcript based on their personal knowledge, experience and assumptions of context. [Viaggio (1991), 37] emphasises that "[t]ranslation, as any other kind of communication, still succeeds as long as sense is conveyed, while it fails completely and inescapably if it is not." It follows that the originator's intended sense of the intercepted utterances is subject to distortion through the translation process when the translated transcripts are prepared for court. Further interpretation of what is contained in the translated transcript is applied by Counsel during the trial.

The nonsensical extracts from a translated transcript that form examples provided in this paper reveal what happens when sense is not adequately conveyed. The outcome is simply words on a page requiring further interpretation for the jury to understand what those words mean. The word "thingy" is a case in point. The data shows that inappropriate use of the word "thingy" is indicative that systemic mistranslations occur in translated transcripts, and they may remain undetected during court proceedings. This opens the door for expert opinion evidence proffered by police officers to interpret such terms for the jury in a realm of significant uncertainty. It is possible, if not probable, that the probative value of the translated transcripts would have been outweighed by the risk of prejudicial effect on the accused had the translated transcript been adequately evaluated for reliability prior to the commencement of trial. It follows that the probative value of the expert opinion evidence in this case may have also been significantly reduced had the significant errors contained in the translated transcripts been identified prior to the trial commencing. Judicial officers and barristers commented at interview that there is a tendency to expect that translated transcripts presented at trial are accurate. Interviews with Vietnamese court interpreters revealed that significant errors of translation are commonplace in translated transcripts from the Vietnamese language. They stated that they avoid alerting the court to errors contained in translated transcripts citing their ethical obligation to remain impartial forbids them from doing so. Interviews were held with Vietnamese translators and interpreters who had experience in producing translated transcripts for evidentiary purposes. They revealed that the word "thingy" had been misused in translated transcripts. They also commented that the word first appeared in Vietnamese drug-related cases in NSW and Victorian courts at least 14 years prior to the time of interview and reaffirmed that it is a term that appears to be peculiar to Vietnamese drug-related cases.

While collecting data during the observation phase of the research, a Vietnamese interpreter was subpoenaed to assist the court with disputed aspects of a translated transcript. Under cross-examination, the interpreter was asked to explain when the word "thingy" was used in translated transcripts. The interpreter stated:

> Sometimes we have different Vietnamese words we use, but basically the appropriate way is when we don't know for sure what that object is or are and when they use that word and we don't know for sure, then I put the word "thingy," because sometimes they will say, "ấy"—they just use the word "that one" or "cái." It could mean anything so I just put the word "thingy" meaning that we are not so sure of what they are talking about.

The Prosecution had alleged that "thingy" was a code word for heroin. Translated transcripts across three Vietnamese drug-related trials contained numerous occurrences where the word appeared seemingly out of context. It was established that the word "thingy" is a cross-jurisdictional phenomenon frequently occurring in Vietnamese drug-related translated transcripts in NSW and Victorian criminal cases. A search of the AUSTLII database at the time of writing reveals that the word "thingy" appeared in a translated transcript presented as evidence at a Vietnamese drug-related trial in the County Court of Victoria in May 2017 in DPP v Agbayani (2017) VCC 723 (June 8, 2017). Again, the word appeared out of context but was not referred to in the court transcript as a code word for drugs.

The problematic misuse of the word "thingy" has been identified in another language. A Chinese interpreter with experience in producing drug-related translated transcripts who participated in the research stated that the word "thingy" was used in a drug-related translated transcript from Chinese. The interpreter explained that use of the word "thingy" came from advice provided by a Vietnamese interpreter who was a colleague of interviewee and was also working for the same law enforcement agency. It is evident that a genre of discourse specific to the specialist area of producing translated transcripts has been in existence for several years and that not only is it cross-jurisdictional, but it has also been used in translated transcripts from at least one other language than Vietnamese. The research has established that the problem of nonsensical English appearing in translated transcripts arises from the translator attempting to preserve the integrity of the evidence by applying accuracy at word level at the sacrifice of conveying sense.

6.2 The Police Expert Witness

Investigating police officers often proffer expert opinion evidence in relation to the alleged meaning of drug-related jargon and code words. It has been established that drug traffickers' jargon is a specialised body of knowledge allowing police officers to give evidence as experts to explain drug-related terminology. In United States v Boissoneault, the court of appeal held that "experienced narcotics agents may explain the use and meaning of codes and jargon developed by drug dealers to camouflage their activities." Police officers are rarely challenged in relation to the reliability of their expert opinion evidence as the aspect of determining reliability rests with the trier of fact. In Australia, Section 79(1) of the Uniform Evidence Act requires that expert opinion evidence is proffered by a person who has "specialised knowledge"; that the specialised knowledge is based on the person's training, study or experience; and the opinion is "wholly or substantially" based on that specialised knowledge.

The research findings reveal a significant bias towards the Prosecution case as a result of inadequate translation quality control procedures. The High Court considered the issue of expert evidence in Dasreef Pty Ltd. v Hawchar with Heydon J[3] stating:

> Opinion evidence is a bridge between data in the form of primary evidence and a conclusion which cannot be reached without the application of expertise. The bridge cannot stand if the primary evidence end of it does not exist. The expert opinion is then only a misleading jumble, uselessly cluttering up the evidentiary scene.

The dangers of experts proffering their opinion without proper scrutiny of the primary data was discussed in In HG v The Queen (1999) HCA 2; (1999) 197 CLR 414 at [44] Gleeson CJ said:

> Experts who venture "opinions," (sometimes merely their own inference of fact), outside their field of specialised knowledge may invest those opinions with a spurious appearance of authority, and legitimate processes of fact-finding may be subverted.

When determining relevance of expert opinion evidence, it is argued that the monolingual police officer should be required to establish the reliability of their opinion when providing interpretations of terms appearing in translated transcripts alleged to be code words for drugs. On a technical point, the primary evidence comprises the sounds recorded on the audio file. Translated transcripts from LOTE derived from the audio files are termed secondary evidence and are presented to the jury

[3] No relation to the second-named author

with an appropriate direction delivered by the judge. Therefore, the reliability of expert opinion testimony is inextricably linked to the accuracy of the translated transcripts. This raises the prospect that words contained in translated transcripts may mean something other than what a police officer as an expert witness purports them to say. It follows that the notion of factual assumptions drawn from translated transcripts can be challenged on the grounds of reliability of any opinion expressed in relation to sense or intended meaning.

The reliability of expert opinion evidence proffered by police officers in relation to drug-related code words translated from a LOTE has been challenged in appeals cases. In the case of Pham, Van Diep; Tran John Xanvi v R the New South Wales Court of Criminal Appeal considered grounds of appeal relating to the conviction of the appellants found guilty of supplying prohibited drugs including heroin, cocaine and ice (crystalline methamphetamine). The first ground of appeal was that the trial judge erred in allowing a NSW police officer to proffer expert evidence.

At trial, the police officer testified to the meaning to alleged code words contained in translated transcripts of recorded conversations from intercepted telephone conversations in Vietnamese. There was no explicit reference to drugs made in any of the translated transcripts. The Court of Criminal Appeal reported that "[t]he Crown's case was that when one appreciated the code was present one could interpret the conversations as ones relevant to the dealing in drugs in question."

The police officer proffered evidence that the word "cabinet" is commonly used as a drug-related term referring to the prohibited drug ice. The officer relied upon his experience and a number of reference sources. The officer stated that his opinion was based on "a translation of a Vietnamese word which literally [led him] to believe the word cabinet is another word for fridge." The police officer referring to his notes explained that "[t]he word "fridge" in Vietnamese is in my knowledge is made up of two words being To and Lun, now I don't profess to have the tone marks or the pronunciation correct in those words."

During cross-examination, the police officer stated that the information upon which he has provided an opinion is consistent with drug-related terminology relating to the drug ice or crystalline methamphetamine. The officer also gave evidence that the word "to" in isolation is consistent with references to the drug ice or crystalline methamphetamine and added that the words "to" and "to lun" are interchangeable. During cross-examination, the officer also stated that the words "old man" refer to heroin. He stated that his opinion was based on previous calls he had seen.

The police officer informed the court that he was unable to properly write or pronounce Vietnamese words. However, the officer's expert opinion evidence was allowed and he provided expert opinion evidence on the meaning of individual Vietnamese words and their meanings when combined with other lexical units. From the information made available in the court report, the police officer relied heavily upon the discretion of the translator when making critical choices during the translation process. At trial, the police officer cited the word "to" (properly written as tủ) as being interchangeable with "to lun" (properly written as tủ lạnh). He stated that both Vietnamese terms are consistent with reference to the drug ice.

The word tủ forms part of many other words in Vietnamese relating to any box-shaped container. For example, a wardrobe in Vietnamese is a tủ aó and a safe is a tủ sắt. The term "cabinet" in English may well be drug jargon used to refer to the drug ice. However, what appears in the English transcript will depend on what lexical choices are made by the translator when translating the word tủ into English. The words "cabinet," "container," "box" or "trunk" are all acceptable translations. In Vietnamese, the words "fridge" and "refrigerator" are written and spoken the same way in the Vietnamese language. While the Vietnamese compound word tủ lạnh may be translated into English as either "fridge" or "refrigerator," it is not possible to abbreviate the Vietnamese word so that one or the other component of the disyllabic word only means "fridge" instead of "refrigerator." The Vietnamese word đồ is a further example of a Vietnamese word that is often skewed during translation to the advantage of the Prosecution. The word is often translated as "gear" in drug-related trials when optimally it means "stuff" or "things." Police officers giving expert opinion evidence have referred to the word "gear" as being consistent with drug-related terminology. The weight of the evidence is arguably diminished if the translator translates the Vietnamese word đồ as "stuff" for inclusion in the translated transcript. The word "stuff" instead of "gear" would not provide a monolingual police officer the leverage the officer requires to inform the jury that "stuff" is consistent with drug-related terminology, whereas the word "gear" is most likely to go unchallenged. It has been clearly demonstrated that expert opinion evidence proffered by police officers hinges upon the choices the translator makes when producing the translated transcript. The key question is whether the monolingual police officer as expert witness is "wholly or substantially" basing their expert opinion on specialised knowledge, training, and experience or whether the expert opinion is a further interpretation based on their understanding of lexical choices made by a translator who produced the translated transcript.

In Nguyen v R the NSW Criminal Court of Appeal considered expert opinion evidence proffered by a NSW police officer who was a native speaker of Vietnamese. The officer gave his opinion relating to the meaning of drug-related code words. The police officer was reported to have had extensive experience listening to recordings of conversations about the supply of prohibited drugs and "had become extremely familiar with drug related terminology, drug related prices and the methods of operation of drug-dealers." It was held that the police officer "could give evidence to the meaning of words and expressions recognised as argot of the drug trade." However, the trial judge also stated that

> . . . it is impermissible to give evidence of what a person means when he uses certain words and phrases, that is a

> witness cannot give evidence of what is in the mind of the person who is speaking or speculate as to what he is meaning.

The judge's statement is interesting when considering that the translator who produces the translated transcript does so without knowing all there is to know about the context within which the intercepted conversation is taking place. The translator must speculate at some point in relation to the meaning of words and phrases he or she hears and the sense that the originator of the utterances intends to convey. The translator is compelled to speculate because he/she is not a party to the conversation but simply a witness to it. The translator produces a translated transcript in a context-deficient environment and therefore will need to speculate. Available extra-linguistic information such as intelligence support is withheld from the translator to maintain the integrity of the evidence relied upon by the Prosecution. Should the translator be provided with all background and intelligence information available, the Defence may argue that the law enforcement translator was primed through the provision of extrinsic information relating to drug-related activity under investigation. Translators and interpreters interviewed during the research cited this as a reason why translators apply a literal approach to translation when producing translated transcripts. As shown in this article, this results in awkward sentence structures and significant distortions of meaning. [Viaggio (1991), 32] clearly summarises the translator's dilemma when producing translated transcripts while attempting to preserve evidentiary value:

> Every single utterance can have countless senses. Sense is, basically, the result of the interaction between the semantic meaning of the utterance and the communication situation, which in turn is its only actualiser. Out of situation, and even within a linguistic context, any word, any clause, any sentence, any paragraph, and any speech may have a myriad of possible senses; in the specific situation—only one (which can include deliberate ambiguity). The translator ideally has to know all the relevant features of the situation unequivocally to make out sense.

The translator's dilemma described above is arguably inescapable and can only be resolved through an agreed consistent approach to the task of producing translated transcripts. There will always remain reasonable doubt in relation to the accuracy of translated transcripts while context underpinning intercepted conversations is not made known to the translator responsible for producing them. Context is an integral part of translation and is based on assumptions of meaning. [Gutt (1998), 46] states that successful communication is predicated upon values of consistency and is context dependent. This is because the author of the source text has intentionally crafted the communication produced in a format that is optimally relevant to the intended audience. Without access to all available contextual information surrounding the intercepted utterances, the translator can only assume the contextual framework within which the conversation takes place between the author and intended recipient. The translator will therefore inevitably intervene during the translation process bringing their own understanding of reality to the translated transcript. The lexical choices made by translators when producing translated transcripts are likely to have a significant impact on the outcome of court decisions noting that a further layer of interpretation is usually provided by police officers as expert witnesses.

6.3 Causal Factors of Mistranslation

Analysis of courtroom interactions and court records indicates that errors in the translated transcripts of recorded conversations have the potential to undermine the integrity of evidence in drug-related cases. Causal factors attributed to these errors include the absence of a nationally recognised standard providing guidance for the production of translated transcripts for evidentiary purposes, inadequate interpreter/translator specialised training in producing translated transcripts, workplace influences, and inadequate quality control measures used to check translated transcripts for correctness and reliability prior to being presented as evidence in court. Systemic misuse of the word "thingy" by law enforcement translators of Vietnamese conversations is evidence of deficiencies in appropriate specialised training and skills recognition. Restricted access to essential background information providing context to the translator is also a contributing causal factor. Translators and interpreters referred to the transcriber's dilemma as being one where they are required to produce an "accurate" translation in the absence of extra-linguistic information to assist them when determining context. This explains why translated transcripts often do not make sense as evidenced by the data used in this paper (cf. **Section 5**).

7 CONCLUSION

This paper provided evidence that translated transcripts presented in court frequently contain significant errors that distort evidence used to prosecute serious and organised crime. Key causal factors that adversely affect the reliability of translated transcripts were identified as deficiencies in areas of specialised training, skills recognition and workplace practices. The reliability of evidence supported by translated transcripts may be further diminished through expert witness testimony provided by police. Effective national policy-making is required to establish appropriate forensic translation training and skills recognition to meet national security objectives and to provide for a fair judicial system. The implications of deficiencies in Australia's forensic translation capability increases the risk of innocent people being convicted and the guilty set free. It is a timely call for national policy-making concerning forensic translation.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by College Human Ethics Advisory Network, RMIT University. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

DG devised the project and carried out the data collection and analysis. GH took responsibility for the ethical integrity of the project. DG took the lead in writing the manuscript and GH provided critical feedback and helped shape the research, analysis and manuscript.

ACKNOWLEDGMENTS

The authors would like to acknowledge the School of Graduate Research, RMIT University and the many research participants for their significant contributions that made this research possible.

REFERENCES

Agbayani, D. P. P. v. (2017). *DVCC 723*. Available at http://www.austlii.edu.au/cgi-bin/viewdoc/au/cases/vic/VCC/2017/723.html?context=1;query=agbayani%20;mask_path= (Accessed on: November 24, 2021).

Baker, M. (2011). *In Other Words: A Course Book on Translation*. New York: Routledge.

Capus, N., and Griebel, C. (2021). The (In-)Visibility of Interpreters in Legal Wiretapping – A Case Study: How the Swiss Federal Court Clears or Thickens the Fog. *Int. J. Lang. L.* 10, 73–98. doi:10.14762/jll.2021.73

Darwish, A. (2012). *Forensic Translation: An Introduction to Forensic Translation Analysis*. Paterson Lakes, Victoria: Writescope.

Fishman, C. S. (2006). Recordings, Transcripts, and Translations as Evidence. *Wash. L. Rev.* 81, 473–523.

Fraser, H., Stevenson, B., and Marks, T. (2011). Interpretation of a Crisis Call: Persistence of a Primed Perception of a Disputed Utterance. *Ijsll* 18 (2), 261–292. doi:10.1558/ijsll.v18i2.261

Gilbert, D. (2017). "Electronic Surveillance and Systemic Deficiencies in Language Capability: Implications for Australia's Courts and National Security," in *Proof in Modern Litigation: Evidence Law and Forensic Science Perspectives*. Editors D. Caruso and Z. Wang (Adelaide: Barr Smith Press), 127–128.

Gilbert, D. (2014). *Electronic Surveillance and Systemic Deficiencies in Language Capability: Implications for Australia's National Security*. [dissertation]. [Melbourne, Vic]: RMIT University.

Gutt, E. (1998). "Pragmatic Aspects of Translation: Some Relevance-Theory," in *The Pragmatics of Translation*. Editor L. Hickey (Great Britain: Cromwell Press Ltd).

Home Office (2021). Criminal Investigations: Use of Interpreters. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1007407/criminal-investigations-use-of-interpreters-v6.0-ext.pdf (Accessed September 16, 2021).

House, J. (2009). *Translation*. New York: Oxford University Press.

Laster, K., and Taylor, V. L. (1995). The Compromised "Conduit": Conflicting Perceptions of Legal Interpreters. *Criminology Aust.* 6, 9–14.

Moreno, J. A. (2005). Strategies for Challenging Police Drug Jargon Testimony. *Criminal Justice* 20, 28–37.

Moreno, J. A. (2004). What Happens when Dirty Harry Becomes an (Expert) Witness for the Prosecution? *Tulane L. Rev.* 79, 1–55.

NAJIT (2009). NAJIT Position Paper: General Guidelines and Minimum Requirements for Transcript Translation in Any Legal Setting, Atlanta, GA: National Association of Judiciary Interpreters & Translators. Available at: https://najit.org/wp-content/uploads/2016/09/Guidelines-and-Requirements-for-Transcription-Translation.pdf (Accessed September 18, 2021).

Nakane, I. (2009). The Myth of an "Invisible Mediator": An Australian Case Study of English–Japanese Police Interpreting. *Portal J. Multidisciplinary Int. Stud.* 6 (1), 1–16. doi:10.5130/portal.v6i1.825

Nunn, S. (2010). "Wanna Still Nine-Hard?": Exploring Mechanisms of Bias in the Translation and Interpretation of Wiretap Conversations. *Surveill. Soc.* 8, 28–42. doi:10.24908/ss.v8i1.3472

Queen, H. G. v. (1999). *HCA 2; (1999) 197 CLR 414*. Available at https://jade.io/article/68118 (Accessed on: November 24, 2021).

Tilley, S. A. (2003). "Challenging" Research Practices: Turning a Critical Lens on the Work of Transcription. *Qual. Inq.* 9, 750–773. doi:10.1177/1077800403255296

Viaggio, S. (1991). Contesting Peter Newmark. *Rivista Internazionale di Tecnica della Traduzione* 0, 27–58.

Wood, L. A., and Kroger, R. O. (2000). *Doing Discourse Analysis: Methods for Studying Action in Talk and Text*. Thousand Oaks, California: Sage Publications.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Communication

ORIGINAL RESEARCH
published: 22 December 2021
doi: 10.3389/fcomm.2021.797448

Specifying Challenges in Transcribing Covert Recordings: Implications for Forensic Transcription

Robbie Love[1]* *and David Wright*[2]

[1]Department of English, Languages and Applied Linguistics, School of Social Sciences and Humanities, College of Business and Social Science, Aston University, Birmingham, United Kingdom, [2]Department of English, Linguistics and Philosophy, School of Arts and Humanities, Nottingham Trent University, Nottingham, United Kingdom

Covert audio recordings feature in the criminal justice system in a variety of guises, either on their own or accompanied by video. If legally obtained, such recordings can provide important forensic evidence. However, the quality of these potentially valuable evidential recordings is often very poor and their content indistinct, to the extent that a jury requires an accompanying transcript. At present, in many international jurisdictions, these transcriptions are produced by investigating police officers involved in the case, but transcription is a highly complex, meticulous and onerous task, and police officers are untrained and have a vested interest in the influence of the transcript on a case, which gives rise to potential inaccuracy. This paper reports the design and results of a controlled transcription experiment in which eight linguistically trained professional transcribers produced transcripts for an audio recording of a conversation between five adults in a busy restaurant. In the context of covert recordings, this recording shares many of the typical features of covert forensic recordings, including the presence of multiple speakers, background noise and use of non-specialist recording equipment. We present a detailed qualitative and quantitative comparison of the transcripts, identifying areas of agreement and disagreement in (a) speaker attribution and (b) the representation of the linguistic content. We find that disagreement between the transcriptions is frequent and various in nature; the most common causes are identified as (i) omission of speech that is included in other transcripts, (ii) variation in the representation of turns, (iii) orthographic variation seemingly motivated by phonetic similarity, and (iv) orthographic variation seemingly not motivated by phonetic similarity. We argue that the variable nature of the transcription of "challenging" audio recordings must be considered in forensic contexts and make recommendations for improving practice in the production of forensic transcriptions.

Keywords: forensic transcription, covert recordings, speaker attribution, transcription variation, inter-rater agreement analysis

OPEN ACCESS

Edited by:
Helen Fraser,
The University of Melbourne, Australia

Reviewed by:
Vincent Hughes,
University of York, United Kingdom
Georgina Brown,
Lancaster University, United Kingdom

****Correspondence:***
Robbie Love
r.love@aston.ac.uk

Specialty section:
This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Communication

Received: *18 October 2021*
Accepted: *30 November 2021*
Published: *22 December 2021*

Citation:
Love R and Wright D (2021) Specifying Challenges in Transcribing Covert Recordings: Implications for Forensic Transcription.
Front. Commun. 6:797448.
doi: 10.3389/fcomm.2021.797448

1 INTRODUCTION

Covert audio recordings feature in the criminal justice system in a variety of guises, either on their own or accompanied by video. This can include clandestine 'undercover' recordings made by police, serendipitous recordings captured incidentally and recordings made by victims or witnesses on their mobile devices. If legally obtained, such recordings can provide important forensic evidence. They can capture a criminal offence being committed or can contain incriminating (or exculpating) material, including admissions of guilt, involvement, or knowledge of criminal activity. In other words, they can help in determining if a crime has been committed, what that crime is and who might be responsible. However, the quality of these potentially valuable evidential recordings is often very poor and their content indistinct, to the extent that a jury needs an accompanying transcript to assist in two tasks (i) working out what is being said (e.g. in cases of disputed utterances), and (ii) in multi-speaker recordings, working out who is saying what (cf. Fraser 2021a: 416).

At present, in many international jurisdictions, these transcriptions are produced by investigating police officers involved in the case "who are given the status of "ad hoc experts" to facilitate admission of their transcripts as opinion evidence" (French and Fraser 2018: 298). As is now well-documented, most comprehensively in the work of Fraser (e.g., Fraser, 2018a; Fraser, 2018b), current practice is problematic and risks producing unreliable evidence that can mislead the jury and result in miscarriages of justice. Transcription is a highly complex, meticulous and onerous task (Jenks 2013: 259). In a forensic context, although trained linguists and phoneticians can be involved in the production of transcripts, it is often the case that the police are responsible for producing transcripts for potentially incriminating audio, and this gives rise to some important problems (see Fraser 2021b for a nuanced discussion of the relative roles of experts and police in transcription). Police officers are untrained and have a vested interest in the influence of the transcript on a case. At best, this renders their transcripts as liable to being inaccurate. At worst, the effects of cognitive bias are such that they may "perceive something they expect, assume or want to be present" (Fraser 2014: 11).

Fraser (2021a: 428) provides an overview of the challenges facing forensic transcription and offers a solution to these problems:

> [T]hat all audio admitted as evidence in criminal trials is accompanied by a demonstrably reliable transcript that sets out the content, provides translations where necessary and attributes utterances reliably to participants in the conversation.

The first step towards achieving this, according to Fraser (2021a: 429), is to ensure that appropriately trained experts in linguistic science create and evaluate forensic transcripts rather than the police. In turn, this requires a branch of linguistic science dedicated specifically to the study of transcription (Fraser 2021a: 429). The current study shares this belief and aims to make a contribution in this direction. The position taken in this paper is that any science of transcription must be committed to observing transcription in practice; describing and explaining the processes and products of transcription; and predicting factors that influence and affect transcription and transcribers. To that end, the analysis conducted in this paper reports on a controlled transcription experiment comparing the transcripts of the same speech recording produced by eight different professional transcribers. It proposes different approaches to comparing transcripts in terms of their similarity and difference and applies these approaches to provide empirical evidence of the extent of variation across transcripts and a categorisation of different sources of this variation. The results of the experiment and the findings of the analysis can be used by forensic transcribers in reflections on their professional practice, to identify any key areas of focus in transcription and provide a basis for future transcription research. The direction of this study is guided by two research questions:

1. To what extent are the eight transcripts different from one another and what are the main sources of variation?
2. What implications do the results have for the practice of forensic transcription?

Prior to the analysis there is a review of relevant literature from linguistics and forensic linguistics, before a necessarily detailed description and justification of the methodological decisions taken. The paper ends with a discussion of findings and implications and a look forward towards future research in the scientific study of transcription.

2 LITERATURE REVIEW

2.1 The Process of Transcription

Linguistic transcription can be characterised simply as the "transfer from speech to writing" (Kirk and Andersen 2016: 291). It is a common procedure in many approaches to linguistic research as well as a range of professional contexts outside of academia, including forensics. Its ubiquity as a method for preparing data in linguistic research has given rise to the identification of a range of challenges that researchers have been contemplating for several decades (see Davidson, 2009, for a review of early transcription literature). For instance, it has been posited that transcription is not an objective process but rather a subjective and selective one: "because it is impossible to record all features of talk and interaction from recordings, all transcripts are selective in one way or another" (Davidson 2009: 38). As such, while some consider transcription as the process of producing "data" (for analysis), others consider transcription to be the first step of analysis in and of itself (Tessier 2012: 447).

The inherent subjectivity and interpretivism of transcription allows for both macro and micro variations among transcribers in terms of the representation of spoken language in written form. Our use of "variation" (rather than "inconsistency") in this instance follows Bucholtz (2007), who argues that transcription is simply one of many forms of the entextualisation of speech into writing and that, therefore, "there is no reason to expect or demand that it must remain unchanged throughout this process of recontextualization" (p. 802). While we do adopt Bucholtz' view that variation in transcription should not be viewed as the exception but rather the norm, we do, unlike Bucholtz (2007), seek to "problematize variability" (p. 785) insofar as minimizing the chance that such variability may interfere with evidential processes, for instance by misrepresenting the contents of evidential recordings.

At the macro level, we can consider transcription as a political exercise that interfaces with the transcriber's world-view, cultural experiences and sociolinguistic biases (Jaffe 2000). There exists also the continuum between what has been termed "naturalism" and "denaturalism" (Oliver et al., 2005); these concepts relate to the extent to which transcription should aim to capture as much of the detail from the speech signal as possible (naturalism) as opposed to the transcription only capturing what is deemed necessary for a particular purpose (denaturalism). Naturalism, which may be considered "excessive" for some purposes (Clayman and Teas Gill 2012: 123), is commonly found in heavily qualitative approaches such as conversation analysis (CA), while transcription lower on the scale of naturalism (e.g., simple orthographic transcription) tends to be preferred in relatively quantitative approaches such as corpus linguistics (Love 2020) (however, even in this context, transcripts are not highly denaturalised, as there is an explicit focus on recording in orthography the exact linguistic content that was uttered, avoiding paraphrasing). This distinction lends itself to variation in transcription notation and formats according to the style of the transcription, as discussed by Bucholtz (2007). As such, there appears to be a consensus that transcription style should vary according to the purpose of the work: "transcriptions should provide the level of detail required for the job they have to do" (Copland and Creese 2015: 196).

At the micro level, there are issues such as the transcriber's ability to decipher the spoken signal (e.g. due to poor audio quality; see Loubere, 2017), the question of how to select the appropriate orthographic representation of speech signals for which there may be multiple variants, and other sources of potential transcription error (Tessier 2012: 450). These challenges are well-documented, and researchers have discussed the difficulties of transcribing phenomena such as "non-standard" speech (Jaffe 2000), semi-lexical items (Andersen 2016) and the structure of dialogue (Nagy and Sharma 2013), among many others (see Bucholtz, 2007, for a discussion of "orthographic variation"). A crude example of such "orthographic choices" (Nagy and Sharma 2013: 238) is the question of how to transcribe contractions, such as *gonna* (a contraction of *going to*). Whether to represent the contraction orthographically *(gonna)* or standardise it *(going to)* depends upon the purpose of the transcription. Either way, the transcriber(s) should apply the convention consistently. Typically, it is recommended that transcription conventions be developed prior to transcription, to anticipate such issues and prescribe standards so that transcribers may apply such conventions consistently, thus maximising rigour (Lapadat and Lindsay 1999). For example, in the context of the transcription of filled pauses in orthographic spoken corpora, Andersen (2016: 343) advocates for "a 'reductionist approach' in which unmotivated variability is eliminated for the sake of consistency". Conventions may be reviewed and revised during transcription in an iterative manner, as additional unmotivated variability is discovered; as Copland and Creese (2015) discuss (in the context of ethnographic research), "transcription requires the researcher to be reflective and reflexive so that decisions about transcription are consciously made and can be discussed and defended" (p. 191).

However, while transcription conventions may help to reduce unwanted variability, what they cannot control for is the transcriber's perception of the original speech signal; "speech perception involves not recognising sounds but constructing them, via a suite of complex (though almost entirely unconscious) mental processes" (Fraser and Loakes 2020: 409). In other words, a convention about whether to transcribe *gonna* or *going to* assumes that the transcriber actually perceives the production of the word *gonna* in the first place, but this might not always be the case. The transcriber may simply mistake one word for another (Easton et al., 2000), and errors like this may be made more likely if there are complicating factors such as multiple speakers, background noise and/or poor audio quality (Love 2020: 138).

2.2 The Problem of Forensic Transcription

It is known that transcription is a highly challenging and subjective process that is influenced by many factors that are unique to (a) individual transcribers and (b) individual speakers. This has potential implications in contexts where the "accuracy" of a transcript is of critical importance, such as in legal cases. In a forensic context, covert recordings can provide powerful evidence, but are often too low quality to be understood by the jury without the assistance of a transcript. Usually, when transcripts are required they are produced by police officers investigating the case who are granted "ad-hoc expert" status (French and Fraser 2018: 298). The production of such transcripts and their presentation to juries can pose a risk to the delivery of justice in two main ways. The first relates to issues of accuracy and reliability of the transcript produced by the police; the second relates to the impact any (inaccurate) transcript can have on jurors' perception of the content of the recording.

Regarding accuracy and reliability, as has been discussed, producing transcripts of recordings is not a straightforward task, particularly when the recording is of low quality. Therefore, since there is a wide range of factors affecting the accuracy and reliability of forensic transcripts (see Fraser, 2003, for a full discussion of these factors), it is very possible that a police-produced transcript may contain inaccuracy. Notwithstanding the difficulty of perceiving low-quality recording, the skill level and the relationship that police officers have with the material can lead to an inaccurate transcription (Fraser 2014: 10–11). On the one hand, although police officers may be highly trained and skilled in a range of different areas, they likely have no training in linguistics or phonetics and have a lack of reflective practice on speech perception. At the same time, although detailed knowledge of the case, exposure to the material and potential familiarity with the speakers on the recording can be valuable when used in the appropriate way, it can mislead police transcribers rather than help them when producing a transcript (Fraser 2018a: 55; French and Fraser 2018: 300). In the same way as anyone else tasked with listening to and transcribing a spoken recording, police officers rely on "cues" to help them construct words and phrases (Fraser 2021a: 418); that is, they draw on precisely their contextual knowledge of the case, the evidence and the speakers involved when determining what is being said. This can lead to a cognitive bias, over which they have little to no control, which leads transcribers to perceive what they think the recording contains, rather than what it necessarily does contain. Therefore, the police are not independent or impartial transcribers (Fraser 2014: 110) and this can lead to the resultant transcript including content that biases in favour of the prosecution case. This is the argument made by Bucholtz (2009), who demonstrates the ways in which recordings of wire-tapped phone calls between drug dealers are recontextualised in the FBI's "logs" of these conversations. She states that this process is one which "systematically and dangerously disadvantages the speakers whose words are subject to professional representation" (Bucholtz 2009: 519).

The main challenge facing forensic transcription is that "'ground truth' (i.e., indisputable knowledge) regarding the content of the recording cannot be known with certainty" (Fraser 2021a: 428). That is to say that there is no way of knowing precisely what is said in the speech recording, and therefore how this is to be represented or reflected in any transcription. Indeed, it is uncertainty over the content of a recording that is very often the rationale for producing a transcript in the first place. So-called "disputed utterance" cases centre around a section (or sections) of a recording that (1) is potentially evidential or incriminating and (2) causes some disagreement over its content. Fraser (2018b) details a case of this kind in Australia in which a police transcript of an indistinct covert recording included the phrase *at the start we made a pact* and the defendant in question was convicted of being party to a joint criminal enterprise and sentenced to 30 years in prison. However, after being asked to re-examine the audio recording, Fraser (2018b: 595) concluded that "the police transcript was inaccurate and misleading throughout" and "the 'pact' phrase was not just inaccurate but phonetically implausible". Therefore, this transcript, produced with the intention of assisting the jury, is likely to have misled them. This builds on earlier work by Fraser et al. (2011), who clearly demonstrate the extent of influence that transcripts can have on people's perceptions and interpretations of ambiguous or disputed recordings. Their experiments, using a recording from a New Zealand murder case, found that participants' opinions of what was said in the recording changed when they were exposed to different "evidence", including expert opinions on suggested interpretations as to what the recording said. In other words, once the jury were "primed" to hear certain things in the recording, this had a significant impact on their perception and interpretation of the recorded evidence. It is not only disputed utterances that can be the source of dangerous inaccuracies in forensic transcripts; speaker attribution also causes difficulties. As well as transcribing the content of the talk, police transcripts also attribute specific, potentially incriminating, utterances to specific speakers (Fraser 2018a: 55). This challenge is investigated by Bartle and Dellwo (2015: 230), who report a case from the UK Court of Appeal in which police officers' identification of speakers in a recording differed from that of two phoneticians. The police officers' attributions, which were important evidence in the original trial, were ruled as inadmissible and the conviction was overturned.

In summary, it is known that transcription is a highly subjective task that is vulnerable to the influence of transcribers' level of skill, cultural awareness and internal biases. In the context of forensic transcription, this has the potential to lead to errors in the judicial process. In this paper, we seek to explore how variation in transcription manifests linguistically in the written record of what was said and by whom, with the aim of making recommendations to improve the practice of forensic transcription.

3 METHODOLOGY

3.1 Data

This paper reports on the design and results of a controlled transcription experiment in which eight linguistically trained, professional transcribers each transcribed the same audio recording using the same transcription conventions. The transcriptions were generated in the pilot phase of data collection for a large corpus of orthographically transcribed audio recordings known as the Spoken British National Corpus 2014 (Spoken BNC 2014; Love et al., 2017), which was gathered by Lancaster University and Cambridge University Press. The audio recording selected for our experiment is 4 minutes and 4 seconds in length and comprises five adult speakers (3 F, 2 M) having a conversation while dining in a busy restaurant in the north east of England. The recording itself, while not completely indecipherable, contains lots of background noise from other guests in the restaurant, and our assessment of its overall intelligibility is that the recording presents a challenging transcription task. The conversation was recorded using the in-built audio recording function on a smartphone. In the context of covert recordings, this recording shares many of the typical features of covert forensic recordings, including the presence of multiple speakers, background noise and the use of non-specialist recording equipment. Furthermore, the recording was transcribed orthographically, which is a technique commonly used in criminal investigations. It is important to acknowledge that there are some elements of forensic covert recordings that are not simulated here–for example, the device was visible to all speakers (rather than being concealed); all speakers were aware they were being recorded; and, despite the presence of some background noise, the speech signals were not affected by poor quality arising from the recording device being distant from the speakers. Furthermore, the context of transcription is not identical either; our recording was transcribed in a lower-stakes environment than would be the case for forensic transcription, and the transcribers were told beforehand that the recording features five speakers. Therefore, although the recording was not obtained–nor transcribed–in a forensic context, and some elements of our choice of recording may seem advantageous when compared to forensic recordings–we believe there to be enough similarity between our experimental conditions and real-world conditions to warrant use in this study.

As part of the pilot phase of the Spoken BNC2014 compilation, the recording was transcribed independently by eight highly experienced professional transcribers employed by Cambridge University Press. All transcribers are L1 speakers of British English and specialise in producing transcripts for linguistic contexts, for example the English language teaching (ELT)

TABLE 1 | Length of transcripts (words and turns).

Transcript	Length (words)	Turns
A	686	87
B	833	89
C	693	90
D	883	117
E	871	134
F	846	106
G	656	82
H	733	97
Mean	775.13	100.25

industry. They are based in the south of England and do not share the same accent or dialect as the speakers in our recording; however, they were selected for the Spoken BNC2014 project on the basis that they have proficiency in transcribing a diverse range of varieties of English from across the United Kingdom. All transcribers were trained to transcribe the recordings orthographically and received specialist linguistic training in common features of casual British English speech that can be difficult to transcribe (e.g., contractions). Although the transcribers do not possess forensic or phonetic expertise, they are to be considered the industry standard with regard to detailed orthographic transcription.

Consent for the transcriptions to be used in future research was gained from the transcribers at the time of this work in accordance with the ethical procedures of Cambridge University Press, and permission was granted from Cambridge University Press to re-use the transcripts for the present study.

As shown by **Table 1**, the length of the transcripts alone ranges from 656–883 words (mean 775) and 82–134 turns (mean 100), demonstrating that there appears to be substantial variation among the transcripts in terms of the amount of linguistic content transcribed.

3.2 Analytical Procedure

In order to gain a nuanced understanding of the nature and possible causes of the apparent variation–not only in quantity but also in quality–we compared the transcripts against each other, identifying areas of agreement and disagreement in (a) the attribution of the speakers and (b) the representation of the linguistic content. What we do not seek to measure in our analysis is accuracy, since no "ground truth" transcript of the recording exists, i.e. there is no set of "correct answers" with which to compare the transcripts. Our analysis is divided into three parts.

3.2.1 Speaker Attribution

In the first part of our analysis, we investigate the consistency with which transcribers performed speaker attribution, which refers to "the annotation of a collection of spoken audio based on speaker identities" (Ghaemmaghami et al., 2012: 4185). Based on previous research on the manual transcription of casual spoken interactions by Love (2020), we expect speaker attribution to be an area of potential difficulty when transcribing a recording comprising more than two speakers, such as the recording used in this study, which has five speakers. Specifically, Love (2020) found that transcribers tend to attribute speaker ID codes with a high degree of confidence, even when inter-rater agreement and accuracy are only at moderate levels; in other words, it is possible (and perhaps likely, with several speakers) that transcribers will unknowingly attribute the incorrect speaker ID codes to a turn on a routine basis–they will "regularly and obliviously get it wrong" (Love 2020: 156). The main reasons for this are likely to be similarities in the accent and/or voice quality of two or more speakers, and insufficiently clear audio quality. In our recording, four of the five speakers (three of which are females of a similar age) have similar northeast English accents, so we expect accent similarity to be a potential cause of difficulty with regard to speaker attribution.

The first step of this part of our analysis involved aligning the turns in each transcript, so that the speaker attribution of each turn could be compared. We did this firstly by separating the turns in the original transcripts from their corresponding speaker ID codes (labelled 1–5), so that they could be viewed alongside each other as columns in a spreadsheet. Secondly, due to differences in the presentation of turns in the transcripts (which we explore in detail in **Section 4.3**), it was not the case that each turn constituted the same row in the spreadsheet. Some transcribers, for example, split a turn across two lines, with an intervening turn from another speaker–for instance a backchannel–in between; representing a multi-unit turn (Schegloff 2007), while others represented the entire turn on one line. Therefore, the transcripts required editing manually in order to align the turns row by row and facilitate a comparison of the speaker attributions.

The transcripts were produced according to the Spoken BNC2014 transcription conventions (Love et al., 2018), which afforded transcribers three types of speaker attribution to represent the level of confidence with which transcribers could attribute each turn to a speaker:

(1) CERTAIN

- mark the turn using a speaker ID code (e.g. "<0211>")

(2) BEST GUESS

- mark the turn using a 'best guess' speaker ID code (e.g. "<0211?>")

(3) INDETERMINABLE

- mark the turn according to the gender of the speaker (i.e. "<M>" or "<F>") or show that many speakers produced the turn (i.e. "<MANY>")

(Love 2020: 137).

For the sake of analysing inter-rater agreement in this study, the "best guess" codes (those marked with a question mark to indicate lower confidence in their own attribution) were merged with the "certain" codes, i.e., we did not make a distinction between a turn attributed to speaker "4" as opposed to speaker "4?"; we considered both of these as positive attributions of the turn to speaker 4, which contribute to agreement.

Once aligned, we compared the speaker ID codes on a turn-by-turn basis in order to calculate inter-rater agreement for speaker attribution. Using the online tool ReCal OIR (Freelon 2013), we calculated Krippendorff's alpha (Krippendorff 1970), which, in many fields, is a widely applied measure of inter-rater reliability (Zapf et al., 2016), i.e., it can tell us the extent to which the transcribers are in agreement about speaker attributions. Unlike other commonly used measures of inter-rater reliability between three or more coders (e.g., Fleiss' kappa, Fleiss 1971), Krippendorff's alpha (α) accounts for cases where the coders (transcribers, in our case) did not provide a speaker ID code at all. This occurred due to variation among transcribers in terms of the inclusion or omission of entire turns (as discussed in **Section 4.3**), meaning that there are many cases where some (but not all) transcribers included a particular turn, and therefore indicated a speaker ID code. In other words, some turn "slots" in the aligned transcripts are empty and thus were not assigned a speaker ID code.

Krippendorff's α ranges from 0.0 to 1.0, indicating the percentage of the speaker ID codes that are attributed with agreement better than chance. Krippendorff (2004: 241) makes two clear recommendations for the interpretation of the alpha:

- Rely only on variables with reliabilities above α = 0.800.
- Consider variables with reliabilities between α = 0.667 and α = 0.800 only for drawing tentative conclusions.

Based on this, an α of less than 0.667 is to be considered poor inter-rater agreement.

3.2.2 Frequency-Based Lexical Similarity

In the second part, we investigated the extent to which the content of the transcripts, measured in both types and tokens, are shared across the transcripts. Starting with types, we used the detailed consistency relations function in WordSmith Tools (Scott 2020) to calculate the number of types that are present in each pair of transcripts and, among those, the number of types that are shared between each pair. We then calculated the Dice coefficient (Dice 1945) for each pair, which indicates the extent of the overlap between each pair. The Dice coefficient is calculated by dividing the number of types or tokens that is shared among two transcripts by the total number of types or tokens present in both transcripts taken together, as per the following formula:

$$(J \times 2)/(F1 + F2)$$

where J = shared types or tokens; F1 = transcript 1 total types or tokens; F2 = transcript 2 total types or tokens (adapted from Scott 2007).

The resulting Dice coefficient ranges from 0.0 to 1.0 and can be taken as a proportion of overlap between the two transcripts, i.e. the closer the coefficient to 1.0, the more overlap in the types or tokens present in the two transcripts (where 0.0 is no overlap whatsoever and 1.0 is complete overlap).

An admittedly crude measure of similarity between transcripts, what our approach does reveal is the extent to which transcripts differ in the quantity of content they contain. In an ideal world, each transcript would be identical, and therefore they would each fully overlap with each other in terms of the types and the frequency of tokens present (as indicated by a Dice coefficient of 1.0). Thus, differences in the number of types and tokens in the transcripts would be indicative of differences in the transcriptions.

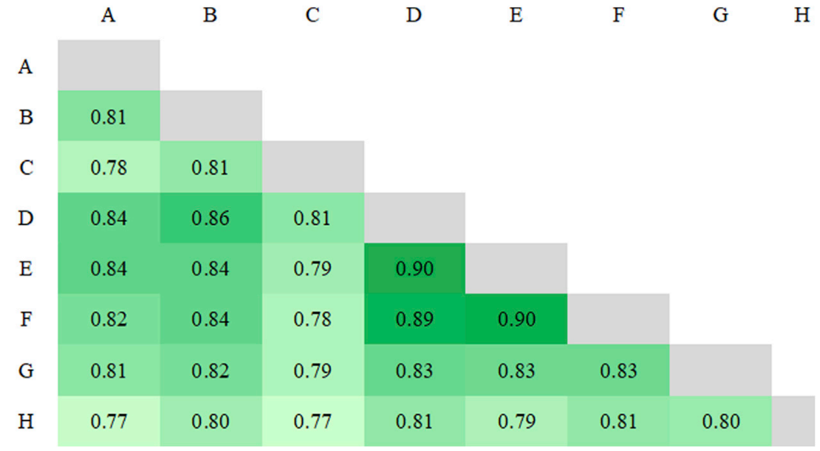
3.2.3 Turn-Based Transcription Consistency

In the final stage of our analysis, we investigated the representation of linguistic content among the transcripts on a turn-by-turn basis. In an ideal world, all eight transcribers would produce identical transcripts of the recording, and this would be maximally desirable in forensic transcription. For that reason, in this analysis, we refer to transcribers being "consistent" with each other when they produce exactly the same linguistic content for a given turn.

Using the aligned transcripts, we compared the linguistic representation of each turn across all transcribers quantitatively and then qualitatively. We started by quantifying the extent to which each version of a given turn was transcribed identically. We did this by comparing the transcription of each turn and counting how many versions of each turn across transcripts were completely identical (out of a possible total of eight, which would indicate perfect agreement across all transcribers). We then counted how many of the turns were matching for each number of transcribers–a match for only one transcriber meant that each version of the transcribed turn was different to the other, i.e., no two (or more) versions matched. In doing so, we considered the presence of empty turn "slots", as caused by the omission of turns by some of the transcribers. If two or more transcribers omitted the same turn, we did not consider this a form of matching, as we cannot prove that the omission of a turn is a deliberate transcription choice, as opposed to being a result of a transcriber simply not having perceived the turn in the audio recording. Therefore, we deemed this an unreliable measure of consistency, and only considered matching among turns that had actually been transcribed.

This approach provides a broad overview of the consistency of transcription, but it is a blunt instrument, making no distinction between minor and major discrepancies between transcribers; nor does it take into account the nature or apparent causes of the discrepancies. Therefore, our next step was to manually examine each set of turns, qualitatively categorising the main cause of variation for each. This was conducted together by both authors in order to maximise agreement in our coding.

To conduct this analysis, we made some further methodological decisions with regard to features of the Spoken BNC2014 transcription scheme (Love et al., 2018). In the transcription scheme, transcribers are instructed to mark the presence of a turn even if they could not decipher the linguistic content of the turn. For the purposes of our analysis, we disregarded such cases and treated them as omissions, as they did not provide any linguistic content to be compared against other versions of the same turn. Of course, in forensic contexts, for an expert transcriber to acknowledge that a section of speech is not transcribable may be meaningful in some cases; however, our focus is on investigating the linguistic content that has been transcribed, and so we chose to omit turns marked as "unclear" from our analysis. Additionally, we decided to disregard the presence or absence of question marks (the only punctuation



	A	B	C	D	E	F	G	H
A								
B	0.81							
C	0.78	0.81						
D	0.84	0.86	0.81					
E	0.84	0.84	0.79	0.90				
F	0.82	0.84	0.78	0.89	0.90			
G	0.81	0.82	0.79	0.83	0.83	0.83		
H	0.77	0.80	0.77	0.81	0.79	0.81	0.80	

FIGURE 1 | Dice coefficient heatmap for types, graded from lower (light green) to higher (dark green) values.

character allowed as part of the transcription scheme, besides tags; Love et al., 2018: 37) as a marker of transcription variation, as we focussed solely on the consistency of the linguistic content.

Once each turn was coded according to the main source of inconsistency (where present), these were categorised to form the basis of our discussion in **Section 4.3**.

4 RESULTS

4.1 Speaker Attribution

Using Krippendorff's alpha (Krippendorff 1970), we calculated the extent of inter-rater agreement for speaker attribution among the eight transcripts. This revealed that across all transcripts and turns, α = 0.408, meaning that only a little over 40% of the turns were attributed to speakers with better-than-chance agreement. While not a direct measure of speaker attribution accuracy (as no 100% correctly attributed "ground truth" transcript exists), the extent of disagreement between transcribers with regards to speaker attribution is a clear indication of inaccuracy; if two (or more) transcribers disagree about a turn, then at least one of the transcribers must have attributed the turn incorrectly.

The possible implications of such a low level of agreement between transcribers in terms of the representation of linguistic content are explored in **Section 5**.

4.2 Frequency-Based Lexical Similarity: Types and Tokens

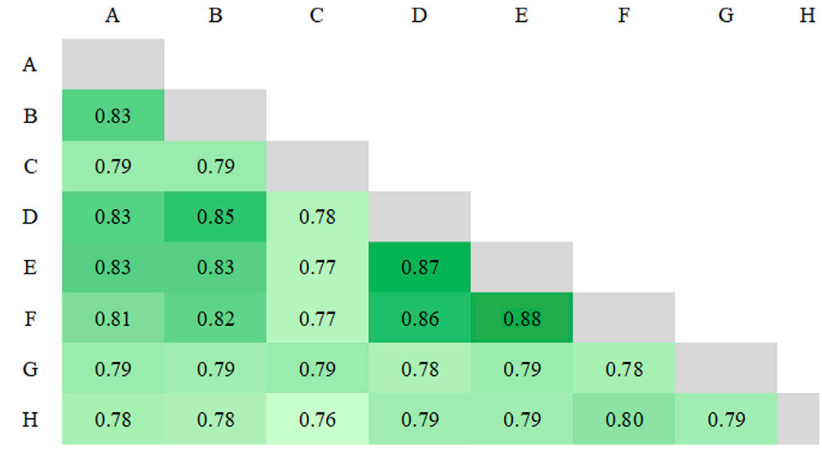
Next, we present the comparison of similarity between transcripts with regard to the types present in each transcript and the number of tokens that are shared. **Figure 1** is a heatmap displaying the Dice coefficient values for each pairwise comparison of type overlap between transcripts. The Dice coefficient results range from 0.77 (pairs AH and CH) to 0.90 (pairs DE and EF), with a mean of 0.82, indicating that a majority of types occur at least once in each transcript pair. However, this also shows that (a) across each pair, there are some types (between 10–23%) that occur in one but not the other transcript, and (b) there is a fair amount of variation between pairs of transcripts, i.e., some transcribers are more consistent with some of their fellow transcribers than others.

Our analysis of similarity in terms of types is limited in that it does not take into account the frequency of each type; it calculates overlap in a binary fashion, based simply on the presence or absence of types (regardless of how many times the type occurs, if present). Therefore, we repeated our analysis using the raw frequencies of each individual token in the transcripts. The heatmap displaying the Dice coefficients results for each pairwise comparison of token frequency are shown in **Figure 2**. The values range from 0.76 (pair CH) to 0.88 (pair EF), with a mean of 0.80, indicating a slightly lower range of overlap when compared to that of the comparison of types. Again, while the values indicate a majority overlap between each pair, between 12–24% of tokens that are present in a given transcript are absent in another.

These comparisons provide a crude indication that there are substantial differences in the content of the transcripts, the specific nature of which requires qualitative examination, which we discuss in the next section.

4.3 Turn-Based Transcription Consistency

Finally, we present the findings of our analysis of the linguistic content on a turn-by-turn basis. Starting with a broad measure of the extent to which turns matched exactly, we found generally low levels of consistency across the eight transcribers in terms of how they transcribed each of the 170 turns. Only five of the 170 turns (2.94%) are transcribed identically by all eight transcribers. All five of these represent minimal speech, with the longest consistently transcribed turn being *yeah it is*. There are two instances where *yeah* was transcribed by all eight transcribers and the remaining two turns are the non-lexical agreement token *mm*. Therefore, this leaves 165 of the 170 turns in which there was inconsistency across the eight transcribers. This ranges from cases in which there was consistency across seven of the eight transcribers, with only one transcriber differing from the



	A	B	C	D	E	F	G	H
A								
B	0.83							
C	0.79	0.79						
D	0.83	0.85	0.78					
E	0.83	0.83	0.77	0.87				
F	0.81	0.82	0.77	0.86	0.88			
G	0.79	0.79	0.79	0.78	0.79	0.78		
H	0.78	0.78	0.76	0.79	0.79	0.80	0.79	

FIGURE 2 | Dice coefficient for tokens, graded from lower (light green) to higher (dark green) values.

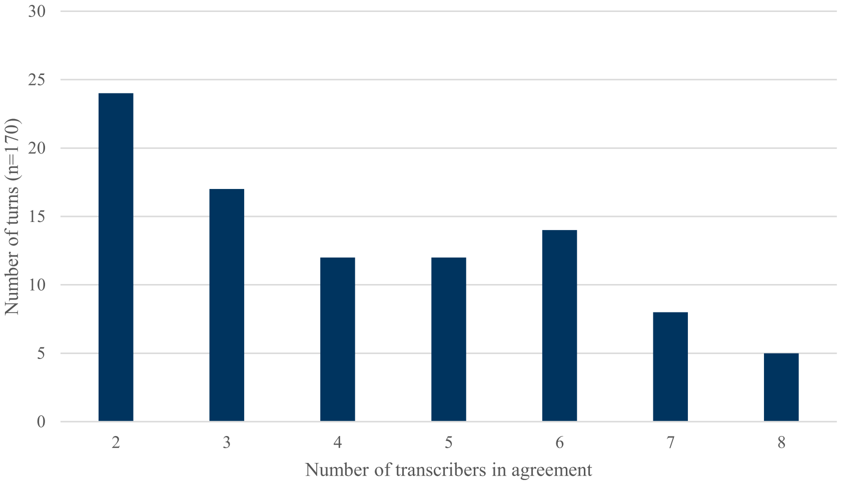


FIGURE 3 | Number of turns transcribed consistently by transcribers.

others, to cases where all eight transcribers transcribed a given turn differently. **Figure 3** shows that the lack of consistency between transcribers is striking. By far the most common occurrence, accounting for 78 of the 170 turns (45.88%), sees only one transcriber "in agreement", meaning in reality that each of the eight transcribers transcribed the turn differently to the other. In fact, in only 24 of 170 turns (14.12%) do any two of the eight transcribers agree on the content of the recording, and this number reduces as the number of transcribers increases. To generalise, only 39 out of the 170 turns (22.94%) were transcribed consistently by the majority of transcribers (i.e., more than four of the eight).

This binary measuring of (in)consistency on the basis of transcribers producing an identical transcription for each turn masks the fact that, while some versions of the transcribed turns produced by different transcribers are very similar, others vary substantially. In turn, this variation and difference is manifest in a number of different ways–what we refer to here as "sources of variation". In each of the 165 turns where there was some variation among the transcribers, we qualitatively identified and categorised the source of variation in terms of precisely how the transcripts differed or on what basis they disagreed with one another. We identified the following sources of variation:

- Omitted or additional speech
- Splitting of turns
- Phonetic similarity
- Lexical variation

There is also one instance of inconsistency based on the transcription convention itself; this relates to a part of the recording in which a place name was mentioned, and some transcribers anonymised the place name while some did not. Because this inconsistency relates to the parameters of the

TABLE 2 | Extract 1 (S= Speaker).

Transcriber B		Transcriber C		Transcriber D	
S	**Turn**	**S**	**Turn**	**S**	**Turn**
1	why you're ruining it	1	why? you're ruining it		
4	ooh because I can't eat it any other way	4	because I can't eat it any other way	F	because I can't eat it any other way
		1	it's like eating an old boot	1	it's like eating the boot of your

TABLE 3 | Extract 2.

Transcriber B		Transcriber C	
S	**Turn**	**S**	**Turn**
4	that's what I was thinking	4	that's what I said
3	don't get too excited		
1	it must be like a shot glass of chicken tikka masala	1	like a shot glass of chicken tikka masala

transcription set out in the experiment, rather than the content of the recording itself, we will not consider this instance any further. The remainder of this analysis will describe and demonstrate each of the other types of inconsistency, drawing on examples in the data to show how transcribers varied in their transcriptions of the same recording.

4.3.1 Omitted or Additional Speech

Some transcriptions of the turn contained more or less speech content than others. The most straightforward example of this is in turns where some of the transcribers identify and transcribe a speaker turn while others do not. In some cases, there is a high level of consistency across transcribers, and the amount or nature of omitted or additional speech is minimal. In one turn, shown in **Table 2**, all eight transcribers agreed on the transcription *because I can't eat it any other way*. The only variation here is that Transcriber B included an *ooh* as a preface to the utterance and this is something that was not found in any of the other transcripts.

In other cases, however, there is less consistency across transcribers. In one turn, for instance, four of the eight transcribers agreed that the turn in the recording *was don't get too excited*, while the other half of the transcribers not only left that turn blank but did not include *don't get too excited* anywhere in their transcript. **Table 3** shows an example of this by comparing two of the transcribers. In cases such as this, it is evident that some transcribers are hearing some talk that others are not, or are at least including talk in their transcripts that is absent from others'. This is perhaps the starkest type of difference or inconsistency between transcribers. When tasked with representing the same recording in a transcript, some identify elements of talk that others do not, including full utterances. The implications of this in a forensic context are clear and problematic; it might be that an evidentially significant utterance that is identified in one transcript is missing altogether from another.

Even obtaining two transcripts of a given recording may not suffice in insuring against omitted utterances. There are other instances in our data where an utterance is transcribed by only one of the eight transcribers. For example, **Table 4** compares the work of two transcribers and shows that, not only is there a lack of agreement on who spoke the second turn (albeit the transcription of this turn is very similar in terms of content), but each transcript sees an utterance transcribed that does not appear in any of the other seven transcripts. For Transcriber E, this is an attribution of Speaker 2 saying *it's hard to find exactly what this stuff is*, while Transcriber H represents Speaker 1 as saying *if you just count it you just count the calories*. The fact that these utterances are only found in the transcripts of one of the eight transcribers reflects the extent of the problem of omitted/additional speech and the discrepancies in the output of different transcribers. However, it also raises an important question as to which is the best interpretation of such instances. It is unclear whether cases such as these should be viewed as seven transcribers missing talk that one hears, or whether one transcriber is contaminating their transcript with talk that only they (think) they hear. In other words, in a forensic context, a question arises as to whose transcript(s) do we trust the most. There is a judgement to be made as to whether more weight is given to the one transcript that does include an utterance or the fact that seven other transcribers do not report hearing that utterance.

4.3.2 Splitting of Turns

The omission of speech that we have seen above can have further consequences for the transcription. Namely, the decision to include an utterance or not can affect the representation of the turn sequences in the transcript. **Table 5** is a case in point. Here, transcribers C and D choose to represent overlapping speech by an unidentifiable but "female" speaker in *yeah*. The way in which this overlapping speech is included is such that it splits the turn of Speaker 1 before *seven hundred and ninety six calories*, and this is the same in both transcripts. Transcriber A and B, on the other hand, do not choose to represent the overlapping *yeah*. Therefore, for them, Speaker 1's utterance is represented in full and uninterrupted, forcing a difference between their version and those of transcriber C and D.

TABLE 4 | Extract 3.

Transcriber E		Transcriber H	
S	Turn	S	Turn
5	can you please tell me how every raffle you seem to go into at the minute you win but we win jack shit on the lottery?	5	can you please tell me how every raffle you seem to go into at the minute you win but we win jack shit on the lottery?
4	it's it's quite big (.) and especially if you go large (.) I'm sure if you I if you go large you've gotta add the extra on but	3	it's it's quite big and especially if you go large I am sure if you if you go large you've got to add the extra on but
2	It's hard to find exactly what this stuff is	1	if you just count it you just count the calories

TABLE 5 | Extract 4.

Transcriber A		Transcriber B		Transcriber C		Transcriber D	
S	Turn	S	Turn	S	Turn	S	Turn
1	I also think unless that bowl of chips is huge it's not gonna be seven hundred and ninety-six calories	1	I also think unless that bowl of chips is huge it's not going to be seven hundred and ninety six calories	1	I also think that unless that bowl of chips is huge it's not gonna be	1	I also think that unless that bowl of chips is huge it's not gonna be
				F	yeah	F	yeah
				1	seven hundred and ninety six calories	1	seven hundred and ninety-six calories

TABLE 6 | Extract 5.

Transcriber A		Transcriber C		Transcriber D		Transcriber H	
S	Turn	S	Turn	S	Turn	S	Turn
4	chicken tikka masala	4	chicken masala chicken balti	F	chicken tikka masala	4	chicken tikka masala chicken balti
F	mm	5	mm	5	mm	5	mm
4	chicken balti			F	chicken balti		

TABLE 7 | Extract 6.

Transcriber B		Transcriber C		Transcriber F	
S	Turn	S	Turn	S	Turn
5	but I'm just looking at	5	cos I'm struggling can't read any of it	5	cos I was looking at it I can't I can't read any of it
4	no I really struggled with it it's like [place] but visualised	3	no I really struggled with it it's like a may get in visualised	4	no I really struggled with it
				3	it's like [place] but visualized

Such differences in turn splitting do not only appear as a result of the inclusion or omission of overlapping speech. In **Table 6**, for example, all transcribers transcribed the *mm* feedback by Speaker 5 (for reasons of space, only four transcripts are shown here). However, despite all transcribers agreeing that some overlapping speech can be heard, they disagreed on how they represented the initial turn; while transcribers A and D chose to place *chicken balti* as a new turn, transcribers C and H did not. The inclusion and/or placement of overlapping speech in a transcript is an important element of the talk being represented in terms of the implications that it has for other turns and the chronology of the unfolding talk.

A final factor that can result in transcriptions varying in terms of turn completion and turn splitting is variation in speaker attribution. **Table 7** shows three transcribers–B, C and F–who vary in terms of to which speaker they attribute a turn. With transcriber B and C, this is a straightforward disagreement; the speaker is identified as Speaker 4 and Speaker 3 respectively. Even though the transcribers disagree on which speaker uttered the turn, they do agree that the full turn was spoken by the same speaker. Transcriber F, in contrast, believes this not to be one turn, but in fact two turns spoken by two different speakers (Speaker 4 and then Speaker 3). Disagreement in terms of "who said what" can have clear implications in a forensic context, and

TABLE 8 | Extract 7.

Transcriber A		Transcriber B		Transcriber C	
S	Turn	S	Turn	S	Turn
1	super food pasta	1	super food pasta	1	super food pasta
2	cos that looks ostensively like how we'd be able to have it	2	cos that looks ostensibly like how we'll be able to have it	3	cos that looks extensively like how we'd be able to have it
4	oh she's starting already	3	ooh she's starting already	4	oh she's starting already

TABLE 9 | Extract 8.

Transcriber B		Transcriber D	
S	Turn	S	Turn
3	Yeah	F	yeah
2	So I'm gonna try it cos then if I like it I can have it if I'm out	F	so I'm gonna try it cos then if I like it I can have it every night
1	what's in the chicken breast	1	want some chicken breast in there

an example such as this brings into sharp focus how differing speaker attributions can result in problematically different transcripts.

4.3.3 Phonetic Similarity

The phonetic similarity between words that gives rise to ambiguity and the resultant challenges to transcription are well-documented. Coulthard et al. (2017: 132) describe a drug case in which there was a dispute over whether a word in a recording was *hallucinogenic* or *German* in a police transcript. A second example from Coulthard et al. (2017) is a murder case which involved a transcript of talk from a murder suspect in which the utterance *show[ed] a man ticket* was erroneously transcribed as the phonetically similar *shot a man to kill*. The mistaking of one word (or phrase) for another that shares some sound similarities with another word can have serious implications in a forensic transcript, particularly when the words have different meanings and, in the context of the case, those differences are significant. It may be, for example, that an innocuous word is transcribed as an incriminating word.

In our data, we found many instances of transcripts containing different but similar-sounding words in the same turn. For our purposes, phonetic similarity was determined impressionistically on the basis of a judgement of two words sharing phonemes. **Table 8** is an example of this, showing a turn in which the same word is transcribed three different ways: **ostensively, ostensibly* and *extensively*. Across all eight transcribers, five transcribed this word as *ostensibly*, two as *extensively* and one as **ostensively*. It is worth noting that, besides the variation in this word, the content of three transcripts is very similar. Notwithstanding that **ostensively* is not a word, although *ostensibly* and *extensively* sound similar, they have very different meanings. In this experimental context, this difference is not of great significance, but in a forensic context this difference could have serious implications.

In the case of *ostensibly/extensively* the choice of either word has implications for the meaning of the full turn. However, the variation across the transcripts is essentially restricted to one word. There are other cases in our data in which longer phrases with phonetically similar properties are found to differ across transcripts. An example of this is in **Table 9**, where two transcribers vary in their transcription of *what's in* and *want some*. This shows that the influence of phonetic similarity can stretch beyond individual words and affect the perception and transcription of multi-word utterances. In deciding between *ostensibly* and *extensively*, contextual cues can be used by transcribers to determine which of the two words makes the most "sense" within the given utterance, and this can influence the choice between two words which sound similar, but which match the semantics of the sentence to different degrees. In the case of **Table 8**, it might be that *ostensibly* makes more semantic sense than *extensively* in the broader context of the talk. In contrast, neither *what's in* or *want some* is the obvious candidate in the context of the turn in **Table 9**. In such cases, the ambiguity may be insurmountable, and to choose one option over the other would do more damage than marking the word as indecipherable or inaudible.

Finally, where phonetic similarity accounts for variation in transcription between different transcribers, this variation not only has the potential to affect individual words or larger multi-word units (changing the semantics of the utterance in the process), but can also change the perceived pragmatic purpose or force of a given turn. This is exemplified in **Table 10**, in which the phonetic indistinguishability of *can* and *can't* and *light* and *late* can see the same turn be transcribed as a statement by some transcribers (B and C) and a question by others (A). As we saw above, these three transcripts are generally very similar, but diverge on the basis of phonetic similarity. In almost all communicative contexts, the pragmatic difference between a question and a statement is significant in terms of speaker intent and knowledge, both of which can be central to (allegedly) criminal talk.

TABLE 10 | Extract 9.

Transcriber A		Transcriber B		Transcriber C	
S	**Turn**	**S**	**Turn**	**S**	**Turn**
3	twenty-fourth of the fourth in the wallet getting drunk	3	twenty-fourth of the fourth in the <place > getting drunk	3	twenty fourth of the fourth in the wallow getting drunk
4	er some of us are	4	er some of us are	4	some of us are
2	can you see in this light? or maybe my eyes just don't see (.) how can chicken tikka masala only be four hundred and fifty calories?	2	I can't see in this light or maybe my eyes just don't see (.) how can chicken tikka masala only be four hundred and fifty calories?	3	seeing this late or maybe my eyes just don't see (.) how can chicken tikka masala only be four hundred and fifty calories

TABLE 11 | Extract 10.

Transcriber A		Transcriber B		Transcriber D	
S	**Turn**	**S**	**Turn**	**S**	**Turn**
1	that was have some huge like deep fried three times the calories	1	that was absolutely huge and like deep fried three times to	1	that was absolutely huge and like deep fried three times to bring up the calories (.) nice of them
2	I could quite go for that pasta	2	I think I might go for that pasta	F	I quite like the look of that pasta
4	which one?	4	mm?	F	which one?
2	that one			F	it's that one

TABLE 12 | Extract 11.

Transcriber D		Transcriber E		Transcriber G	
S	**Turn**	**S**	**Turn**	**S**	**Turn**
5	cos I'll buy one as well	5	buy one aswell		
F	yeah				
F	or well no that's it there he reckons if you go large and add the samosa and a large onion bhaji	F	it reckons if you go large and add the samosa and a large onion bhaji it's only two hundred and forty calories	F	I reckons if you go large and add the small onion bhaji it's only two hundred and forty calories
F	Mm	F	mm that sounds nice		

4.4.4 Lexical Variation

In the previous section, we showed how transcripts can include different versions of the same utterance and how those differences can be accounted for by some sound similarity between the different versions. However, in our data, we also found many instances where the lexical content of the transcribed turns differed in contexts where there was seemingly no phonetic explanation for that difference. We have called this lexical variation.

In **Table 11**, for example, we see three versions of the same turn across three transcribers. The location of the variation here is in the verb phrase, *I could quite go for*, *I think I might go for* and *I quite like the look of*. The versions by transcribers A and B at least share the same main verb *go for*, but there is variation in the premodification. What is key here is that there is a clear lexical intrusion between *could quite* and *think I might* in the latter that cannot be straightforwardly accounted for by phonetic similarity.

Another, possibly more noteworthy, example of this is shown in **Table 12**. Here, the three transcripts are consistent in their inclusion of *reckons if you go large*. However, the key lexical difference is that each of the transcripts has a different pronoun as the subject of reckons: *he* (D), *it* (E) and *I* (G). Although this is a very small lexical difference, it has significant consequences insofar as it attributes agency to different people or things. In a casual conversation such as that recorded here, this may not be important, but the implications of the difference between *he, it* and *I* in a forensic context are clear in terms of responsibility and agency.

In terms of agency and action, we not only see inconsistencies in subject allocation but also main verbs themselves. **Table 11** above saw variation in the premodification of main verbs, but **Table 13** shows how, while six of the eight transcripts include one verb, another includes a different, unrelated verb. There is no phonetic similarity that would explain a disagreement between *said* and *was thinking*, and both make sense in context. Incidentally, the difference between saying something and thinking something could be the difference between committing and not committing a criminal offence. Although inconsequential in this recording, the (mis)identification of one verb as another could have substantial consequence in criminal and forensic contexts.

5 DISCUSSION

Forensic transcription faces many difficult challenges regarding the accurate and reliable representation of spoken recordings and

TABLE 13 | Extract 12.

Transcriber C		Transcriber D		Transcriber E	
S	Turn	S	Turn	S	Turn
F	I won the raffle	F	I won the raffle		
F	only be four hundred and fifty calories?				
4	that's what I said	F	that's what I was thinking	F	that's what I was thinking
		1	must be like		

the effect that transcriptions have on juries' perception of the evidence presented. Fraser (2021a) proposes that, in order to address these issues, and to ensure that transcripts used in forensic contexts are reliable, a branch of linguistic science dedicated specifically to the study of transcription is required. This study has aimed to move in this direction by providing empirical evidence from a transcription experiment that observes the extent and nature of variability across transcripts of the same recording. The primary motivation of this experiment and subsequent analysis has been to inform reflective practice and shed light on the process of transcription in new ways.

We have made the argument that the recording used for this experiment shares important similarities with the types of (covert) recordings that are likely to be central to forensic evidence. Relevant factors are that there are multiple speakers and the recording was taken on a smartphone in a busy environment with background noise. However, it should also be emphasised that the eight transcribers compared here did not anticipate their transcriptions to be analysed from a forensic perspective. For example, they were not directed to produce a transcript as if it were to be used as evidence in court. Had such an instruction been given, this may have motivated greater care and attention than was used (or indeed required) for the original task.

In terms of developing methodologies for a science of transcription, this paper proposes three ways in which different transcriptions of the same recording can be compared. We acknowledge that each of these methods have their own unique caveats and areas for refinement, but they are offered here as foundations for future work. They are: (i) measures of inter-rater reliability to evaluate speaker attribution, (ii) the use of the Dice coefficient to measure lexical similarity across transcripts in terms of types and tokens, and (iii) a qualitative approach to identifying patterns in variation at the level of the turn.

The findings of the analysis revealed that, generally, there is a substantial level of variation between different transcripts of the same recording. In terms of speaker attribution, agreement of who said what was just over 40%. In terms of lexical overlap, transcripts averaged 82% similarity in terms of word types, and 80% in terms of tokens. Finally, in terms of consistency across transcripts at the level of the turn, transcribers varied in terms of the speech included or omitted, the representation of overlapping speech and turn structure, and the representation of particular words or phrases, some of which seems to be motivated by phonetic similarity, while for others the source of difference is more difficult to ascertain.

It is clear that the interpretation of (indistinct) audio recordings, forensic or otherwise, is not simply a case of 'common knowledge', that can be left in the hands of the police or, indeed, the jury (Fraser 2018c: 101). Our results suggest that even trained transcribers do not produce transcripts "bottom-up", and that disagreements between transcripts are common. Our interpretation of these findings is emphatically not that transcription is too difficult to be useful, or that forensic transcription should not be carried out at all. Rather, we believe our findings reveal that even professional transcribers vary in their perception and interpretation of recorded talk. The task of improving the practice of forensic transcription should not lie in attempting to completely eliminate variation, but rather to minimize the influence of variation on evidential and judicial processes. As such, at the most basic level, our findings emphasise and underline the argument that transcription should not be undertaken solely by police officers who are untrained in linguistics.

Our aim here is to take into consideration the findings of this work and use them to begin to develop frameworks and protocol for the management of forensic transcription. The extent to which this is achieved or achievable, in many ways, will be determined by future research and practice in this area.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: The data used in this study are part of the pilot phase data from the Spoken British National Corpus 2014 project. Requests to access these datasets should be directed to RL, r.love@aston.ac.uk.

ETHICS STATEMENT

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

RL: Conceptualization, Methodology, Formal Analysis, Investigation, Data Curation, Writing–Original Draft, Writing–Review and Editing, Project Administration, Funding Acquisition. DW: Conceptualization, Methodology, Formal Analysis, Investigation, Writing–Original Draft, Writing–Review and Editing, Visualization, Funding Acquisition.

ACKNOWLEDGMENTS

We are grateful to Dr Claire Dembry for granting us permission on behalf of Cambridge University Press for the re-use of pilot data from the Spoken British National Corpus 2014 project.

REFERENCES

Andersen, G. (2016). Semi-lexical Features in Corpus Transcription: Consistency, Comparability, Standardisation. *Int. J. Corpus Linguistics* 21 (3), 323–347. doi:10.1075/ijcl.21.3.02

Bartle, A., and Dellwo, V. (2015). Auditory Speaker Discrimination by Forensic Phoneticians and Naive Listeners in Voiced and Whispered Speech. *Int. J. Speech, Lang. L.* 22 (2), 229–248. doi:10.1558/ijsll.v22i2.23101

Bucholtz, M. (2007). Variation in Transcription. *Discourse Stud.* 9 (6), 784–808. doi:10.1177/1461445607082580

Bucholtz, M. (2009). Captured on Tape: Professional Hearing and Competing Entextualizations in the Criminal Justice System. *Text & Talk.* 29 (5), 503–523. doi:10.1515/text.2009.027

Clayman, S. E., and Teas Gill, V. (2012). "Conversation Analysis," in *The Routledge Handbook of Discourse Analysis*. Editors J. P. Gee and M. Hanford (London: Routledge), 120–134.

Copland, F., and Creese, A. (2015). *Linguistic Ethnography: Collecting, Analysing and Presenting Data*. London: SAGE.

Coulthard, M., Johnson, A., and Wright, D. (2017). *An Introduction to Forensic Linguistics: Language in Evidence*. London: Routledge.

Davidson, C. (2009). Transcription: Imperatives for Qualitative Research. *Int. J. Qual. Methods* 8, 35–52. doi:10.1177/160940690900800206

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology* 26 (3), 297–302. doi:10.2307/1932409

Easton, K. L., McComish, J. F., and Greenberg, R. (2000). Avoiding Common Pitfalls in Qualitative Data Collection and Transcription. *Qual. Health Res.* 10 (5), 703–707. doi:10.1177/104973200129118651

Fleiss, J. L. (1971). Measuring Nominal Scale Agreement Among many Raters. *Psychol. Bull.* 76 (5), 378–382. doi:10.1037/h0031619

Fraser, H. (2003). Issues in Transcription: Factors Affecting the Reliability of Transcripts as Evidence in Legal Cases. *Int. J. Speech, Lang. L.* 10 (2), 203–226. doi:10.1558/sll.2003.10.2.203

Fraser, H. (2014). Transcription of Indistinct Forensic Recordings: Problems and Solutions From the Perspective of Phonetic Science. *Lang. L. / Linguagem e Direito.* 1 (2), 5–21.

Fraser, H. (2018a). Covert Recordings Used as Evidence in Criminal Trials: Concerns of Australian Linguists. *Judicial Officers' Bull.* 30 (6), 53–56. doi:10.3316/INFORMIT.728989125075618

Fraser, H. (2018b). 'Assisting' Listeners to Hear Words that Aren't There: Dangers in Using Police Transcripts of Indistinct Covert Recordings. *Aust. J. Forensic Sci.* 50 (2), 129–139. doi:10.1080/00450618.2017.1340522

Fraser, H. (2018c). Thirty Years Is Long Enough: It's Time to Create a Process That Ensures covert Recordings Used as Evidence in Court Are Interpreted Reliably and Fairly. *J. Judicial Adm.* 27, 95–104.

Fraser, H. (2021a). "Forensic Transcription: The Case for Transcription as a Dedicated Area of Linguistic Science," in *The Routledge Handbook of Forensic Linguistics*. Editors M. Coulthard, A. Johnson, and R. Sousa-Silva. 2nd edn. (London: Routledge), 416–431.

Fraser, H. (2021b). The Development of Legal Procedures for Using a Transcript to Assist the Jury in Understanding Indistinct covert Recordings Used as Evidence in Australian Criminal Trials A History in Three Key Cases. *Lang. L.* 8 (1), 59–75. doi:10.21747/21833745/lanlaw/8_1a4

Fraser, H., and Loakes, D. (2020). Acoustic Injustice: The Experience of Listening to Indistinct Covert Recordings Presented as Evidence in Court. *L. Text Cult.* 24, 405–429.

Fraser, H., Stevenson, B., and Marks, T. (2011). Interpretation of a Crisis Call: Persistence of a Primed Perception of a Disputed Utterance. *Int. J. Speech, Lang. L.* 18 (2), 261–292. doi:10.1558/ijsll.v18i2.261

Freelon, D. (2013). ReCal OIR: Ordinal, Interval, and Ratio Intercoder Reliability as a Web Service. *Int. J. Internet Sci.* 8 (1), 10–16. Available at: http://dfreelon.org/utils/recalfront/recal-oir/.

French, P., and Fraser, H. (2018). Why 'Ad Hoc Experts' Should Not Provide Transcripts of Indistinct Forensic Audio, and a Proposal for a Better Approach. *Criminal L. J.* 42, 298–302.

Ghaemmaghami, H., Dean, D., Vogt, R., and Sridharan, S. (2012). Speaker Attribution of Multiple Telephone Conversations Using a Complete-Linkage Clustering Approach. *Speech Signal. Process. (Icassp).*, 4185–4188. doi:10.1109/icassp.2012.6288841

Jaffe, A. (2000). Introduction: Non-Standard Orthography and Non-Standard Speech. *J. Sociolinguistics.* 4 (4), 497–513. doi:10.1111/1467-9481.00127

Jenks, C. J. (2013). Working With Transcripts: An Abridged Review of Issues in Transcription. *Lang. Linguistics Compass.* 7 (4), 251–261. doi:10.1111/lnc3.12023

Kirk, J., and Andersen, G. (2016). Compilation, Transcription, Markup and Annotation of Spoken Corpora. *Int. J. Corpus Linguistics.* 21 (3), 291–298. doi:10.1075/ijcl.21.3.01kir

Krippendorff, K. (1970). Estimating the Reliability, Systematic Error and Random Error of Interval Data. *Educ. Psychol. Meas.* 30 (1), 61–70. doi:10.1177/001316447003000105

Krippendorff, K. (2004). *Content Analysis: An Introduction to its Methodology*. Thousand Oaks, California: SAGE.

Lapadat, J. C., and Lindsay, A. C. (1999). Transcription in Research and Practice: From Standardization of Technique to Interpretive Positionings. *Qual. Inq.* 5 (1), 64–86. doi:10.1177/107780049900500104

Loubere, N. (2017). Questioning Transcription: The Case for the Systematic and Reflexive Interviewing and Reporting (SRIR) Method. *Forum Qual. Soc. Res.* 18 (2), 15. doi:10.17169/fqs-18.2.2739

Love, R., Dembry, C., Hardie, A., Brezina, V., and McEnery, T. (2017). The Spoken BNC2014. *International Journal of Corpus Linguistics* 22 (3), 319–344. doi:10.1075/ijcl.22.3.02lov

Love, R., Hawtin, A., and Hardie, A. (2018). The British National Corpus 2014: User Manual and Reference Guide. Lancaster: ESRC Centre for Corpus Approaches to Social Science. Available at: http://corpora.lancs.ac.uk/bnc2014/doc/BNC2014manual.pdf.

Love, R. (2020). *Overcoming Challenges in Corpus Construction: The Spoken British National Corpus 2014*. New York: Routledge.

Nagy, N., and Sharma, D. (2013). "Transcription," in *Research Methods in Linguistics*. Editors R. Podesva and D. Sharma (Cambridge, 235–256.

Oliver, D. G., Serovich, J. M., and Mason, T. L. (2005). Constraints and Opportunities With Interview Transcription: Towards Reflection in Qualitative Research. *Social Forces.* 84 (2), 1273–1289. doi:10.1353/sof.2006.0023

Schegloff, E. A. (2007). *Sequence Organisation in Interaction: A Primer in Conversation-Analysis*. Cambridge: Cambridge University Press.

Scott, M. (2007). Formulae. Retrieved From WordSmith Tools. Available at: https://lexically.net/downloads/version5/HTML/index.html?formulae.htm.

Scott, M. (2020). *WordSmith Tools Version 8*. Stroud: Lexical Analysis Software.

Tessier, S. (2012). From Field Notes, to Transcripts, to Tape Recordings: Evolution or Combination? *Int. J. Qual. Methods.* 11 (4), 446–460. doi:10.1177/160940691201100410

Zapf, A., Castell, S., Morawietz, L., and Karch, A. (2016). Measuring Inter-Rater Reliability for Nominal Data - Which Coefficients and Confidence Intervals Are Appropriate? *BMC Med. Res. Methodol.* 16 (93), 93–10. doi:10.1186/s12874-016-0200-9

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Communication

ORIGINAL RESEARCH
published: 02 February 2022
doi: 10.3389/fcomm.2021.797485

Doing the Organization's Work—Transcription for All Practical Governmental Purposes

Alex Holder[1], Christopher Elsey[2], Martina Kolanoski[3], Phillip Brooker[1] and Michael Mair[1]*

[1]Department of Sociology, Social Policy and Criminology, University of Liverpool, Liverpool, United Kingdom, [2]Institute of Allied Health Sciences Research, Faculty of Health and Life Sciences, De Montfort University, Leicester, United Kingdom, [3]Department of Sociology, Goethe-University, Frankfurt, Germany

By comparing two distinct governmental organizations (the US military and NASA) this paper unpacks two main issues. On the one hand, the paper examines the transcripts that are produced as part of work activities in these worksites and what the transcripts reveal about the organizations themselves. Additionally, the paper analyses what the transcripts disclose about the practices involved in their creation and use for practical purposes in these organizations. These organizations have been chosen as transcription forms a routine part of how they operate as worksites. Further, the everyday working environments in both organizations involve complex technological systems, as well as multi-party interactions in which speakers are frequently spatially and visually separated. In order to explicate these practices, the article draws on the transcription methods employed in ethnomethodology and conversation analysis research as a comparative resource. In these approaches audio-video data is transcribed in a fine-grained manner that captures temporal aspects of talk, as well as how speech is delivered. Using these approaches to transcription as an analytical device enables us to investigate when and why transcripts are produced by the US military and NASA in the specific ways that they are, as well as what exactly is being re-presented in the transcripts and thus what was treated as worth transcribing in the interactions they are intended to serve as documents of. By analysing these transcription practices it becomes clear that these organizations create huge amounts of audio-video "data" about their routine activities. One major difference between them is that the US military selectively transcribe this data (usually for the purposes of investigating incidents in which civilians might have been injured), whereas NASA's "transcription machinery" aims to capture as much of their mission-related interactions as is organizationally possible (i.e., within the physical limits and capacities of their radio communications systems). As such the paper adds to our understanding of transcription practices and how this is related to the internal working, accounting and transparency practices within different kinds of organization. The article also examines how

OPEN ACCESS

Edited by:
Helen Fraser,
The University of Melbourne, Australia

Reviewed by:
Ken Tann,
The University of Queensland, Australia
Jenifer Ho,
City University of Hong Kong, Hong Kong SAR, China

****Correspondence:***
Martina Kolanoski
Kolanoski@em.uni-frankfurt.de

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *18 October 2021*
Accepted: *31 December 2021*
Published: *02 February 2022*

Citation:
Holder A, Elsey C, Kolanoski M, Brooker P and Mair M (2022) Doing the Organization's Work—Transcription for All Practical Governmental Purposes. Front. Commun. 6:797485. doi: 10.3389/fcomm.2021.797485

Abbreviations: AR 15-6, Army Regulation (AR) 15-6 investigation; CA, Conversation Analysis; CCs, Capcoms; CDR, Commander (NASA); DOD, Department of Defense (US); EM, Ethnomethodology; IO, Investigating Officer; JTAC, Joint Terminal Attack Controller; LOS, Loss of signal; MQ-1B Predator, Armed, multi-mission, medium-altitude, long-endurance remotely piloted aircraft or drone; NASA, National Aeronautics and Space Administration; mIRC, Military internet relay chat; ODA, Operational Detachment Alpha; PLT, Pilot (NASA); SPT, Science Pilot (NASA).

the original transcripts have been used by researchers (and others) outside of the organizations themselves for alternative purposes.

Keywords: military, NASA (national aeronautics and space administration), ethnomethodology and conversation analysis, transcription, inquiries

1 INTRODUCTION

This article compares two distinct governmental organizations (the US military and NASA) as perspicuous worksites that produce written transcripts as part of their routine work activities and practices. It examines the transcription practices of these organizations with respect to everyday working environments made up of complex, multiple-party interactions in which speakers are frequently spatially and visually separated while engaged in collaborative work. These are technical worksites with multiple communication channels open and in-use to co-ordinate disparate and varied courses of action. How these complexities are re-presented in the transcripts produced provides researchers with a window into the priorities and purposes of transcription, and the "work" transcripts are produced to do in terms of these organizations' tasks. This paper thus examines how transcription fits within the accounting practices of the organizations and how these serve various internal and external purposes. Above all, then, it is interested in how transcripts make the practices they detail "accountable" in Harold Garfinkel's terms (Garfinkel, 1967:1), that is, differently observable and reportable, in their specific contexts of use. By attending to transcription practices in these terms, it becomes possible to draw out lessons about the internal working, accounting and transparency practices within different kinds of organization. With our focus on transcription practices in organizational contexts, this represents a particular kind of "study of work" (Garfinkel, 1986). To aid this comparative exercise the transcription practices routinely used in ethnomethodology and conversation analysis will be deployed as an analytical device to consider decisions made about the level of detail included in any given transcript and the consequences of these decision-making processes.

1.1 Transcription: Theoretical Implications

As with all social scientific research methods and tools, transcription is built upon a set of assumptions about the social settings and practices under investigation. Whether in academia or professional contexts, the work of transcription always requires that a set of decisions be made—explicitly acknowledged or otherwise—in accordance with the goals and purposes of the work, the background understandings which underpin it, and prior knowledge about transcribed interactions. As Bucholtz (2000) argues, these decisions can be grouped into two categories: "interpretive" decisions concerning the content of the transcription and "representational" decisions concerning the form they take. In this regard, written transcripts are never "natural data", neutral imprints of the transcribed interaction, but professional artifacts whose production is ultimately contingent upon organization-specific ways of maintaining and preserving what happened "for the record" for particular practical purposes.

The methodological research literature in this area has suggested that transcription rarely receives the same level of scrutiny and critique applied to research topics or data collection processes, which are frequently the focus of accusations of bias, subjectivity, selectivity, and so on (Davidson, 2009). As Lapadat (2000) frames the issue, transcription is too often treated as holding a "mundane and unproblematic" position in the research process, characterised as being neutral, objective, and concerned solely with re-presenting the spoken words presented in the original recorded data. In the vast majority of cases, little to no effort is made to account for the transcription practices which have been employed, with their reliability usually "taken for granted", a process in which the "contingencies of transcription" are often hidden from view (Davidson, 2009).

For those seeking to open those contingencies up, a key feature of transcription is how original audio/visual data is converted into text for analytical and practical purposes (Ochs, 1979; Duranti, 2006). As Ochs (1979) has demonstrated, the very "format" and re-presentation of audio and/or video-recorded data directly impacts how researchers and readers "interpret" the communication transcribed so that, in her field for instance, talk between adults and children is almost automatically compared to adult-adult interactional practices. Likewise, seemingly trivial omissions of spoken words can considerably shift the readers' understanding of the overall interaction and situation, as Bucholtz shows in a highly consequential analysis of how transcription of a police interview can impact legal proceedings and outcomes (Bucholtz, 2000). However, when taking a practice-based view on transcripts, the work/act of reading and *interpreting* a written transcript is just as important to consider as the work/activities involved in *producing* the transcript. Crucially, both activities are part of the organizational work of accounting for and preserving organizational actions (Lynch and Bogen, 1996). Just as presuppositions and organizational purposes influence the production of the transcripts, they also guide the use of the transcripts, where the transcribed situations are woven into broader narratives. In military-connected investigations these narratives include legal assessments based on assumptions of normal/regular soldierly work and the defining operational context. For NASA, these narratives center on communicating the significance of their missions to domestic public and political audiences as more or less direct stakeholders on whom future funding depends, alongside underlining organizational contributions to scientific and technical knowledge.

Transcription practices are, on the whole, then, opaque. A notable exception in this regard, however, is the discipline of conversation analysis, which, in its perennial focus on transcription techniques and conventions, tends to be more

TABLE 1 | Key features of two incidents involving US Military.

Feature of incident	Baghdad airstrike, aka "collateral murder"	Uruzgan incident
Year	2007	2010
Location	Baghdad, Iraq	Uruzgan, Afghanistan
Casualties	11 civilian casualties (inc. 2 Reuters journalists), 2 children seriously injured	16-23 civilian deaths. Serious injury to men, women and children
Investigations	AR 15-6 investigation of the incident (2007); Investigative work by WikiLeaks (2010)	AR 15-6 investigation of the incident in general and Command Directed Investigation into the conduct of the Predator drone crew (both 2010)
Transcript and original record	WikiLeaks leaked audio-video file (full and edited versions); Transcript produced by WikiLeaks doesn't ascribe speakers	Transcripts of talk from Predator crew cockpit and Kiowa helicopter cockpit produced as part of the original AR 15-6 Investigation
Who produced the transcript?	Not transcribed by US military in 2007	US Military
When was it produced?	Transcribed by WikiLeaks in 2010	2010. Report was complete within a couple of months of the incident, though not publicly available until 2011
When/how was it made public?	Uploaded onto the Collateral Murder webpage with leaked video of incident in 2010	Freedom of information requests by the Los Angeles Times and American Civil Liberties Union. Released to the public in April 2011
Purpose of the transcripts production (if known)	Sub-titling Part of dossier of "evidence" released by WikiLeaks	To provide an account of what happened during the incident. To provide an evidentiary basis for claims made in the AR15-6 reports. The transcripts were also used during interviews with those involved
Redactions present?	N/A	Minor redactions for the purpose of censoring swearing, preserving anonymity of those involved, and obscuring the names of certain technologies and procedures
Author publications	Mair et al. (2016), Elsey et al. (2018)	Holder et al. (2018), Holder (2020)

transparent with regards to the contingencies, challenges and compromises which are an unavoidable feature of transcription (**section 2.2** for full details). Tellingly, for the current analysis, when set against the example of conversation analysis, we find that the US military and NASA also do not explain their transcription practices in any of the documents created. The assumption is that the "work" of explicating the transcription method is not necessary to the organizations' actual work. However, one reason why these worksites represent "perspicuous" settings for comparison is because it is possible to learn lessons from the "complexities" inherent in the production of transcripts in technology-driven, spatially/visually separated, multi-party interactions (Garfinkel, 2002; Davidson, 2009, 47). That is why, after some additional background, we want to unpack what is involved below (**sections 3**, **4**).

2 MATERIALS AND METHODS

2.1 Overview of the Organizational Settings

2.1.1 The US Military

This paper draws together our findings regarding the transcription processes and practices employed by US military personnel following a range of high-profile incidents and accidents that led to the death and injury of civilians during operations involving a combination of ground force and air force units (e.g., planes, helicopters and drones). **Table 1** provides an overview of the key military incidents covered in this paper (listed in chronological order of occurrence).

What unites these tragic incidents for the purposes of our comparison is that they each resulted in formal internal investigations, Army Regulation or AR 15-6s, and because transcripts of both events were produced using the original audio-visual recordings to capture the various parties speaking, though by different parties in each case. Given the loss of civilian life involved, these incidents achieved notoriety when the incidents were eventually made public and thus require careful scrutiny. How transcripts help in that regard is worth some consideration.

2.1.2 National Aeronautics and Space Administration

The National Aeronautics and Space Administration (NASA) is an independent agency of the American government that oversees the US national civilian space program as well as aeronautics and space research activity. From their earliest human-crewed spaceflights, NASA have kept detailed "Air-to-Ground" conversation transcripts covering every available minute of communications throughout human-crewed missions. These transcription practices mobilise a vast pool of human resources in their production—from the crew and ground teams themselves, to technical operators of radio/satellite communications networks across Earth, to teams of transcribers tasked with listening to the recorded conversational data and putting them to paper. This makes it all the more impressive that NASA have been consistently able to produce such transcripts within approximately 1 day of the talk on which they were based. Even with NASA's Skylab program—America's first space station, which was occupied by nine astronauts throughout the early 1970s—it was possible to record and transcribe every available minute of talk occurring when the vehicle was in range of a communications station, amounting to approximately 246,240 min of audio and many thousands of pages of typed transcripts. Though granular detail is difficult to acquire, the annual NASA budget indicates the size of the enterprise, with the mid-Apollo peak of close to $60 billion levelling out to between $18 billion and $25 billion since the 1970s to today (between 0.5 and 1% of all U.S. government public spending) (Planetary Society, 2021). Just why such a huge transcribing machine has been constructed and put to work as part of that effort remains, however, curiously unclear. Ostensibly, the transcripts capture talk for various purposes: to support journalistic reportage of missions, as a kind of telemetry that allows a ground team to learn more about space missions in operation, for its scientific functions (e.g., astronaut crews reporting experimental results) and as a matter of historical preservation. Yet as these transcripts are not drawn on in their fullness for any of these purposes, an exploration of the transcripts themselves is required to learn more about their practical organizational relevance.

2.2 Jefferson Transcription Conventions as Analytical Tools

This study is informed by the principles and practices of ethnomethodology (hereafter EM) and conversation analysis (hereafter CA). These sociological traditions have had an enduring connection with transcription practices and processes as a matter of practical and analytical interest. Given their preoccupation with them, how transcripts fit into these academic enterprises is worth exploring.

In outlining what gave CA its distinctive creative spark, Harvey Sacks (1984: 25-6, our emphasis) suggested CA's novel approach to sociology needed to be understood in the following way:

> [This kind of] research is about conversation only in this *incidental* way: that conversation is something that we can get the actual happenings of on tape and that we can get more or less transcribed; that is, conversation is something to *begin* with.

Yet despite this emphasis on transcripts as something to begin analytical investigations with, for researchers working in these areas the re-production and re-presentation of audio/visual data has, in part, also been a technical issue. While it was Sacks who instigated the focus on conversations as data, it was Gail Jefferson who worked to develop and revise transcription techniques and conventions that reflected the original recordings as closely as possible (Schegloff 1995; Jefferson, 2015). The now established Jeffersonian transcription conventions were designed to capture the temporal or sequential aspects of talk (e.g., overlap, length of pauses, latched utterances) and the delivery of the utterances (e.g., stretched talk/cut-off talk, emphasis/volume, intonation, laughter). For analysts in these fields, transcripts were intended to re-present the original recordings as accurately as possible in order for the resulting analysis to be open to scrutiny by the reader, even if the recording was not available.

In this paper, we use these same transcription techniques as an analytical resource to investigate the transcription practices of a specific set of organizational and institutional settings. Unusually compared with those transcribing verbatim, researchers working under the aegis of EM and/or CA routinely document the transcription procedures and processes applied to any given dataset (audio and/or video). Using these conventions as comparative tools allows us, at least partially, to recover the sense-making and reasoning practices which shaped how transcripts were produced and to what ends in the organizations we examine. A key issue we will take up in this paper is why a specific transcript was created and disseminated in a particular form, something which, we will argue, the transcript itself as an organizational artifact gives us insight into.

Using transcription conventions as an analytical device and method allows researchers to explore the following issues (Davidson, 2009:47):

- What is included in a transcript?
- What is considered pertinent? What is missing (e.g., speaker identifiers and utterance designations)?
- What is deliberately missing or omitted?
- What is/was the purpose/use of the transcript?
- When was it originally produced?
- What is the wider context of the transcripts production and release (e.g., legal/quasi-legal inquiry, inquest, leak)?
- Who is/was the intended audience?
- Is the original recording available? Is the transcript an aid to follow the audio/video or intended to replace it?

These research questions will be applied to the transcription practices in two contrasting work contexts, namely US military investigative procedures and the documentary work of space agencies, in order to provide a window into these settings and to explore issues of record-keeping, self-assessment and accountability. These organizations' transcription practices are compared as they adopt different approaches as to what is transcribed and when. For instance, whereas, NASA operates a "completist" approach to transcription (i.e., with a setup for recording and transcribing all interactions relating to day-to-day space activities within the limits of the physical capacity of their communication setups), the US military audio/video record all missions conducted, but only selectively transcribe when there is a military "incident" requiring formal investigation. This is an important distinction as it speaks to the motives for transcribing and the practical purposes that transcripts are used for. The relevance of this distinction and its implications will be unpacked below.

3 RESULTS—HOW DO THESE ORGANIZATIONS USE TRANSCRIPTS?

In this section of the paper we outline how and why the US military and NASA use the "data" they collect as part of their work. It will also unpack how this data is re-presented, what is transcribed and the transcription practices that are recoverable from transcripts as artifacts alongside their uses within these worksites.

3.1 US Military AR 15-6 Investigations

All airborne military missions and a growing number of ground missions are routinely audio-video recorded. Alongside training and operations reviews, this is done for the purposes of retrospectively collecting evidence in case of the reporting of incidents that occur during operations. As outlined above, such incidents include actions resulting in the injury or death of civilians. However, it is normally only when an incident is declared and a formal internal inquiry is organized that the audio-video recording will be scrutinized for the purposes of producing a transcript. Fundamental differences between the cases we have previously analysed become apparent at this stage. First, not all types of inquiries require transcripts for their investigative work. Depending on the objective, scope and purpose of the investigation, the recorded talk may be treated as more (or less) sufficient on its own. Secondly, the transcripts produced can, at times, be made available *either* as a substitute for the original audio-video data *or* as a supplement to it. To demonstrate the relevance of these issues, we will examine two cases in which transcription was approached in divergent ways. By describing, explicating and scrutinizing the transcription practices used in each case, we can contrast the "work" these practices accomplish. The analysis in this section focuses on the Uruzgan incident as it provides documentary evidence of transcription practices in conjunction with how military investigators read, interpret, and use transcripts as part of their internal accounting practices. The "Collateral Murder" case will be taken up more fully in **sections 3.1.2**, **4.2**.

3.1.1 The Uruzgan Incident

The Uruzgan incident, which took place in Afghanistan in 2010, was the result of a joint US Air Force and US Army operation in which a special forces team, or "Operational Detachment Alpha" (ODA), were tasked with finding and destroying an improvised explosive device factory in a small village in Uruzgan province. Upon arriving in the village, however, the ODA discovered that the village was deserted. Intercepted communications revealed that a Taliban force had been awaiting the arrival of US forces and were preparing to attack the village under cover of darkness. As the situation on the ground became clearer, three vehicles were identified travelling towards the village from the north, and an unmanned MQ-1 Predator drone crew were tasked with uncovering evidence that these vehicles were a hostile force and thus could be engaged in compliance with the rules of engagement. In communication with the ODA's Joint Terminal Attack Controller (JTAC)—the individual responsible for coordinating aircraft from the ground—the Predator crew surveilled the vehicles for well over 3 hours as they drove through the night and early morning. Despite their journey having taken the vehicles *away* from the special forces team for the vast majority of this period, the vehicles were eventually engaged and destroyed by a Kiowa helicopter team at the request of the ODA commander. It did not take long for the reality of the situation to become clear. Within 6 minutes the first call was made that women had been seen nearby the wreckage, and within 25 min the first children were identified. The vehicles had not been carrying a Taliban force. In fact, the passengers were a group of civilians seeking safety in numbers as they drove through a dangerous part of the country. Initial estimates claimed that as many as 23 civilians had been killed in the strike, though subsequent investigations by the US would conclude there had been between fifteen and sixteen civilian casualties. Though investigations into what took place identified numerous shortcomings in the conduct of those involved in the incident, the strike was ultimately deemed to have been compliant with the US rules of engagement and, by extension, the laws of war.

3.1.1.1 The Role of Transcripts in Investigations of the Uruzgan Incident

In this first section of analysis, we will approach the investigative procedures which took place following the Uruzgan incident, identifying the ways in which investigators made use of transcripts in order to: re-construct the finer details of what unfolded; make assessments of the conduct of those involved in the incident; make explanatory claims about the incident's causes; and, finally, contest the adequacy and relevance of other accounts of the incident. The Uruzgan incident is distinctive as a military incident because of the vast body of documentation which surrounds it. There are two publicly available investigations into the incident which not only provide access to the details of the operation itself, but also make visible the US armed forces' mechanisms of self-assessment in response to a major civilian casualty incident. The analysis will exhibit how the three transcripts that were produced following the incident were employed within the two publicly available investigations in order to achieve different conclusions.

The first investigation to be conducted into the Uruzgan incident was an "Army Regulation (AR) 15-6 investigation" (United States Central Command, 2010). AR 15-6 investigations are a type of administrative (as opposed to judicial) investigation conducted internally to the US armed forces concerning the conduct of its personnel. Principally, AR 15-6 investigations are structured as fact-finding procedures, with investigating officers being appointed with the primary role of investigating "the facts/circumstances" surrounding an incident (Department of the Army, 2016: 10). In order to tailor specific investigations to the details of each case, the appointing letter by which a lead investigator is selected includes a series of requests for information. AR 15-6 investigations are intended to serve as what Lynch and Bogen might call the "master narrative" of military incidents, providing a "plain and practical version" of events "that is rapidly and progressively disseminated through a relevant community" (1996: 71). Within this process AR 15-6's represent initial investigations that are routinely conducted where possible mistakes or problems have arisen (see the Collateral Murder analysis in **sections 3.1.2**, **4.2** for another example).

The task of conducting the AR 15-6 investigation into the Uruzgan incident was given to Major General Timothy P. McHale, whose appointing letter stated that he must structure his report as a response to 15 specific requests for information listed from a-t. These questions included:

1) "what were the facts and circumstances of the incident (the 5 Ws: Who, What, When, Where, and Why)?"
2) "was the use of force in accordance with the Rules of Engagement (ROE)?",
3) "what intelligence, if any, did the firing unit receive that may have led them to believe the vans were hostile?" (United States Central Command, 2010: 14–15)

In producing responses to these requests, the appointing letter clearly stated that McHale's findings "must be supported by a preponderance of the evidence" (United States Central Command, 2010: 16). In accumulating evidence during the AR 15-6 investigation, McHale travelled to Afghanistan to conduct interviews with US personnel, victims of the incident, village elders, members of local security groups, and others. He reviewed an extensive array of documents relating to the incident, including personnel reports, battle damage assessments, intelligence reports, and medical records alongside the video footage from aerial assets involved in the operation. Crucially, he also analyzed transcripts of communications that were recorded during the incident. In this way, it can be said that McHale's investigative procedures were demonstrative of concerns similar to those of any individual tasked with producing an account of an historical event. That is, he sought to "use records as sources of data... which permit inferences... about the real world" (Raffel, 1979: 12). Transcripts of recordings produced during the incident were central among McHale's sources of data and, before making assessments of the character of their use in the AR 15-6, it is necessary to introduce the three different transcripts to which McHale refers in the course of his report: the Predator, Kiowa and mIRC Transcripts.

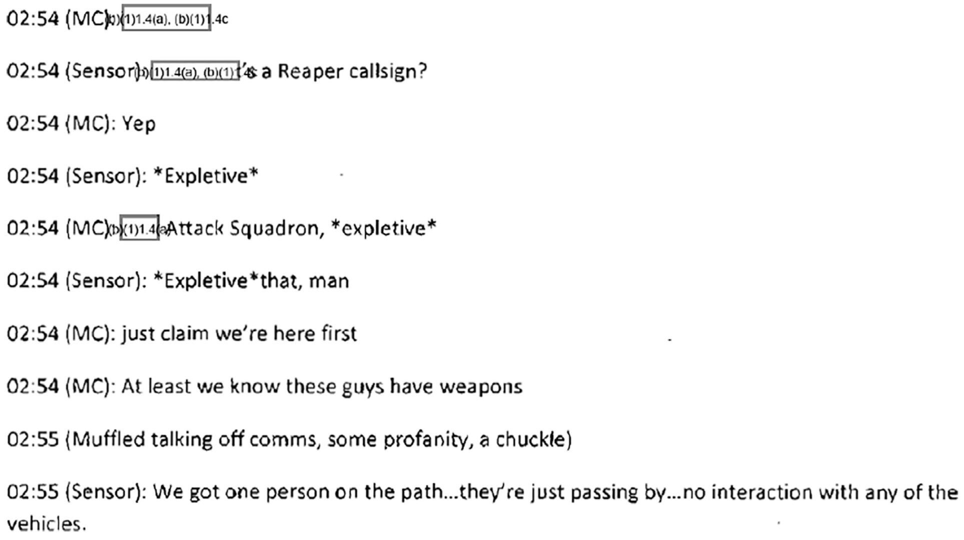
The first transcript, which will be referred to as "the Predator transcript", was produced using recordings from the Predator drone crew's cockpit. This transcript documents over four of hours of talk and includes almost a dozen individuals. That said, as the recordings were made in the Predator crew's cockpit, the bulk of the talk takes place between the three crew members who are co-located in Creech Air Force Base in Nevada. The crew includes the pilot, the mission intelligence coordinator (also known as MC/MIC), and the camera operator (also known as "sensor"). Though the conversations presented in this transcript cover a diversity of topics, they are broadly unified by a shared concern for ensuring that the desired strike on the three vehicles could be conducted in compliance with the rules of engagement. This involved, but was not limited to, efforts to identify weapons onboard the vehicles, efforts to assess the demographics of the vehicles' passengers, and efforts to assess the direction, character, and destination of the vehicles' movements. In terms of format, the Predator transcript is relatively simple—containing little information beyond the utterances themselves, the speakers, and the timing of utterances—though the communications themselves are extremely well preserved as **Figure 1** shows.

The second transcript is "the Kiowa transcript". As above, this document was produced using recordings from the cockpit of one of the Kiowa helicopters which conducted the strike. This document is far more restricted than the Predator transcript in several important ways. For one thing it is far shorter, around six pages, and largely documents the period immediately surrounding the strike itself. There are far fewer speakers, with only two members of the Kiowa helicopter crew, the JTAC, and some unknown individuals being presented in the document. Additionally, the subject matter of the talk presented is far more focused, almost exclusively concerning the work of locating and destroying the three vehicles. In terms of transcription conventions, the Kiowa transcript is far more rudimentary than the Predator transcript, crucially lacking the timing of utterances and—in the publicly available version—the identification of speakers (see WikiLeaks' Collateral Murder transcript in **sections 3.1.2**, **4.2** for comparison). As such, the transcript offers a series of utterances separated by paragraph breaks which do not necessarily signify a change of speaker, as exhibited in **Figure 2**.

Though the Kiowa transcript presents significant analytic challenges in terms of accessing the details of the incident, our present concern lies in the ways in which this transcript was used in McHale's AR 15-6 report, and as such the opacity of its contents constitutes a secondary concern in the context of this paper.

Where the Kiowa transcript is opaque, the final transcript to which McHale refers in the AR 15-6 report is almost entirely inaccessible. That transcript, known as "the mIRC transcript", is constituted by the record of typed chatroom messages sent between the Predator crew and a team of image analysts, known as "screeners", who were reviewing the Predator's video feed in real time from bases in different parts of the US. "mIRC" (or military internet relay chat) communications are text-based messages sent in secure digital chatrooms which are used to distribute information across the US intelligence apparatus. Excepting some small fragments the mIRC transcripts in the AR 15-6 report are entirely classified, and as such, the only means of accessing their contents is through their quotation in the course of the AR 15-6 report. As it happens, McHale frequently makes reference to the contents of the mIRC transcript because, as we shall see, he considers faulty communications between the image analysts and the Predator crew to have played a causal role in the incident.

Though the transcripts which are present in the Uruzgan incident's AR 15-6 investigation each, in different ways, fall short of the standards established by the Jeffersonian transcription conventions, the following sections will identify three ways in which investigators made use of transcripts in order to make, substantiate, and contest claims about what took place.



02:54 (MC)(b)(1)1.4(a), (b)(1)1.4c

02:54 (Sensor)(b)(1)1.4(a), (b)(1)1's a Reaper callsign?

02:54 (MC): Yep

02:54 (Sensor): *Expletive*

02:54 (MC)(b)(1)1.4(a)Attack Squadron, *expletive*

02:54 (Sensor): *Expletive*that, man

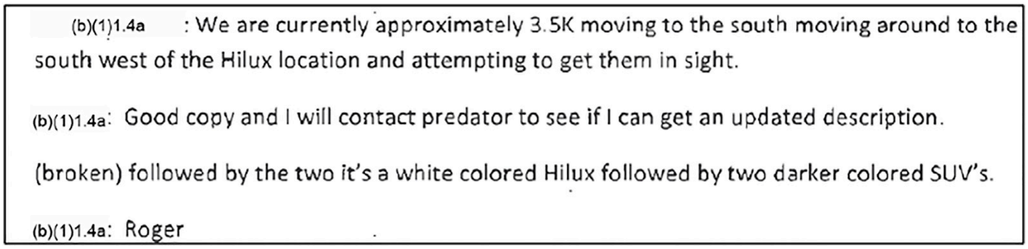
02:54 (MC): just claim we're here first

02:54 (MC): At least we know these guys have weapons

02:55 (Muffled talking off comms, some profanity, a chuckle)

02:55 (Sensor): We got one person on the path…they're just passing by…no interaction with any of the vehicles.

FIGURE 1 | Excerpt from the Predator transcript.



(b)(1)1.4a : We are currently approximately 3.5K moving to the south moving around to the south west of the Hilux location and attempting to get them in sight.

(b)(1)1.4a: Good copy and I will contact predator to see if I can get an updated description.

(broken) followed by the two it's a white colored Hilux followed by two darker colored SUV's.

(b)(1)1.4a: Roger

FIGURE 2 | Excerpt from the Kiowa transcript.

3.1.1.2 Three Uses of the Kiowa, Predator and mIRC Transcripts

The first and most straightforward manner in which transcripts were used in the AR 15-6 investigation was as a means of reconstructing the minutia of the incident. This usage of the transcript is most straightforwardly evident in the response to the request number 2 of the appointing letter, which asked that McHale "describe in specific detail the circumstances of how the incident took place". In response to this question McHale provides something akin to a timeline of events—though not a straightforward one. It does not contain any explicitly normative assessments of the activities it describes and makes extensive reference to various documentary materials which were associated with the incident, including both the Kiowa and the Predator transcripts. In the following excerpt, McHale uses the Kiowa transcript to provide a detailed account of the period during which the strike took place:

> "The third missile struck immediately in front of the middle vehicle, disabling it. After the occupants of the second vehicle exited, the rockets were fired at the people running from the scene referred to as "squirters"; however, the rockets did not hit any of the targets. (Kiowa Radio Traffic, Book 2, Exhibit CC). The females appeared to be waving a scarf or a part of the burqas. (Kiowa Radio Traffic, Book 2. Exhibit CC). The OH-58Ds immediately ceased engagement, and reported the possible presence of females to the JTAC. (Kiowa Radio Traffic, Book 2, Exhibit CC)." (United States Central Command, 2010: 24).

Passages such as this are a testament to the ability of the US military to produce vast quantities of information regarding events which only become significant in retrospect. Though the fact that every word spoken by the Kiowa and Predator crews was recorded is a tiny feat in the context of the US military's colossal data management enterprise (Lindsay, 2020), McHale's ability to reconstruct the moment-by-moment unfolding of the Uruzgan incident remains noteworthy. Where the task of establishing the "facts and circumstances of the incident" is concerned, the transcripts provide McHale with a concrete resource by which "what happened" can be well established, and the AR 15-6's status as a "master narrative" can be secured. As we shall see, however, in those parts of the report where McHale proceeds beyond descriptive accounts of what took place, and into causal assessments of *why* the Uruzgan incident happened, allowing the transcript to "speak for itself" is no longer sufficient. As such, the second relevant reading of the transcripts in the AR 15-6 report was as an evidentiary basis by which causal claims could be substantiated.

Though McHale's AR 15-6 report identified four major causes for the incident, our focus here will be upon his assertion that "predator crew actions" played a critical role in the incident's tragic outcome. The following excerpt is provided in response to the appointing letter's request that McHale establish "the facts and circumstances surrounding the incident (5 Ws)":

> "The predator crew made or changed key assessments to the ODA (commander) that influenced the decision to destroy the vehicles. The Predator crew has neither the training nor the tactical expertise to make these assessments. First, at 0517D, the Predator crew described the actions of the passengers of the vehicles as "tactical maneuvering". At that point, the screeners located in Hurlburt field described the movement as adult males, standing or sitting [(redacted) Log, book 5, Exhibit X, page 2]. At the time of the strike "tactical maneuver" is listed by the ODA Joint Tactical Air Controller (JTAC), as one of the elements making the vehicle a proper target [(Redacted) Logbook 5, Exhibit T, page 57" (United States Central Command, 2010: 21-22)."

In this section, the citation of "[(redacted) Log, Book 5, Exhibit X, page 2]" is a reference to the mIRC transcript. As such, though it is not explicitly stated, the communications at 0517D took the form of typed messages between the Predator crew and the Florida-based image analysts[1]. It should be immediately clear that this passage is of a different character to our previous excerpt. Most notably, the assertion of a causal relation between the Predator crew's assessments of the vehicles' movements and the commander's decision to authorize the strike is rooted in McHale's own interpretation of events. In line with the appointing letter's request that McHale's assertion be based upon a "preponderance of the evidence", McHale seeks to use the mIRC transcript to substantiate that claim as this section proceeds.

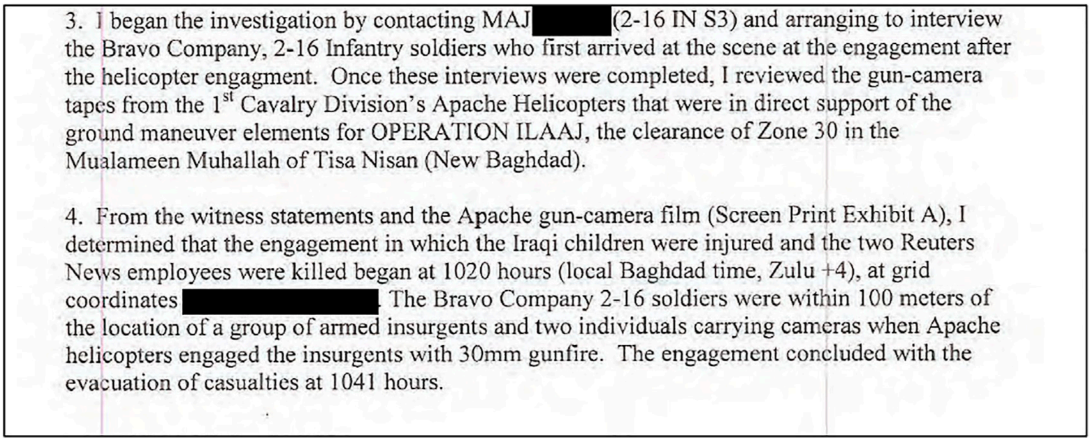
As a first step towards doing so, McHale sets up a contrast between the Predator crew's assessment that the vehicles were engaged in "tactical maneuvering" and the image analysts' apparently contradictory assessment that there were "adult males, standing or sitting". In establishing the incongruity between these conflicting assessments, McHale presents tactical maneuvering as a contestable description that the Predator crew put forward without the requisite training or tactical expertise. As McHale proceeds, he proposes a link between the Predator crew's use of the term and its appearance in the JTAC's written justification for the strike. In this way, McHale not only makes use of the transcript as a mechanism by which assessments of the Predator crew's inadequate conduct could be made, but also as a means by which a causal relationship between the Predator crew's actions and the incident's outcome could be empirically established. As we shall see, however, assessments which are secured by reference to the record of what took place ultimately open to contestation, and McHale's own analysis in this regard would be open to criticism from elsewhere.

Following the completion of the AR 15-6 investigation, McHale recommended that a Command Directed Investigation be undertaken to further examine the role of the Predator crew in the incident. This was undertaken by Brigadier General Robert P. Otto. At that time Otto was the Director of Surveillance and Reconnaissance in the US Air Force and, in Otto's own words, the investigation took a "clean sheet of paper approach" to the Predator crew's involvement in the operation (Department of the Air Force, 2010: 34). Despite McHale's initial findings, Otto's commentary on the incident resulted in a different assessment of the adequacy and operational significance of the Predator crew's actions. One particularly notable example concerns McHale's criticism of the Predator crew's use of the term 'tactical maneuvering'. Otto writes:

> "The ground force commander cited "tactical maneuvering with (intercepted communications) chatter as one of the reasons he felt there was an imminent threat ... Tactical maneuvering was identified twice before Kirk 97 began tracking the vehicles. Although not specifically trained to identify tactical maneuvering, Kirk 97 twice assessed it early in the incident sequence. However, for 3 hours after Kirk 97's last mention of tactical maneuvering, the (commander) got frequent reports on convoy composition, disposition, and general posture (...) I conclude that Kirk 97's improper assessment of tactical maneuvering was only a minor factor in the final declaration". (Department of the Air Force, 2010: 36)

In this passage, McHale's causal claim regarding the significance of the Predator crew's reference to tactical maneuvering is rejected, initially on the grounds that the Predator crew were not responsible for introducing the concept. As Otto observes, "Tactical maneuvering was identified twice before Kirk 97 began tracking the vehicles" (ibid.). Interestingly, this counter-analysis charges McHale with having straightforwardly misread the record of what took place. Recall that McHale's analysis of the term tactical maneuvering cited the mIRC transcript as evidence of the Predator crew's shortcomings without making any reference to the Predator transcript. As Otto observes, analysis of the Predator transcript reveals that the first reference to tactical maneuvering took place at 0,503, where the term was used by the JTAC himself. With this being the case, McHale's causal claim regarding the

[1] For clarity, it is worth noting that the Predator crew made a radio call to the JTAC identifying the vehicles' tactical maneuvering at 0,512, just a couple of minutes before the mIRC message to which McHale refers was sent.



3. I began the investigation by contacting MAJ [redacted] (2-16 IN S3) and arranging to interview the Bravo Company, 2-16 Infantry soldiers who first arrived at the scene at the engagement after the helicopter engagment. Once these interviews were completed, I reviewed the gun-camera tapes from the 1^{st} Cavalry Division's Apache Helicopters that were in direct support of the ground maneuver elements for OPERATION ILAAJ, the clearance of Zone 30 in the Mualameen Muhallah of Tisa Nisan (New Baghdad).

4. From the witness statements and the Apache gun-camera film (Screen Print Exhibit A), I determined that the engagement in which the Iraqi children were injured and the two Reuters News employees were killed began at 1020 hours (local Baghdad time, Zulu +4), at grid coordinates [redacted]. The Bravo Company 2-16 soldiers were within 100 meters of the location of a group of armed insurgents and two individuals carrying cameras when Apache helicopters engaged the insurgents with 30mm gunfire. The engagement concluded with the evacuation of casualties at 1041 hours.

FIGURE 3 | Redacted extract from the US military AR 15-16 investigation (Iraq, July 12, 2007).

Predator crew's characterization of the vehicles' movements as tactical maneuvering is problematic and significantly weakened.

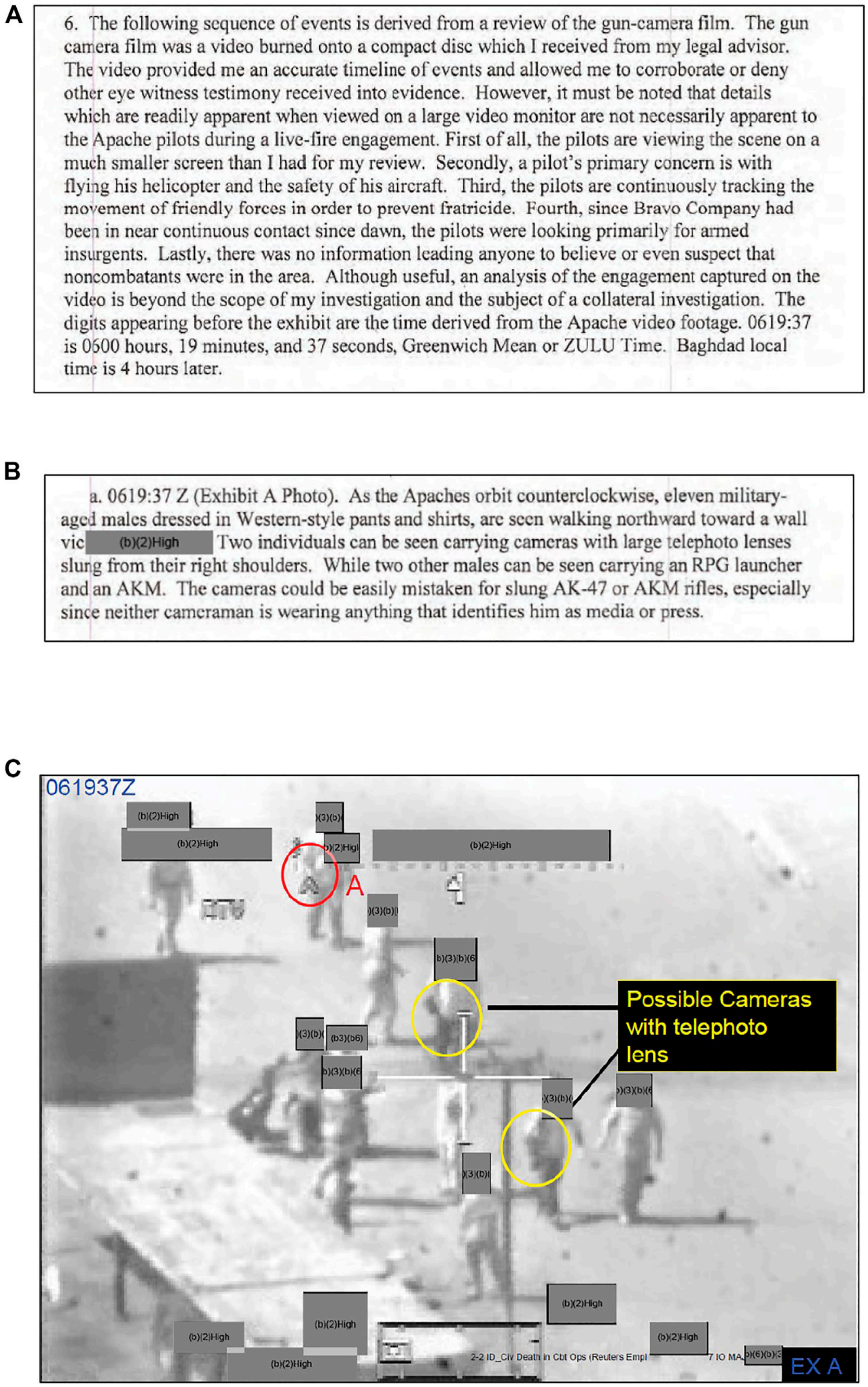
This is not the end of Otto's criticism, however. As the passage goes on, Otto also rejects the McHale account as having overstated the operational relevance of the Predator crew's reference to tactical maneuvering. Though Otto doesn't cite the Predator transcript explicitly, he notes that in the hours following the final use of the term the crew routinely provided detailed accounts of the "composition, disposition, and general posture" (ibid.) of the vehicles. The proposal here is that by the time the strike took place, so much had been said about the vehicles and their movements that the reference to tactical maneuvering hours previously was unlikely to have been a crucial element in the strike's justification. Again, Otto's criticism is rooted in an accusation that McHale's account misinterprets what the transcript reveals about the Uruzgan incident. On this occasion, it was not a misreading which led to error, rather it was a failure to appreciate the ways in which transcripts warp the chronology of events. There is a lesson to be learned here: though transcripts effectively preserve the details of talk, they do not provide instructions for assessing their *relevance*. The relevance of particular utterances within broader courses of action depends upon a considerable amount of contextualizing information, as well as the place of that utterance within an on-going sequence of talk. Of course, Otto does not articulate McHale's error in these terms—he has no reason to—but his critical engagement with McHale's analysis has clear corollaries with conversation analytic considerations when working with transcripts.

3.1.2 Investigations Without Transcripts: The Collateral Murder Case

Not all military investigations seek to use transcripts as the primary means by which the details of what took place can be accessed. The "Collateral Murder" case—so named following the infamous Wikileaks publication of video footage from the incident under that name—took place in 2007 and involved the killing of 11 civilians, of whom two were Reuters journalists, following a US strike conducted by a team of two Apache helicopters (Reuters Staff, 2007; Rubin, 2007). It took 3 years for the incident to make its way to the public eye. On April 5th, 2010, Wikileaks published a 39-min video depicting the gunsight footage from one of the Apache helicopters involved in the strike. As with the Uruzgan incident, the collateral murder case had been the subject of an AR 15-6 investigation soon after the incident, but the investigations resulting report was not made publicly available until the day the WikiLeaks video was published. Once again, the investigation declared that the strike had taken place in compliance with the laws of war, though it was not nearly so critical of the conduct of those involved as McHale's account of the Uruzgan incident had been.

Based on the completed report, we are able to ascertain what evidence was gathered in support of the investigation (Investigating Officer 2nd Brigade Combat Team 2nd Infantry Division, 2007). Fundamentally, the Investigating Officer (IO) drew on two main forms of evidence: witness testimony from the US personnel involved and the Apache video footage, which was utilized by the IO to produce a *timeline* of what happened on the day (**Figure 3** below). No transcript was produced in support of the investigation. As such, the report displays the ways in which visual materials were used in combination with after the fact interviews to establish how the incident had unfolded.

Instead of making use of a transcript to reconstruct the details of the incident, the IO decided that the combination of timestamps (actual time, taken from the video recording), still images taken from the video (displayed as exhibits in the appendices with IO annotations) and visual descriptions of the action taken from the video could be compiled into a "sequence of events" or timeline covering those actions deemed to constitute the incident. This offers a neat contrast with Sacks' understanding of the analytic value of transcription. For Sacks, in depth transcriptions allowed interaction to be closely examined, forming as "a "good enough" record of what happened" in real-time interactions (Sacks 1984, 25-6). Transcription would become a consistent feature of CA but



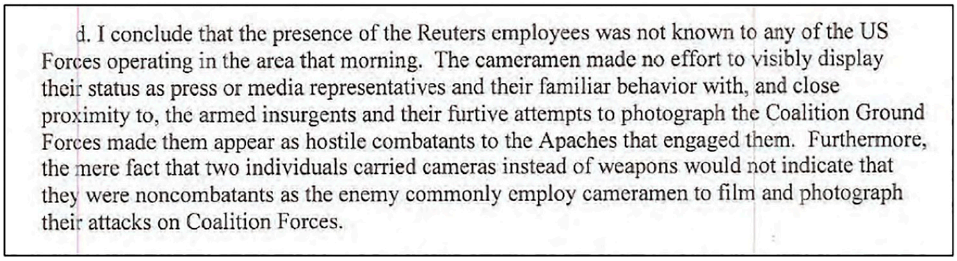
6. The following sequence of events is derived from a review of the gun-camera film. The gun camera film was a video burned onto a compact disc which I received from my legal advisor. The video provided me an accurate timeline of events and allowed me to corroborate or deny other eye witness testimony received into evidence. However, it must be noted that details which are readily apparent when viewed on a large video monitor are not necessarily apparent to the Apache pilots during a live-fire engagement. First of all, the pilots are viewing the scene on a much smaller screen than I had for my review. Secondly, a pilot's primary concern is with flying his helicopter and the safety of his aircraft. Third, the pilots are continuously tracking the movement of friendly forces in order to prevent fratricide. Fourth, since Bravo Company had been in near continuous contact since dawn, the pilots were looking primarily for armed insurgents. Lastly, there was no information leading anyone to believe or even suspect that noncombatants were in the area. Although useful, an analysis of the engagement captured on the video is beyond the scope of my investigation and the subject of a collateral investigation. The digits appearing before the exhibit are the time derived from the Apache video footage. 0619:37 is 0600 hours, 19 minutes, and 37 seconds, Greenwich Mean or ZULU Time. Baghdad local time is 4 hours later.

a. 0619:37 Z (Exhibit A Photo). As the Apaches orbit counterclockwise, eleven military-aged males dressed in Western-style pants and shirts, are seen walking northward toward a wall vic (b)(2)High Two individuals can be seen carrying cameras with large telephoto lenses slung from their right shoulders. While two other males can be seen carrying an RPG launcher and an AKM. The cameras could be easily mistaken for slung AK-47 or AKM rifles, especially since neither cameraman is wearing anything that identifies him as media or press.

FIGURE 4 | (A) and **(B)** Paired extracts from the US military AR 15-16 investigation (Iraq, July 12, 2007). **(C)** Exhibit A Photo' from the US military AR 15-16 investigation (Iraq, 12 July 2007).

not, as we see here, a consistent feature of US military investigations which have various other ways of arriving at a "good enough record" for their own analytic purposes.

An example of the alternative "pairing" of evidence and reporting is provided in the extracts from the official report (**Figure 4**).



d. I conclude that the presence of the Reuters employees was not known to any of the US Forces operating in the area that morning. The cameramen made no effort to visibly display their status as press or media representatives and their familiar behavior with, and close proximity to, the armed insurgents and their furtive attempts to photograph the Coalition Ground Forces made them appear as hostile combatants to the Apaches that engaged them. Furthermore, the mere fact that two individuals carried cameras instead of weapons would not indicate that they were noncombatants as the enemy commonly employ cameramen to film and photograph their attacks on Coalition Forces.

FIGURE 5 | Conclusions—Extract from the US military AR 15-16 investigation (Iraq, July 12, 2007).

The report itself was fairly brief (amounting to 43 pages), and in its course the IO was able to identify the primary features of the incident, all without a transcript. Using the kinds of materials outlined above, the IO was able to provide an adequate account of the mission objectives, who was killed and their status (as either civilian or combatants), and how/why the Reuters journalists were misidentified (i.e., their large cameras could/were reasonably mistaken for RPGs, there were no known journalists in the area, etc.). Within the understood scope of the AR 15-6's administrative parameters and functions, a transcript was not, therefore, required.

The evidence from the witness testimony and the video recording was deemed sufficient to ascertain that the troops had come under fire from a "company of armed insurgents" the Reuters journalists were said to be moving around with. The identities of the journalists were later verified in the report (via the presence of their cameras, the photographic evidence on the memory cards, and the recovered "press identification badges from the bodies"). Despite this, the conduct of the US military personnel (Apache crews and ground forces) was given the all-clear by the report (see **Figure 5** below):

Thus, whilst both the Uruzgan incident and the collateral murder case were deemed legal by their respective investigations, their conclusions differ significantly insofar as the AR 15-6 for the collateral murder case does not identify shortcomings in the conduct of the US personnel involved. In our analysis of the AR 15-6 investigation into the Uruzgan incident, we have clearly demonstrated that McHale's (and subsequently Otto's) assessments of the incident were, to a large extent, pre-occupied with the adequacy of the conduct of those involved. We would here propose that the documentary materials used to reconstruct the facts and circumstances of the incident are reflective of this pre-occupation—with transcripts of talk being treated as a primary means of reconstructing what had taken place in one case but deemed to be superfluous in the latter case.

Even in relation to one of the most seemingly egregious aspects of the incident, the injuries to the two young children, the report concluded that their presence could not have been expected, anticipated or known as they were not known to the Apache crews and could not be identified on the video—the Apache's means of accessing the scene below them—prior to contact. Beyond a short, redacted set of recommendations, these conclusions meant the incident was not deemed sufficiently troublesome to require a more formal legal investigation of the kind that would have generated a transcript.

Having presented two contrasting cases of the use of transcripts with US military AR 15-6 investigations, we will now turn to our other institutional setting, namely NASA's Skylab Program.

3.2 NASA's Skylab Program

As noted previously in **section 2.1.2**, the transcription machinery of NASA that was deployed in the service of their Skylab program forms an extraordinarily large collective effort to meet the needs of NASA's first long-duration missions. NASA's Skylab space station was launched in May 1973, and was occupied on a near-continuous basis for 171 days until February 1974, producing (amongst its scientific achievements) 246,240 min of audio, all of which was transcribed and archived as a legacy of the program. Elaborating the justification for and purpose of such vast collaborative labor inevitably involves tracing NASA's transcription practices back to Skylab's predecessors; NASA's major human spaceflight programs Mercury (1958–1963), Gemini (1961–1966) and Apollo (1960–1972).

The Mercury program was NASA's early platform for researching the initial possibility (technical and biological) of human-crewed orbital spaceflight, hosting a single pilot for missions lasting from just over 15 min to approximately 18 h. Once it was proven that a vessel could be successfully piloted into low Earth orbit and sustain human life there, the Gemini program extended NASA's reach by building craft for two-person crews that could be used to develop human spaceflight capabilities further—for instance, Gemini oversaw the first EVA (extra-vehicular activity, i.e., a "spacewalk" outside of a craft) by an American, the first successful rendezvous and docking between two spacecraft, and testing if human bodies could survive long duration zero gravity conditions for up to 14 days. Building on the successes of Gemini, Apollo's goal—famously—was to transport three-person crews to the Moon, orbit and land on the Moon, undertake various EVA tasks and return safely to Earth, and Apollo mission durations ranged from 6 to just over 12 days. For all three programs—due to the relatively short duration of individual missions and the experimental nature of the missions themselves—not only were spaceflight technical systems tested, so were auxiliary concerns such as food and water provision, ease of use of equipment, various measures of crew health and wellbeing, etc., and all possible communications were tape-recorded and transcribed[2]. In this sense, while live communications with an astronaut crew flying a mission were vital for monitoring health, vehicles and performance, the transcriptions of talk between astronauts and mission control has a different function—they stand as a more or less full record of significant historical moments for journalistic purposes, but also a record of source data for the various experiments that were built into these missions.

The Skylab transcription machine of the 1970s might then be seen as a direct continuation of a system that had already worked to great effect for NASA since the late 1950s. Despite the obvious differences between Skylab and its predecessor programs—far longer duration missions (up to 84 days) and a different substantive focus (laboratory-based scientific experimentation)—Skylab sought to implement a tried-and-tested transcription machinery without questioning its need or purpose in this markedly new context. There are seemingly two interrelated reasons for this: first, NASA's achievements were iteratively built on risk aversion (as the adage goes, "if it ain't broke, don't fix it") (Newell, 1980; Hitt et al., 2008), and second, that in the scientific terms under which Skylab was designed and managed (Compton and Benson, 1983; Hitt et al., 2008) the matter becomes one of merely scaling up a variable (e.g., mission duration) as a technically-achievable and predictable phenomenon rather than being seen as an opportunity or need to revisit the social organization of NASA itself. To some degree, producing full supplementary transcriptions did serve some purposes for Skylab, where mission activities aligned with those of earlier programs—for instance, in scientific work where crews could verbally report such experimental metadata as camera settings which could then be transcribed and linked to actual frames of film when a mission had returned its scientific cache to Earth upon re-entry, or where various daily medical measurements could be read down verbally from crew to ground to be transcribed and passed along to the flight surgeon teams. For these kinds of activities, having a timestamped transcript to recover such details post-mission was useful. However, given the longer duration of Skylab missions generally, and the intention for those missions to help routinise the notion of "Living and Working in Space" (cf. Brooker, forthcoming; Froehlich, 1971; Compton and Benson, 1983), much was also transcribed that seemingly serves very little purpose—for instance, regularly-occurring humdrum procedural matters such as morning wake-up calls, and calls with no defined objective other than keeping a line open between ground and crew.

It is perhaps useful at this point to introduce excerpts of transcriptions that illuminate the ends to which such an enormous collaborative transcribing effort was put, and to provide further detail on just what is recorded and how. The transcripts that follow are selected to represent relevant aspects of the Skylab 4 mission specifically [as this forms the basis of ongoing research covering various aspects of Skylab (Brooker and Sharrock, forthcoming)], reflecting 1) a moment of scientific data capture (**Figure 6**), and 2) a moment where nothing especially significant happens (see **Figure 7**)[3]. Timestamps are given in the format "Day-of-Year: Hour: Minute: Second", and speakers are denoted by their role profile: CDR is Commander Gerald Carr, PLT is Pilot William Pogue, SPT is Science-Pilot Ed Gibson, and the CCs are CapComs Henry "Hank" Hartsfield Jr and Franklin Story Musgrave[4].

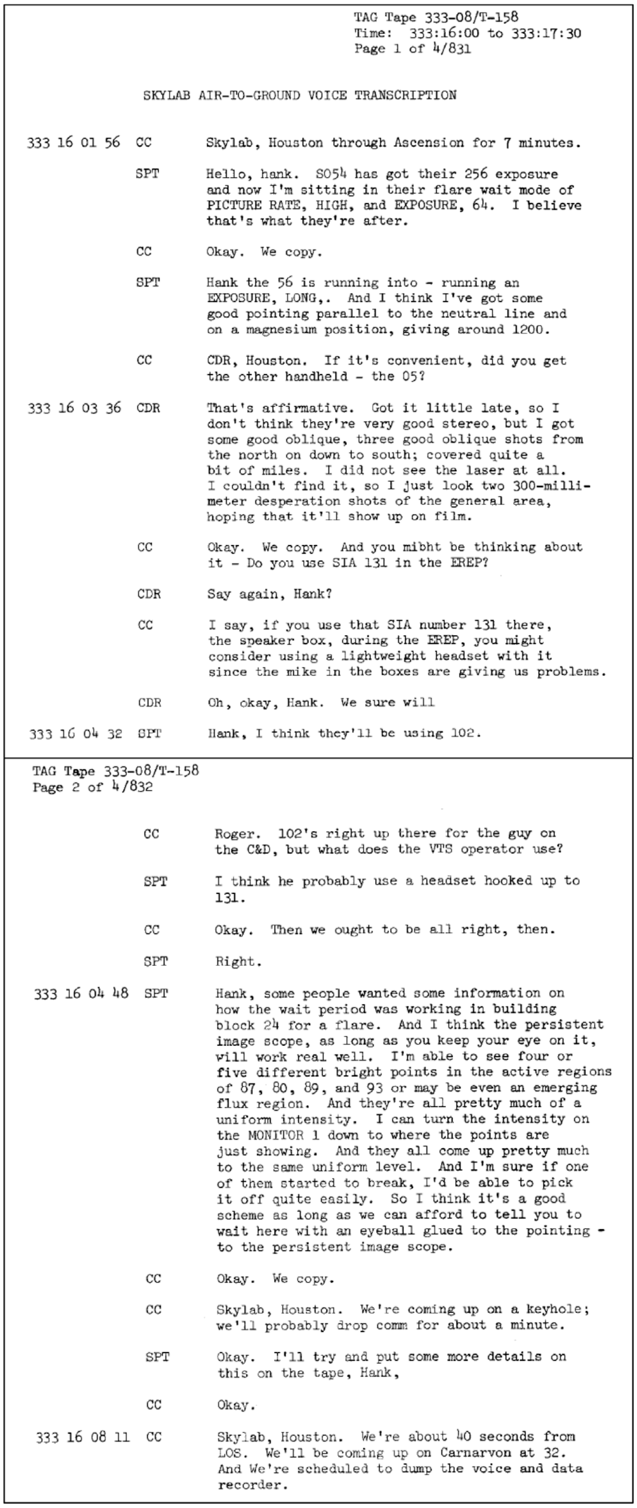
Figure 6 commences with a call at 333 16 01 56 with CC announcing their presence, which communications relay they are transmitting through, and the time they will be available before the next loss of signal (LOS) ("Skylab, Houston through Ascension for 7 min"), and closes at 333 16 08 11 with CC announcing the imminent loss of signal and timings for the next call. In the intervening 7 min, SPT and CDR take turns at reporting the progress of their current, recent and future experimental work in what proves to be a tightly-packed call with several features to attend to here. Immediately, SPT takes an opportunity to report on an ongoing experiment (e.g., "Hello, hank. S054 has got their 256 exposure and now I'm sitting in their flare wait mode of PICTURE RATE, HIGH, and EXPOSURE, 64. I believe that's what they're [the scientists in charge of experiment S054] after."). This report delivers key salient metadata—the experiment designation (S054), and various details pertaining to camera settings. In the transcript, these salient details are all the more visible for being typed out in all-caps; strategically a useful visual marker for science teams on the ground seeking to identify *their* metadata from transcripts replete with all manner of information. That it is SPT delivering this information is also important, as it is he who was designated to perform this particular experiment on this particular day (another clue for transcript readers seeking to gather details of a particular experiment post-hoc)—this provides for specific timestamps to be catalogued by ground-based science teams according to their relevance to any given scientific task.

CC then (333 16 03 36) requests a report from CDR on a recently-completed photography activity, and CDR and CC are able to both talk about the live continuation of that activity (e.g., instruction to use a particular headset in future as opposed to malfunctioning microphones) as well as record, for the benefit

[2] It was not necessarily the case that astronaut crews were in contact with ground control for every minute of a Mercury, Gemini or Apollo mission, owing to the nature of the radio communications used at the time and the network of relay stations that NASA could use to facilitate transmissions. But missions could be planned to maximise time in communication range even for Apollo where astronauts flew almost 250,000 miles away from Earth, meaning that acquisitions and losses of signal were a known and predictable occurrence around which interactions between astronauts and ground control could be organized, even in emergency scenarios (cf. Brooker and Sharrock, forthcoming).

[3] As it is impossible to pick out a "typical" transcript from the vast expanse of Skylab's timespan and range of tasks, these transcripts have been more or less arbitrarily selected. However, they will nonetheless illuminate NASA's transcription machine in different ways and are as such useful points of reference.

[4] The CapCom (Capsule Communicator) is a ground-based role normally taken by a member of the astronaut corps, such that mission control have a single designated contact with an astronaut crew, through which communications can be relayed (though the CapCom role rotates through personnel in 8-h shifts).

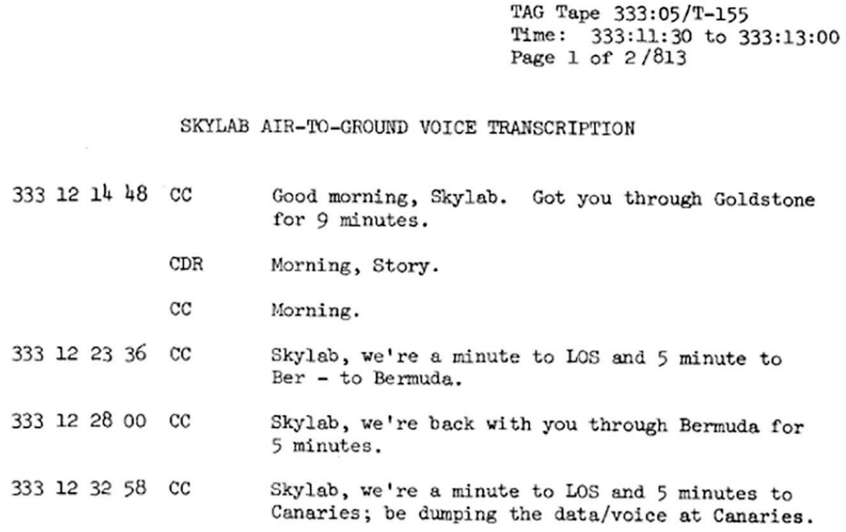


```
                                        TAG Tape 333-08/T-158
                                        Time:  333:16:00 to 333:17:30
                                        Page 1 of 4/831

              SKYLAB AIR-TO-GROUND VOICE TRANSCRIPTION

333 16 01 56  CC       Skylab, Houston through Ascension for 7 minutes.

              SPT      Hello, hank.  S054 has got their 256 exposure
                       and now I'm sitting in their flare wait mode of
                       PICTURE RATE, HIGH, and EXPOSURE, 64.  I believe
                       that's what they're after.

              CC       Okay.  We copy.

              SPT      Hank the 56 is running into - running an
                       EXPOSURE, LONG,.  And I think I've got some
                       good pointing parallel to the neutral line and
                       on a magnesium position, giving around 1200.

              CC       CDR, Houston.  If it's convenient, did you get
                       the other handheld - the 05?

333 16 03 36  CDR      That's affirmative.  Got it little late, so I
                       don't think they're very good stereo, but I got
                       some good oblique, three good oblique shots from
                       the north on down to south; covered quite a
                       bit of miles.  I did not see the laser at all.
                       I couldn't find it, so I just look two 300-milli-
                       meter desperation shots of the general area,
                       hoping that it'll show up on film.

              CC       Okay.  We copy.  And you mibht be thinking about
                       it - Do you use SIA 131 in the EREP?

              CDR      Say again, Hank?

              CC       I say, if you use that SIA number 131 there,
                       the speaker box, during the EREP, you might
                       consider using a lightweight headset with it
                       since the mike in the boxes are giving us problems.

              CDR      Oh, okay, Hank.  We sure will

333 16 04 32  SPT      Hank, I think they'll be using 102.
```

```
TAG Tape 333-08/T-158
Page 2 of 4/832

              CC       Roger.  102's right up there for the guy on
                       the C&D, but what does the VTS operator use?

              SPT      I think he probably use a headset hooked up to
                       131.

              CC       Okay.  Then we ought to be all right, then.

              SPT      Right.

333 16 04 48  SPT      Hank, some people wanted some information on
                       how the wait period was working in building
                       block 24 for a flare.  And I think the persistent
                       image scope, as long as you keep your eye on it,
                       will work real well.  I'm able to see four or
                       five different bright points in the active regions
                       of 87, 80, 89, and 93 or may be even an emerging
                       flux region.  And they're all pretty much of a
                       uniform intensity.  I can turn the intensity on
                       the MONITOR 1 down to where the points are
                       just showing.  And they all come up pretty much
                       to the same uniform level.  And I'm sure if one
                       of them started to break, I'd be able to pick
                       it off quite easily.  So I think it's a good
                       scheme as long as we can afford to tell you to
                       wait here with an eyeball glued to the pointing -
                       to the persistent image scope.

              CC       Okay.  We copy.

              CC       Skylab, Houston.  We're coming up on a keyhole;
                       we'll probably drop comm for about a minute.

              SPT      Okay.  I'll try and put some more details on
                       this on the tape, Hank,

              CC       Okay.

333 16 08 11  CC       Skylab, Houston.  We're about 40 seconds from
                       LOS.  We'll be coming up on Carnarvon at 32.
                       And We're scheduled to dump the voice and data
                       recorder.
```

FIGURE 6 | NASA excerpt 1—scientific reporting on skylab.

of the eventual transcript, CDR's evaluation of the performance of that activity to complement what will eventually be seen on film (e.g. "I did not see the laser at all. I couldn't find it, so I just took two 300-mm desperation shots on the general area, hoping that it'll show up on film."). In this call, SPT also proposes a suggestion on undertaking a continuation of his current



TAG Tape 333:05/T-155
Time: 333:11:30 to 333:13:00
Page 1 of 2/813

SKYLAB AIR-TO-GROUND VOICE TRANSCRIPTION

333 12 14 48	CC	Good morning, Skylab. Got you through Goldstone for 9 minutes.
	CDR	Morning, Story.
	CC	Morning.
333 12 23 36	CC	Skylab, we're a minute to LOS and 5 minute to Ber - to Bermuda.
333 12 28 00	CC	Skylab, we're back with you through Bermuda for 5 minutes.
333 12 32 58	CC	Skylab, we're a minute to LOS and 5 minutes to Canaries; be dumping the data/voice at Canaries.

FIGURE 7 | NASA excerpt 2—a "mundane" call to skylab.

experiment (333 16 04 48)—again, this serves a live function in terms of providing details that CC can pass on to relevant ground teams (mission control and scientific investigators) for consideration, but also records specific parameters that SPT intends to use in that experiment for the transcript (e.g. "I think the persistent image scope, as long as you keep your eye on it, will work real well. I'm able to see four or five different bright points in the active regions of 87, 80, 89 and 92 or may be even an emerging flux region."). On this latter reporting, SPT also notes an intention to "put some more details on this on the tape (which records "offline" notes that can be reviewed and transcribed at a later point)", flagging for the transcript that a future section of the transcribed tape recordings—another set of volumes capturing the talk of astronauts, though not talk that is held on the air-to-ground channel—may contain relevant details for the scientific teams on the ground.

At moments such as these, where scientific work is in-train and there is much to be reported, the transcripts reveal strategies for making that work visible post-hoc, and in doing so, for supporting the analysis of the data that astronauts are gathering through flagging the location and type of metadata that it is known will be transcribed. At other moments however, the between-times of experiments, or during longer-running experiments where little changes minute-by-minute, there may be less of a defined use for the transcripts, as we will see in the following excerpt **Figure 7**.

This excerpt, in fact, features two successive calls with seemingly little content which might be used to elaborate the practical work the astronauts are undertaking at the time of the call. CC announces the opening of a call (333 12 14 48), the transmission relay in-use, and the expected duration of the signal ("Good morning, Skylab. Got you through Goldstone for 9 min"). Good-mornings are exchanged between CDR and CC, but the call is brought to end 9 min later with no other substantive content other than an announcement of loss of signal and a pointer towards when and where the next call will take place (CC at 333 12 23 36: "Skylab, we're a minute to LOS and 5 min to Ber—Bermuda."). The next call (333 12 28 00) opens similarly—CC: "Skylab, we're back with you through Bermuda for 5 min". In contrast to the previous call however, the astronauts remain silent and the call closes shortly thereafter with a similar announcement of the imminent loss of signal from CC, plus the location of the relay for the next call and a note that the next call will begin with the ground team retrieving audio data to be fed into the transcription machine, without the astronaut crews having spoken at all (333 12 32 58: "Skylab, we're a minute to LOS and 5 min to Canaries; be dumping the data/voice at Canaries").

Despite the seeming inaction on display here, the transcripts might still be used to elicit an insight into various features of the ways in which NASA is organized. For instance, we learn that transcribing activity is comprehensive rather than selective—it is applied even when nothing overtly interesting is taking place, to keep the fullest record possible. Communication lines are accountably opened and closed in the eventuality that there might be things worth recording, even if that isn't always the case. There are procedural regularities to conversations between ground and astronauts that bookend periods of communications (e.g., a sign-on and a sign-off), which do not necessarily operate according to the general conventions of conversation (e.g., it would be a noticeable breach for a person not to respond to a greeting on the telephone, but not here) (Schegloff, 1968). However, it is worth noting that what we might learn from these episodes is of no consequence to NASA or their scientific partners—for them, the purpose of transcribing these episodes can only be to ensure their vast transcription machine continues rolling; here, producing an extraordinarily elaborate icing on what could at times be the blandest of cakes.

4 POST-HOC USES OF TRANSCRIPTS

This section will explore the ways in which materials we have introduced up have been put to use for different ends post-hoc by other institutions with differing sets of interests beginning with the NASA case first.

4.1 Post Hoc Uses of NASA Transcripts

Post-mission, various researchers have attempted to tap into the insights contained in Skylab's volumes of transcripts, particularly as part of computationally-oriented studies that process the data captured therein (scientific results and talk alike) to elaborate on the work of doing astronautics and propose algorithmic methods for organising that work more efficiently. Kurtzman et al. (1986), for instance, draw on astronaut-recorded data to propose a computer system—MFIVE—for absolving the need of having insights recorded in transcript at all by mechanising the processes of space station workload planning and inventory management. The addition of a computerised organisational tool, which would record and process information about workload planning and inventory management issues, is envisaged as follows:

> "The utility and autonomy of space station operations could be greatly enhanced by the incorporation of computer systems utilizing expert decision making capabilities and a relational database. An expert decision making capability will capture the expertise of many experts on various aspects of space station operations for subsequent use by nonexperts (i.e., spacecraft crewmembers)." (Kurtzman et al., 1986: 2)

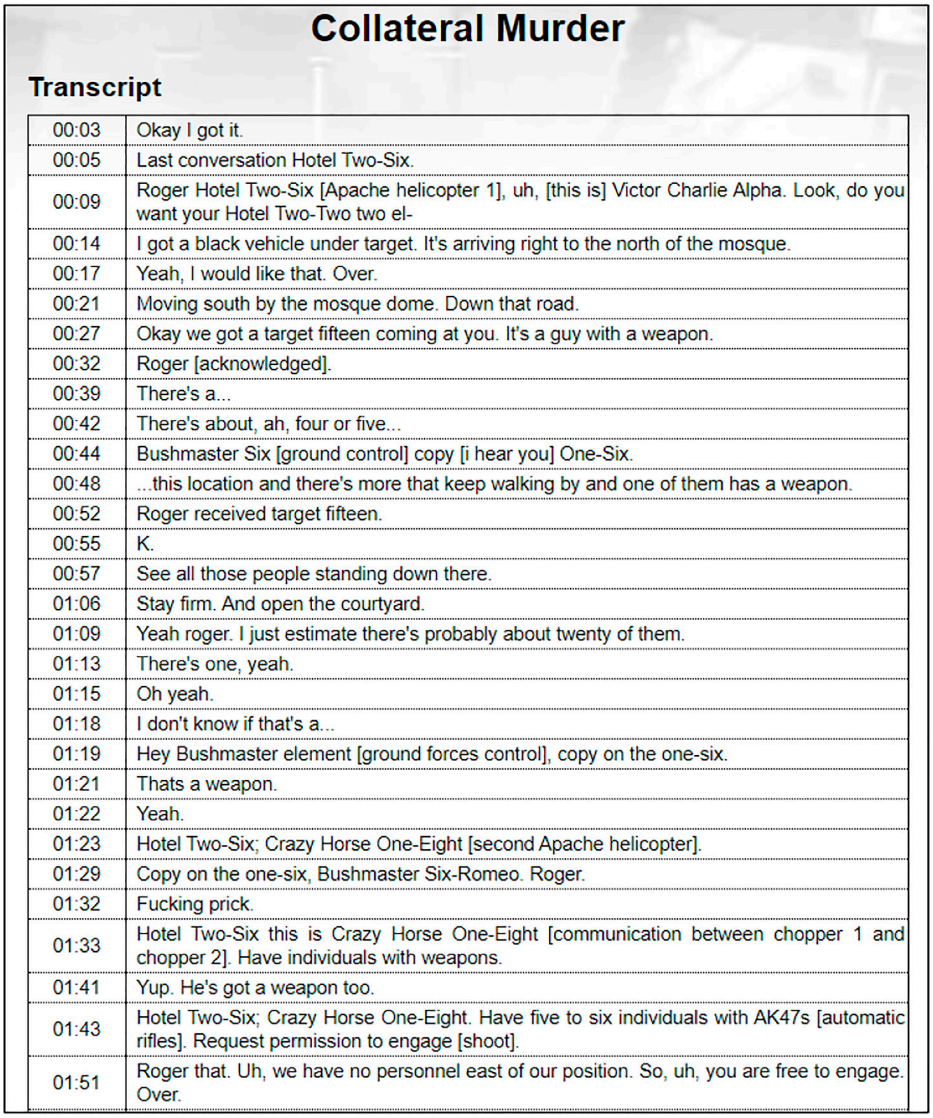
From their report then, we get a sense that what the computer requires and provides is a fixed variable-analytic codification of the work of doing astronautics that can form the basis for artificially-intelligent decision-making and deliver robust instructions on core tasks to astronaut crews. The crew autonomy that is promised, then, is partial, inasmuch as Kurtzman et al.'s (1986) MFIVE system is premised on having significant components of the work operate mechanistically (e.g., with a computer providing decision-making on the optimum ways to complete given core tasks, and astronauts then following the computer-generated instructions). In this sense, we might take their recommendations to be to de-emphasise the need for transcriptions altogether, as they argue that much of the decision-making might be taken off-comms altogether in the first place.

The notion of standardising and codifying the work of astronautics for the benefit of computerised methods (especially in regard to work which has previously been captured in and mediated through talk and its resultant transcriptions) is developed further by DeChurch et al. (2019), who leverage natural language processing techniques to analyze the conversation transcripts produced by Skylab missions. Chiefly, the text corpus is treated with topic modelling—"computational text analysis that discovers clusters of words that appear together and can be roughly interpreted as themes or topics of a document" (DeChurch et al., 2019: 1)—to demonstrate a standardised model of "information transmission" (DeChurch et al., 2019: 1) which can be organised and managed in ways that mitigate communicative troubles between astronaut crews and mission control. As with the Kurtzman et al. (1986) study, the notion embedded in DeChurch et al. (2019) use of the transcripts is one of standardisation; that astronauts' talk can be construed as a topically-oriented, discoverable phenomenon, the verbal content of which directly maps onto the work of doing astronautics. This is problematic for conceptual as well as practical reasons. Conceptually, the talk that is represented in a transcript does not necessarily fully elaborate on the goings-on of the settings and work within which that talk is contextually situated (cf. Garfinkel (1967) on good organisational reasons for bad clinical records). Practically, it is important to recognise that Skylab spent 40 minutes out of every hour out of radio contact with mission control due to its orbital trajectory taking it out of range of communications relay stations (and naturally, there is more to the work of doing astronautics than talking about doing astronautics; the astronauts were of course busy even during periods of loss-of-signal).

An interesting question then might be, if using conversation transcripts in the ways outlined above is problematic in terms of how a transcript maps onto the practices that produce it, how might we use them alternatively? An ethnomethodological treatment might instead focus on how the audio-only communications link is used to make the work of both astronauts and mission control accountable, and where the notion of "life" and "work" in space is defined and negotiated in terms of how it is to be undertaken, achieved and evaluated. The difference being pointed to here is between two positions. First, the approach that follows or more-or-less direct continuation of NASA's own staunchly scientific characterisations of living and working in space: conceptualising the work of astronauts and other spaceflight personnel as if it could be described in abstract universal terms (i.e., as if it can be codified as a set of rules and logical statements connecting them, such that a computer technology—artificial intelligence, natural language processing—can 'understand' this work as well as the human astronauts designated to carry it out). Second, leveraging the transcripts as some kind of (non-comprehensive, non-perfect) record through which we might learn something of what astronauts do and how they do it (which is often assumed a priori rather than described).

4.2 WikiLeaks' Post-Hoc Uses of the Collateral Murder Footage

Earlier, we accounted for the absence of a military-produced transcript documenting the talk of the individuals involved in the Collateral Murder incident by reference to the fact that the IO for the incident's AR 15-6 did not believe that the conduct of US personnel had played a causal role in the deaths of the 11 civilians killed in the strike. As we know, however, the US military were not the only organization to take an interest in the Collateral Murder case. As noted, Wikileaks published leaked gunsight



Collateral Murder

Transcript

00:03	Okay I got it.
00:05	Last conversation Hotel Two-Six.
00:09	Roger Hotel Two-Six [Apache helicopter 1], uh, [this is] Victor Charlie Alpha. Look, do you want your Hotel Two-Two two el-
00:14	I got a black vehicle under target. It's arriving right to the north of the mosque.
00:17	Yeah, I would like that. Over.
00:21	Moving south by the mosque dome. Down that road.
00:27	Okay we got a target fifteen coming at you. It's a guy with a weapon.
00:32	Roger [acknowledged].
00:39	There's a...
00:42	There's about, ah, four or five...
00:44	Bushmaster Six [ground control] copy [i hear you] One-Six.
00:48	...this location and there's more that keep walking by and one of them has a weapon.
00:52	Roger received target fifteen.
00:55	K.
00:57	See all those people standing down there.
01:06	Stay firm. And open the courtyard.
01:09	Yeah roger. I just estimate there's probably about twenty of them.
01:13	There's one, yeah.
01:15	Oh yeah.
01:18	I don't know if that's a...
01:19	Hey Bushmaster element [ground forces control], copy on the one-six.
01:21	Thats a weapon.
01:22	Yeah.
01:23	Hotel Two-Six; Crazy Horse One-Eight [second Apache helicopter].
01:29	Copy on the one-six, Bushmaster Six-Romeo. Roger.
01:32	Fucking prick.
01:33	Hotel Two-Six this is Crazy Horse One-Eight [communication between chopper 1 and chopper 2]. Have individuals with weapons.
01:41	Yup. He's got a weapon too.
01:43	Hotel Two-Six; Crazy Horse One-Eight. Have five to six individuals with AK47s [automatic rifles]. Request permission to engage [shoot].
01:51	Roger that. Uh, we have no personnel east of our position. So, uh, you are free to engage. Over.

FIGURE 8 | Wikileaks' Collateral Murder transcript—Opening sequence.

footage from one of the Apache helicopter's which carried out the strike in 2010. Alongside the video, Wikileaks released a rudimentary transcript of talk (**Figure 8** below) which was produced using recordings from the cockpit of that same Apache helicopter, the audio from which was included in the leaked video (see Mair et al., 2016).

In our previous discussion of the Collateral Murder case, we accounted for the absence of a military-produced transcript by reference to the fact that, in contrast to the Uruzgan incident, the AR 15-6 IO for the collateral murder case did not believe that the conduct of US personnel had played a causal role in the incident's outcome. Wikileaks' subsequent production of a transcript for the Collateral Murder case can be accounted for by examining their organization-specific practical purposes in taking up the video. In approaching the materials surrounding the strike, Wikileaks' objectives were radically different to that of the US military. Most notably, the Wikileaks approach is characterized by a significantly different perspective on the culpability of the US personnel involved in the operation. Though it is noteworthy that Wikileaks had relatively little to say about the incident itself, what little commentary does exist surrounding the transcript and the video footage points clearly towards a belief that the US personnel involved in the incident had acted both immorally and illegally. The first piece of evidence regarding this belief can be found in the incident's given name: Collateral *Murder* (Elsey et al., 2018). Implicit in such a title is an accusation that the strike did not constitute a legitimate killing in the context of an armed conflict. The brief commentary which surrounds the video reinforces such a claim, describing the strike as an "unprovoked slaying" of a wounded journalist (WikiLeaks,

TABLE 2 | Head-turning sign ("Last time memory let you down").

033 (dementia, accompanied)		
1	Neu	And could you, give me an example of the last time your memory, let you down?
2	—	(1.5)
3	Pat	Um: [(turns to AP1)]
4	—	(2.8)
5	AP1	In the car you've lost your sense of direction (.) does that count?
6	Pat	Right [(nods head)]
7	—	[(Pat and AP1 laugh)]

2010). Comparably to the Uruzgan incident, therefore, the production of a transcript has emerged alongside accusations regarding the failures of military personnel, wherein the transcript provides record by which the conduct of those personnel can be assessed in its details. As with the other cases we have presented up to this point, the Wikileaks transcript has several shortcomings—and in this final section of the paper it will be worth giving these apparent inadequacies some serious consideration in light of the Jeffersonian transcription system and Sacks' own reflections on the nature of transcripts.

5 TOPICALIZING THE WORK OF PRODUCING AND USING TRANSCRIPTS

The rudimentary character of the transcripts we have presented up to this point are particularly conspicuous when contrasted with excerpts of transcripts produced using the Jeffersonian transcription conventions. Consider the following transcript excerpt (**Table 2** below) taken from a study of a United Kingdom memory clinic where dementia assessments are conducted by neurologists (Elsey 2020: 201):

If we compare this transcript to the Wikileaks transcript of the collateral murder case (**Figure 8**), we can see various similarities. They both capture the "talk" recorded; they both separate the talk into distinct "utterances" which appear in sequence; and they both preserve the temporal aspects of the talk through the use of time stamps or line numbers. Nevertheless, the Wikileaks transcript differs from the memory clinic transcript insofar as it does not include any reference to the pauses which appear in natural conversation and, crucially, it does not include a distinct column to record "who" is speaking. The audio recordings for collateral murder case include the talk of two Apache helicopter crews, who are communicating both with one another as well as with numerous different parties on the ground, and without speaker identifiers, the action depicted in the Wikileaks transcript is extremely difficult to follow when read on its own. In comparison, the memory clinic interaction notes whether the neurologist (Neu), patient (Pat) or accompanying person (AP) is speaking, albeit the actual identities of the participants are anonymized for ethical purposes in the research findings.

From a CA perspective, therefore, the way in which talk has been presented in the Wikileaks transcript, and indeed in the Uruzgan and Skylab transcripts, fails to preserve a sufficient level of detail for serious fine-grained analysis of the action and interaction to be possible. In rendering speakers indistinguishable from one another, many of CA's central phenomena—most prominently sequentiality and turn-taking—are obscured (Sacks et al., 1978; Heritage, 1984; Jefferson, 2004; Schegloff, 2007; Elsey et al., 2016). This relates to how individual utterances in interaction both rely on and re-produce the immediate context of the on-going interaction. As such the intelligibility and sense of any utterances is tied to what was previously said and who it was addressed to. In military and space settings this is a critical issue given the number of communication channels and speakers involved.

Now, the lesson to be learned here is not that the transcripts presented over the course of this paper are, in any objective sense, inadequate. It might well be said that they are inadequate for the stated objectives of CA, but if this paper has demonstrated anything it is that conversation analysts are by no means the only ones interested in transcripts. The lesson, therefore, is that questions regarding what constitutes an adequate record of "what happened" are asked and answered within a field of organisationally specific relevancies. Over the course of this paper, we have demonstrated that a diversity of transcripts—many of which bear little resemblance to one another—can be adequately put to use towards a variety of ends depending upon the requirements of the organisation in question. Naturally, this same point applies in the context of transcripts produced using the Jeffersonian transcription conventions, which are, ultimately, just one benchmark for adequate transcription amongst countless others (e.g., Gibson et al., 2014 for a discussion). Towards that end, it is worth returning to an earlier quoted passage from Sacks, this time given more fully, in which he outlines his methodological position regarding audio-recordings in research.

> "I started to work with tape-recorded conversations. Such materials had a single virtue, that I could replay them. I could transcribe them somewhat and study them extendedly—however long it might take. The tape-recorded materials constituted a "good enough" record of what happened. Other things, to be sure, happened, but at least what was on the tape had happened."

From the founder of conversation analysis this could be read as a deflationary account of how recordings of talk can be analyzed. However, Sacks' explanation clearly speaks towards precisely the thing that transcripts make possible. In preserving talk and making it available for assessment, transcripts afford analysts the opportunity to make empirical assessments regarding 'what happened'. Thus, the distinctive move that this paper has proposed to make has been to treat the production and use of transcripts as a phenomenon in and of itself, topicalizing their contingent and institutionally produced character in order to gain an insight into the motives and objectives behind the transcription practices of the US Military and NASA. What we are recommending, then, based on our research, is that transcripts be seen as contextually embedded artifacts-in-use. Understanding them, therefore, means understanding the embedding context, how the transcript achieves its specific work of transcription and, crucially, what it allows relevant personnel to subsequently do.

6 CONCLUSION

The wide range of different transcripts (re)-presented in this paper indicate that we are dealing with huge organizations, with staff and technology to match. What also becomes apparent from our research is the huge amounts of "data" that NASA and the US military collect as part of their routine work activities. However, for various reasons (i.e., secrecy, sensitivity and so on) military organizations can be characterised as somewhat reluctant actors in terms of the transparency of their routine operations and procedures or the intelligibility of the materials released. As a result, public access to existing "data" (e.g., mission recordings, transcripts, documents) is severely restricted or difficult to make sense of. NASA's transcription machinery, on the other hand, is more oriented to issues of transparency, although the sheer volume of transcription materials conceivably counteracts that aim.

While a lot of the literature has pointed out the political significance of omitted content—conversational details that had not been included in the transcript—our comparison of NASA and US military transcription work adds a new perspective to that: transcripts can document too little or too much—both creating distinct problems for people relying on/using the transcripts. While in military contexts there is typically too little material, NASA's transcription machinery produced what might in latter-day social science, based on NASA's treatment of them, be construed as "Big Data" (Kitchin, 2014): large corpus interactional datasets that by virtue of their volume must necessarily rely on computational processing for their analyze (cf. DeChurch et al. (2019) and Kurtzman et al. (1986) discussed elsewhere in this paper), which itself embeds the assumption that talk is just one more scientific variable that NASA's scientists have at their analytic disposal. However, these scientistic efforts appear to deepen, rather than diminish, the "representational gap" in NASA's understanding of the work of astronautics, inasmuch as completionist all-in-one one-size-fits-all approaches do not seem to acknowledge the various mismatches between transcript and transcribed interaction. This is an area that EM and CA have a long-standing tradition in drawing attention to, which compounds their relevance here. In contrast to our previous published work (Mair et al., 2012, Mair et al., 2013, Mair et al., 2016, Mair et al., 2018; Elsey et al., 2016; Elsey et al., 2018; Kolanoski, 2017; Kolanoski, 2018), which focused on using the available "data" to describe and explicate military methods and procedures (e.g., communication practices and target identification methods), this study has used the available "data" and, specifically the transcripts produced internally, to demonstrate aspects of how these organizations work. For instance, the available transcripts we have examined here can provide an open door into the accounting practices of these specific organizations. One key use of transcripts in the military examples relates to the insights we gain about how the transcripts are treated as evidentiary documents during investigations following deadly "incidents". Though this may also be the case in how NASA leverages their transcriptions (c.f. Vaughan (1996) on usages of various data including conversation transcripts as diagnostic telemetry for forensically and legally examining disasters such as the 1986 Space Shuttle Challenger explosion), it is more typical that transcripts stand as a record of achievements of various kinds. That said, as we have seen, the transcripts that NASA produces are designed to feed into a broad range of activities (e.g. "doing spaceflight", "doing research", "doing public relations", etc), which dually resists attempts to treat them as standardisable documentation as NASA often conceive of them (cf. DeChurch et al. (2019) and Kurtzman et al. (1986)) and point towards the value of an EM/CA approach which can more carefully attune to the interactional nuance that NASA's own various teams draw on to extract useful information for their specific and discrete purposes (e.g. "doing spaceflight", "doing research", "doing public relations", etc).

One interesting observation that the paper makes plain is the fact that transcripts are rarely, if ever, read and used on their own in any of the examples included in this paper. The transcripts do *not* offer "objective" accounts that can speak for themselves in the way that videos are occasionally treated (Lynch, 2020). To read and make sense of a transcript requires context and background obtained from supplementary sources (e.g., interviews with participants, other documents). This is strongly linked to the veracity of the original recordings themselves.

A key question that this paper has returned to continually relates to the reasons why transcripts are produced by the different organizations. The military-based examples reveal that the transcription of the audio-video recordings is not a routine part of military action. Instead, it is seen as a required step in formal and/or legal investigations of incidents involving possible civilians or friendly fire. The analysis presented here unpacks the relationship between the audio/video and the transcript produced and raises questions about which (re)presentation of a mission takes primacy. In stark contrast, NASA's "transcription machinery" displays a systematic and completist approach to transcript production, ranging from scientific experiments, mundane greeting exchanges and all daily press conferences with mission updates (or lack thereof).

The what's and why's of transcription practices in these contexts are relatively easy to ascertain and describe. In contrast, the transcription methods themselves remain obscured and only recoverable from the documents produced. This applies to both the military and NASA where transcription practices and methods employed are rarely explicitly described or articulated in comparison to the Jeffersonian transcription techniques in CA. As such we do not learn who actually produced the transcripts and there is no account of the "conventions" used to format the transcripts. Arriving at answers to those questions thus requires additional investigative work. In the military cases, we can use the military "logs" to ascertain when they were produced in relation to the original events and the investigations. These logs and timelines document when transcription occurred (including when it was corrected and approved) and what was transcribed (e.g., witness testimony, gunsight camera/comms audio-video).

Transcription has a particular place within ethnomethodological and conversation analytic research traditions. It forms a central methodological tool and part of the analytical process. The techniques and conventions can be taught and can be applied to a wide range of recorded data. Therefore, a researcher who can "read" CA transcripts can effectively read any paper ethnomethodological and/or CA study that uses Jefferson's notations, whilst still being reliant upon the description of the context of the interaction and social setting. In stark contrast, "reading" the transcripts of NASA and the US military requires an ethnographic understanding of the working practices of these organizations. This raises important questions about how an artifact or document, such as the transcripts exhibited here, can be said to re-present the embodied and visual work that the soldiers or astronauts are undertaking through their interactions recorded during their respective missions. As Heritage 1995: 395fn, emphasis added) states in EM and CA:

> The transcript is valuable as a support for memory and as a means for the quick recovery of data segments . . . However, transcription is at best an approximation to the recorded data.

By contrast, and as this paper has demonstrated, the transcripts produced by the US military and NASA re-present an "approximation" of the original recorded "data" for all practical organizational purposes, no more but also no less.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

AH—Data collection (equal), Analysis (equal), Writing—Original draft (equal), Writing—Review and editing (equal) CE—Conceptualization (Equal), Data collection (equal), Analysis, Writing—Original draft, Writing—Review and editing MK—Conceptualization (equal), Writing—Review and editing (equal) PB—Data collection (equal), Analysis (equal), Writing—Original draft (equal), Writing—Review and editing MM—Data collection (equal), Writing—Review and editing (equal).

FUNDING

Alexander Holder's work on this paper was supported by the Economic and Social Research Council *via* the North West Social Science Doctoral Training Partnership, Grant number: ES/P000665/1. Open Access publication fees provided by Goethe University Frankfurt.

REFERENCES

Brooker, P. (Forthcoming). *Living and Working in Space: An Ethnomethodological Study of Skylab*. Manchester: Manchester University Press.

Brooker, P., and Sharrock, W. (Forthcoming). "Bricolage in Astronautics: Talk-In-Interaction in the Construction of Apollo 13's DIY CO2 Scrubber," in *Instructed and Instructive Actions*. Editors M. Lynch and O. Lindwall (London: Routledge).

Bucholtz, M. (2000). The Politics of Transcription. *J. Pragmatics* 32, 1439–1465. doi:10.1016/S0378-2166(99)00094-6

Compton, W. D., and Benson, C. D. (1983). *Living and Working in Space: A History of Skylab [The NASA History Series]*. Washington, D.C: Scientific and Technical Information Branch, National Aeronautics and Space Administration.

Davidson, C. (2009). Transcription: Imperatives for Qualitative Research. *Int. J. Qual. Methods* 8 (2), 35–52. doi:10.1177/160940690900800206

DeChurch, L. A., Schultz, M., and Contractor, N. S. (2019). *Networks as Filters: Selective Information Transmission in Conversations Aboard Skylab, 2019 NASA Human Research Program Investigators' Workshop Annual Conference*. Texas, USA: Galveston. January 22-25 2019: 1.

Department of the Army (2016). *Procedures for Administrative Investigations and Boards of Officers*. Washington, D.C: United States Army: Headquarters.

Department of the Air Force (2010). *Uruzgan Province CIVCAS Incident 21 February 2010: Commander Directed Investigation*. Creech, Nevada: Department of the Air Force Headquarters.

Duranti, A. (2006). Transcripts, like Shadows on a Wall. *Mind, Cult. Activity* 13 (4), 301–310. doi:10.1207/s15327884mca1304_3

Elsey, C. (2020). "Dementia in Conversation: Observations from Triadic Memory Clinic Interactions," in *Atypical Interaction: The Impacts of Communicative Impairments within Everyday Talk*. Editors R. Wilkinson, J. Rae, and G. Rasmussen (Basingstoke: Palgrave Macmillan), 195–221. doi:10.1007/978-3-030-28799-3_7

Elsey, C., Mair, M., and Kolanoski, M. (2018). Violence as Work: Ethnomethodological Insights into Military Combat Operations. *Psychol. Violence* 8, 316–328. doi:10.1037/vio0000173

Elsey, C., Mair, M., Smith, P. V., and Watson, P. G. (2016). "Ethnomethodology, Conversation Analysis and the Study of Action-In-Interaction in Military Settings," in *The Routledge Companion to Military Research Methods*. Editors A. J. Williams, N. Jenkings, R. Woodward, and M. F. Rech (Abingdon: Routledge), 180–195.

Froehlich, W. (1971). *Man in Space: Space in the Seventies*. Washington, D. C.: National Aeronautics and Space Administration.

Garfinkel, H. (1986). *Ethnomethodological Studies of Work*. Abingdon: Routledge & Kegan Paul.

Garfinkel, H. (2002). *Ethnomethodology's Program: Working Out Durkheim's Aphorism*. Lanham, MD: Rowman & Littlefield Publishers, Inc.

Garfinkel, H. (1967). *Studies in Ethnomethodology*. Cambridge: Polity.

Gibson, W., Webb, H., and Vom Lehn, D. (2014). Analytic Affordance: Transcripts as Conventionalised Systems in Discourse Studies. *Sociology* 48 (4), 780–794. doi:10.1177/0038038514532876

Heritage, J. (1995). "*Conversation Analysis: Methodological Aspects," Aspects of Oral Communication*. New York and Berlin: U. QuasthoffWalter de Gruyter, 391–418.

Heritage, J. (1984). *Garfinkel and Ethnomethodology*. Cambridge: Polity Press.

Hitt, D., Garriott, O., and Kerwin, J. (2008). *Homesteading Space: The Skylab Story*. London: University of Nebraska Press.

Holder, A. (2020). The Centrality Of Militarised Drone Operators In Militarised Drone Operations. *Ethnographic Studies* 17, 81–99. doi:10.5281/zenodo.4050543

Holder, A., Minor, E., and Mair, M. (2018). Targeting Legality: The Armed Drone as a Socio-technical and Socio-Legal System. *J. Oxford Centre for Socio-Legal Studies* 1, 1–7.

Investigating Officer 2nd Brigade Combat Team 2nd Infantry Division, (MND-B) (2007). *Investigation into Civilian Casualties Resulting from an Engagement on 12 July 2007 in the New Baghdad District of Baghdad, Iraq*. Arlington, VA: US Department of Defense.

Jefferson, G. (2004). "Glossary of Transcript Symbols with an Introduction," in *Conversation Analysis: Studies from the First Generation*. Editor G. Lerner (Amsterdam and Philadelphia, PA: John Benjamins), 13–31. doi:10.1075/pbns.125.02jef

Jefferson, G. (2015). *Talking about Troubles in Conversation*. Oxford: Oxford University Press.

Kitchin, R. (2014). Big Data, New Epistemologies and Paradigm Shifts. *Big Data Soc.* 1 (1), 1–12. doi:10.1177/2053951714528481

Kolanoski, M. (2018). Trans-Sequential Analysis: A Production-Focused Approach to Procedurally Organized Work. *Ethnographic Stud.* 15, 58–82. doi:10.5281/zenodo.1475777

Kolanoski, M. (2017). Undoing the Legal Capacities of a Military Object: A Case Study on the (In)Visibility of Civilians. *L. Soc. Inq.* 42, 377–397. doi:10.1111/lsi.12284

Kurtzman, C. R., Akin, D. L., Kranzler, D., and Erlanson, E. (1986). *Study of Onboard Expert Systems to Augment Space Shuttle and Space Station Autonomy: Final Report*. Cambridge, MA: Massachusetts Institute of Technology.

Lapadat, J. C. (2000). Problematizing Transcription: Purpose, Paradigm and Quality. *Int. J. Soc. Res. Methodol.* 3 (3), 203–219. doi:10.1080/13645570050083698

Lindsay, J. (2020). *Information Technology and Military Power*. Ithaca, NY: Cornell University Press.

Lynch, M., and Bogen, D. (1996). *The Spectacle of History: Speech, Text and Memory at the Iran-Contra Hearings*. Durham, NC: Duke University Press.

Lynch, M. (2020). "Vernacular Visions of Viral Videos," in *Legal Rules in Practice: In the Midst of Law's Life*. Editors B. Dupret, J. Colemans, and M. Travers (Routledge), 182–204. doi:10.4324/9781003046776-12

Mair, M., Elsey, C., Smith, P. V., and Watson, P. G. (2016). "The Violence You Were/n't Meant to See: Representations of Death in an Age of Digital Reproduction," in *The Palgrave Handbook of Criminology and War*. Editors R. McGarry and S. Walklate (London: Palgrave Macmillan UK), 425–443. doi:10.1057/978-1-137-43170-7_23

Mair, M., Elsey, C., Smith, P. V., and Watson, P. G. (2018). War on Video: Combat Footage, Vernacular Video Analysis and Military Culture from within. *Ethnographic Stud.* 15, 83–105. doi:10.5281/zenodo.1475784

Mair, M., Elsey, C., Watson, P. G., and Smith, P. V. (2013). Interpretive Asymmetry, Retrospective Inquiry and the Explication of Action in an Incident of Friendly Fire. *Symbolic Interaction* 36, 398–416. doi:10.1002/symb.78

Mair, M., Watson, P. G., Elsey, C., and Smith, P. V. (2012). War-making and Sense-Making: Some Technical Reflections on an Instance of 'friendly Fire'. *Br. J. Sociol.* 63, 75–96. doi:10.1111/j.1468-4446.2011.01394.x

Newell, H. E. (1980). *Beyond the Atmosphere: Early Years of Space Science*. Washington, D.C.: National Aeronautics and Space Administration, Scientific and Technical Branch.

Ochs, E. (1979). "Transcription as Theory," in *Developmental Pragmatics*. Editors E. Ochs and B. Schieffelin (New York, N: Academic Press), 43–72.

Planetary Society (2021). Your Guide to NASA's Budget. Available at: https://www.planetary.org/space-policy/nasa-budget (Accessed October 15, 2021).

Raffel, S. (1979). *Matters of Fact*. London & New York: Routledge.

Reuters Staff (2007). *Reuters Photographer and Driver Killed in Iraq*. Reuters. Available at: https://www.reuters.com/article/us-reuters-iraq-deaths-idUSL1280681020070712 (Accessed October 1, 2021).

Rubin, A. J. (2007). *2 Iraqi Journalists Killed as U. S. Forces Clash with Militias*. New York: The New York Times. Available at: https://archive.ph/20120714131743/http://query.nytimes.com/gst/fullpage.html?res=9D0DEED81E3EF930A25754C0A9619C8B63 (Accessed October 1, 2021).

Sacks, H. (1984). "Notes on Methodology," in *Structures of Social Action: Studies in Conversation Analysis*. Editors J. Atkinson and J. Heritage (Cambridge: Cambridge University Press), 21–27.

Sacks, H., Schegloff, E., and Jefferson, G. (1978). "A Simplistic Systematics for the Organisation of Turn-Taking for Conversation," in *Studies in the Organization of Conversational Interaction*. Editor J. Schenkein (New York, NY: Academic Press), 7–55.

Schegloff, E. (1968). Sequencing in Conversational Openings. *Am. Anthropol.* 70, 1075–1095.

Schegloff, E. A. (1995). "Introduction," in *Lectures on Conversation: Volumes I & II (by Harvey Sacks)*. Editor G. Jefferson (Oxfordix-lii: Blackwell).

Schegloff, E. A. (2007). *Sequence Organisation in Interaction: A Primer in Conversation Analysis I*. Cambridge: Cambridge University Press.

United States Central Command (2010). *AR 15-6 Investigation, 21 February 2010 CIVCAS Incident in Uruzgan Province*. Tampa, FL: United States Central Command.

Vaughan, D. (1996). *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA*. London: University of Chicago Press.

WikiLeaks (2010). Collateral Murder. Available at: https://collateralmurder.wikileaks.org/en/index.html (Accessed January 26, 2015).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

From Verbal Account to Written Evidence: Do Written Statements Generated by Officers Accurately Represent What Witnesses Say?

Rebecca Milne[1]*, *Jordan Nunan*[1], *Lorraine Hope*[2], *Jemma Hodgkins*[1] and *Colin Clarke*[1]

[1] Institute of Criminal Justice Studies, University of Portsmouth, Portsmouth, United Kingdom, [2] Department of Psychology, University of Portsmouth, Portsmouth, United Kingdom

Most countries compile evidence from witnesses and victims manually, whereby the interviewer assimilates what the interviewee says during the course of an interview to produce an evidential statement. This exploratory research examined the quality of evidential statements generated in real world investigations. Transcribed witness/victim interviews (*N* = 15) were compared to the resultant written statements produced by the interviewing officer and signed as an accurate record by the interviewee. A coding protocol was devised to assess the consistency of information between what was said by the interviewee in the verbal interview and what was reported in the written statement. Statements contained numerous errors including omissions, distortions, and the inclusion of information not mentioned in the verbal interview. This exploratory work highlights an important area for future research focus.

Keywords: witness, investigative interviewing, evidence, consistency, statements

OPEN ACCESS

Edited by:
Kate Haworth,
Aston University, United Kingdom

Reviewed by:
Miet Vanderhallen,
University of Antwerp, Belgium
Jamal K. Mansour,
Queen Margaret University,
United Kingdom

****Correspondence:***
Rebecca Milne
becky.milne@port.ac.uk

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *11 September 2021*
Accepted: *21 December 2021*
Published: *10 February 2022*

Citation:
Milne R, Nunan J, Hope L, Hodgkins J and Clarke C (2022) From Verbal Account to Written Evidence: Do Written Statements Generated by Officers Accurately Represent What Witnesses Say? Front. Psychol. 12:774322. doi: 10.3389/fpsyg.2021.774322

INTRODUCTION

Witnesses are central to most criminal cases; indeed, some have argued they provide the most critical evidence in court (Zander and Henderson, 1993). Consequently, considerable attention has been paid to developing techniques that elicit reliable, relevant, and detailed information from witnesses during interviews (Gabbert et al., 2016; Milne and Bull, 2016). Traditionally, witnesses provide their accounts at two separate points of the criminal justice process; first when interviewed during the investigation and later when giving evidence during criminal proceedings (Westera et al., 2011). The information provided initially as part of the investigation not only informs investigative decision making (e.g., what lines of inquiry to pursue and prioritize), it is also central to legal decision-making, for example, whether to proceed with the case (or not). The written statement, produced when the interviewer assimilates the information provided by the witness in the course of the interview, is also key in any resultant court-case, informing legal strategy and likely serving as a memory aid for the witness. Clearly, the written product of the witness interview should thus be an accurate representation of what the witness reports about the event in question. The criminal justice system relies on the accuracy of this statement to avoid ill-informed investigative and legal decisions. This exploratory research examined the quality of evidential statements taken in real world investigations and, specifically, assessed the extent to which the written statements produced were in fact consistent with the content of the associated verbal interviews.

The purpose of an investigation is to establish what, if any, criminal offending has taken place and the identity of those who may be culpable (Milne and Bull, 2006). To answer these primary investigative questions the police seek information from a number of sources, including witnesses. The most common way to formalize witness accounts across the world is for an officer to produce a written (hand-written or typed) statement reflecting the information obtained during an interview. Statement production is often conducted at the same time as interviewing the witness, however this is dependent on circumstance (e.g., dynamic nature of the event), crime type (e.g., seriousness of the offense), officer training, and individual preference, i.e., there is limited evidenced based practice guidance (though see Smith and Milne, 2018 for a United Kingdom example- WISCI- Witness Interview Strategies for Critical Incidents). After being endorsed and signed by the witness, the statement is then used as the basis for investigative and legal decision-making. Often, the interview or the process of transferring the verbal content of that interview into a written statement is not electronically recorded or otherwise documented.

To date, psychological research has concentrated on enhancing our understanding of how the interview process can affect a witness's memory recall of events and the development of techniques to enhance the quality and quantity of information obtained in witness interviews (Vrij et al., 2014). These advances have influenced police practices in many jurisdictions (Milne et al., 2019) and there is now growing consensus with respect to witness interviewing best practice (see Meissner, 2021 and associated special issue). Interviewers are encouraged to start such interviews with a free recall, followed by open-ended prompts and questions, and finishing with appropriate non-leading closed questions if necessary (see e.g., Achieving Best Evidence Guidance, Home Office, 2011). Open questions such as "tell me what happened..." are generally considered the best type of question to use because they encourage a detailed and unrestricted answer. As questions become more specific or interviewer-driven, responses become less accurate (Oxburgh et al., 2010; Boon et al., 2020; Kontogianni et al., 2020). In practice however, the usual method of recording the witness-police interaction relies on the interviewer's own memory of what the witness said and there is typically no actual record of the questions used by the interviewer to obtain the witness's account.

Indeed, Barristers Heaton-Armstrong and Wolchover (1992) were one of the first to argue that written statements are mistakenly treated by the criminal justice system as a verbatim record of interview:

> *"There is a certain coyness on the part of most officers, when asked how they "took" a statement, in admitting that the narrative was obtained by questioning. The fiction is perpetuated that for the most part statements are the product of straight dictation." (p. 161).*

The production of a written statement involves the interviewer, both deliberately and inadvertently, filtering the information generated by the witness during the interview, and deciding what should and should not be included in the statement (Westera et al., 2011). The cognitive demands of this task make it susceptible to distortion at many stages and the resulting statement is an abridged and often inaccurate version of what was said within the interaction. Further, in the United Kingdom (and many other countries) there is no legal requirement to make a record of the utterances of the interviewer (e.g., questions used) within the resultant statement. Given the importance of witness statements within the criminal justice system, there has been very limited research examining the accuracy of this witness statement-taking process.

Kohnken et al. (1994) examined the statement-taking process in a mock-witness experimental paradigm and found statements written by the interviewer immediately after the interview contained only about two thirds of the information reported by the witness. Similarly, Hyman Gregory et al. (2011) examined notes made by 13 US police investigators during a single mock witness interview and compared them to their subsequent reports. This comparison revealed that 68% of the information reported by the witness was omitted with 40% of the omitted information being deemed crime-relevant. In a US sample of 20 real-life interviews with child witnesses/victims, Lamb et al. (2000) found the interviewer's "verbatim" notes were missing 25% of the forensically relevant details reported by the witness. In the United Kingdom, McLean (1995) examined 16 formal witness-police interviews and found that none of the statements contained all the relevant information reported by witnesses. These types of omission errors may be due to the cognitive load inherent in the multitude of tasks that constitute the statement taking process, for example, actively listening to the interviewee, formulating which questions to ask, assimilating the information reported, and taking comprehensive notes (Fisher et al., 2014; Kleider-Offutt et al., 2015; Hanway et al., 2021). Indeed, the cognitive load associated with the conduct of interviews is well recognized by police interviewers (Hanway and Akehurst, 2018). One possible result of reduced cognitive resources is that interviewers may, unwittingly, prioritize information that fits with their existing expectations or schema for the reported event. When information provided by a witness is not consistent the interviewer can: (i) include the information in full; (ii) distort the information to make it more consistent, or (iii) omit the information altogether (McLean, 1995). Furthermore, and worryingly, it would appear that witnesses fail to detect such revisions or errors in their own statements (Sagana et al., 2017).

Using cases drawn from two forces in the United Kingdom, the current research examined the consistency between information provided in verbal interviews with the resultant evidential statement. Specifically, we sought to identify any inconsistencies emerging in this translative process and describe the nature of those inconsistencies using a comprehensive coding protocol.

MATERIALS AND METHODS

Case Materials

As part of the national evaluation of PEACE in the United Kingdom (Clarke and Milne, 2001) police officers were asked to record their interviews with real-life witnesses/victims, including the statement taking segment of the interview. Six forces (of 43 force areas) in England and Wales agreed to participate in the research. In order to gain a representative sample across the country, forces were selected based on willingness to participate, geographical location, and size of force (for a full outline of the National evaluation, see Clarke and Milne, 2001). At the time, two forces also gathered the resultant hand-written statement and submitted them to the research team as part of the project materials, but these were not included as part of the original evaluation, which focussed on the quality of the interview process. For the current research, 15 cases where the recorded interview with a witness including the statement-taking segment and the resultant hand-written statement were available and were analyzed. The cases analyzed included ten thefts, three criminal damage cases, one assault, and one public order incident. The statement length varied in length from 1 to 6 pages (*M* = 2.7, *SD* = 1.4).

Coding Protocol

Drawing on the existing literature on consistency across reporting in investigative settings (e.g., Fisher et al., 2009), a coding protocol was developed to determine the extent to which what the witness reported in the interview was consistent with what the officer recorded as their evidence, at the time, in the form of a hand-written statement.

The following categories were included in the coding protocol: (i) consistent details (mentioned by the interviewee and included in the hand-written statement); (ii) omissions (mentioned by the interviewee and omitted from the hand-written statement); (iii) distortions (mentioned by the interviewee and written down incorrectly by the interviewer); (iv) contradictions (written in the statement but directly contradicts what was said by the interviewee), and (v) intrusions (not mentioned by the interviewee at any point but included in the statement). Omissions, distortions, contradictions and intrusions all reflect error in the translation of a verbal account into a written statement. We also coded for the category "known information" which reflects factual information known to the interviewer (mainly demographic) but not necessarily mentioned in the interview (e.g., address of interviewee). Following common interview coding approaches (e.g., Milne and Bull, 2002; Gabbert et al., 2009), each category was also coded with respect to type of detail i.e., persons, actions, objects, surroundings, conversation, and temporal information.

The second author coded the data, which comprised 15 hand-written statements and their partnering interview transcripts for comparison. The procedure involved comparing each hand-written statement to the counterpart transcript of what the witness actually said within the interview. Firstly, the coder examined each detail in the transcript and ascertained whether it was included in the statement. If not, then it was coded as an omission (1 point per item of information). If it was included within the statement then it was determined if it was included accurately and coded as either, "known," "consistent," "distortion" or "contradictory" (1 point per item of information). Any information within the statement but not in the transcript, was coded as an inclusion (1 point per item of information). The final step was to code each piece of information with respect to detail type- person etc. (as outlined above). An independent coder was randomly assigned four of the hand-written statements and their partnering interview transcripts and followed the same coding procedure. Across the two raters there was only one minor discrepancy (i.e., in one statement one rater scored 28 for consistent person details whereas the other rater scored 29). Thus, the overall inter-rater reliability agreement across all the thirty-six independent variables within the four statements was 99.3%.

RESULTS AND DISCUSSION

All 15 final statements contained errors in that their content diverged from the original verbal account provided by the witness in at least one of the ways captured by our coding protocol. Descriptive results across each of the 15 statements are presented in **Table 1**. Consistent detail percentages ranged from 19.28 to 86.97%. Known facts accounted for 1.30–20.00% of the statements. The most commonly observed type of error were omission errors which ranged from 4.76% of a statement to 51.81%, followed by distortions ranging from 1.85 to 19.28% of a statement. The intrusion of new (i.e., previously unmentioned) information had a range of 0.00–20.51% of details. Only two statements did not include any intrusion errors. Finally, three statements contained contradictory information (range 0.00–5.00%). Examples of each error category observed in statements are presented in **Table 2**.

To summarize, every statement examined contained errors, primarily omissions, followed by distortions and then intrusions (new) information. Thus, in this sample, the evidential product (i.e., the witness statement) was never an exact replication of what the witness actually said at interview. Worse, in some cases there were sizeable discrepancies between the original verbal account provided by the witness and what the officer recorded in the statement. There are a number of possible reasons for such discrepancies. First, there are significant cognitive demands associated with both interviewing and statement-taking. Recent research by Hanway et al. (2021) observed that when people complete tasks intrinsic to investigative interviewing (such as listening, remembering, judging the information provided, and generating follow-up questions to ask) not only do they experience a higher cognitive burden than those who simply have to listen to a witness's statement, but they also make more recall errors when asked to recall what the witness actually said. Further, research examining memory for conversation has found that it tends to be gist as opposed to verbatim due to competing demands (Brown-Schmidt and Benjamin, 2018). In the current sample of statements, recall errors may well be reflected in the omission errors (information the interviewer did not remember when writing the statement) and distortion errors (information the interviewer remembered incorrectly when writing the statement). Second, when writing the statement, information that fits with an existing schema for the reported event (e.g., an archetype; Shepherd and Milne, 1999) may have inadvertently been prioritized over non-schematic information, particularly when cognitive resources were limited. Finally, some discrepancies may reflect the preconceptions or beliefs officers hold about what constitutes "a good statement"

TABLE 1 | Number of details per interview transcript and hand-written statement pair across coded consistency category (% of transcript); illustrates discrepancy across each of the fifteen statements.

Consistency category	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Consistent details	102 (69.39)	74 (63.25)	95 (58.64)	99 (43.04)	130 (68.78)	96 (41.03)	121 (66.12)	249 (72.59)	37 (52.86)	320 (85.33)	267 (86.97)	245 (71.33)	16 (19.28)	167 (60.51)	347 (67.77)
Known details	8 (05.44)	9 (07.69)	7 (04.32)	3 (01.30)	7 (03.70)	26 (11.11)	27 (14.75)	33 (09.62)	14 (20.00)	6 (01.60)	10 (03.26)	30 (08.72)	6 (07.23)	10 (03.62)	20 (03.91)
Intrusions	16 (10.88)	7 (05.98)	1 (00.62)	9 (03.91)	22 (11.64)	32 (13.68)	3 (01.64)	8 (02.33)	0 (00.00)	3 (00.80)	0 (00.00)	25 (07.27)	2 (02.41)	19 (06.88)	105 (20.51)
Distortions	14 (09.52)	6 (05.13)	3 (01.85)	17 (07.39)	8 (04.23)	15 (06.41)	9 (04.92)	46 (13.41)	6 (08.57)	8 (02.13)	8 (02.61)	20 (05.81)	16 (19.28)	21 (07.61)	12 (02.34)
Contradictions	0 (00.00)	0 (00.00)	0 (00.00)	0 (00.00)	0 (00.00)	0 (00.00)	1 (00.55)	5 (01.46)	0 (00.00)	0 (00.00)	0 (00.00)	0 (00.00)	0 (00.00)	4 (01.45)	0 (00.00)
Omissions	7 (04.76)	21 (17.95)	56 (34.57)	102 (44.35)	22 (11.64)	65 (27.78)	22 (12.02)	2 (00.58)	13 (18.57)	38 (10.13)	22 (07.17)	24 (06.98)	43 (51.81)	55 (19.93)	28 (05.47)
Total details	147	117	162	230	189	234	183	343	70	375	307	344	83	276	512

and what information is relevant or appropriate to include. In such instances, officers may have edited or distorted the information accordingly.

Thirteen of the statements included information that was not mentioned by the interviewee. In other words, "new" intruded information (beyond known facts) was introduced by the officer when writing the statement. This new information may be the result of a source monitoring error whereby the officer misremembered the original source of the information and accidentally attributed it to the witness interview when in fact the information was obtained elsewhere (e.g., another witness; see Source Monitoring Framework; Johnson et al., 1993; Hanway, 2021). As the number of witnesses the interviewer deals with increases, this type of error is likely to be more prevalent. It could also be the case that interviewers incorporate this "new" information to increase the plausibility of the witness's account. Indeed, visually recorded police interviews (often used as evidence in chief) are regularly critiqued by legal practitioners for not being succinct and not taking the form of a coherent chronological narration (Westera et al., 2017). Worryingly, interviewers striving for a "good" statement in the eyes of the justice system may result in evidence that is distorted, has intrusions, and with omissions. Future research should further explore the extent to which preconceptions about what constitutes a "good statement" and how any pre-existing beliefs distort the production and evaluation of statements and other evidence. For these reasons, psychological, legal, and linguistic professionals alike have criticized the justice system for an over-reliance on the statement-taking process as it lacks accuracy, legitimacy, and transparency (e.g., Heaton-Armstrong and Wolchover, 1992; Milne and Shaw, 1999; Rock, 2001; Westera et al., 2011).

TABLE 2 | Examples of discrepancies across the interview transcripts and hand-written statements.

Consistency category	Interview transcript—verbal evidence	Hand-written statement—written evidence
Distortions	1. "Few of the lads." 2. "One of them" (carrying TV).	1. "Gang of youths." 2. "They were carrying TV."
Contradictions	1. "Couldn't hear what was being said."	1. "I recall the conversation during this."
Omissions	1. "Car was definitely a Metro." 2. "I didn't actually see any damage." 3. "No caps, no glasses on youths."	1, 2, and 3 omitted from written evidence.
Intrusions	1 and 2 not mentioned by the witness during the interview.	1. "There were no obstructions to my view." 2. "Brown hair."

To examine the nature of the errors in more depth, errors were examined with respect to detail type (person, action, object, surrounding, conversation, and temporal). Means (and SDs) across detail type were calculated for each variable (omissions, distortions, and intrusions) across the 15 statements. Only three statements contained contradictory information. An example can be seen in **Table 2** and worryingly it concerned witness reliability with regard their visibility at the time the event was witnessed. **Table 3** shows that every type of detail was omitted and this occurred in each of the 15 pairings. Omission errors primarily pertained to the objects and people involved in the incident and their actions. Distortions also primarily concerned the people in the events, where the event took place, what people did, and the objects involved. With respect to intrusions, the largest mean number of these errors related to the objects involved, followed by the key players in the incidents and what they did. In sum, errors identified in this sample pertained to forensically relevant details, including information about the perpetrator, their actions, where the incident took place and the objects used.

Overall, given that this exploratory work identified clear discrepancies in all of the statements that were examined with reference to the original account given by the witness, it appears this issue may be relatively commonplace. If that is indeed the case, what are the implications? First, given that the criminal justice system relies on accurate witness statements to both pursue investigations and inform subsequent legal decision making, statements that contain errors of any kind may not only

TABLE 3 | Means and standard deviations for consistency categories by detail type across interview-statements (*N* = 15).

	Consistent details *M* (*SD*)	Known details *M* (*SD*)	Intrusions *M* (*SD*)	Distortions *M* (*SD*)	Contradictions *M* (*SD*)	Omissions *M* (*SD*)
Person	41.80 (30.93)	4.13 (6.58)	4.13 (6.58)	4.13 (3.48)	0.27 (0.59)	9.33 (7.86)
Action	35.00 (23.24)	2.00 (2.98)	3.53 (5.89)	3.27 (3.63)	0.20 (0.56)	7.93 (7.78)
Object	40.87 (24.85)	1.93 (2.40)	5.47 (13.25)	2.40 (1.92)	0.07 (0.26)	8.67 (4.78)
Surroundings	28.53 (27.19)	3.33 (2.47)	2.20 (2.88)	2.93 (2.79)	0.07 (0.26)	5.80 (7.70)
Conversations	2.47 (4.22)	0.07 (0.26)	0.07 (0.26)	0.20 (0.77)	0.00 (0.00)	0.93 (1.75)
Temporal	9.00 (9.45)	3.60 (2.64)	1.40 (1.76)	1.00 (1.36)	0.07 (0.26)	2.00 (2.67)

result in wasted time and resources but also jeopardize the pursuit of justice. Secondly, given that cases can take some time to come to court, witnesses may rely on reviewing their statement before testifying. If that statement contains erroneous information, then it is entirely possible that the witness's memory of their original experience will be distorted accordingly (e.g., Misinformation Effect; see Frenda et al., 2011, for a review).

There is a simple solution to address such concerns: visually record all evidence gathering interactions, harnessing technology, such as a body-worn video recording device, to legitimize the process and allow reliability assessment. Indeed, some jurisdictions now favor visually recording the process, especially for vulnerable groups (Davies et al., 2016). However, a move toward more accurate witness testimony through visual recording also requires an understanding and adoption of basic memory principles (i.e., that memory is both fallible and easily contaminated) and that the written statement is not the verbatim record it was previously assumed to be. In addition, the raw product of memory, such as recall, may not emerge in the form of a chronologically narrated, comprehensively detailed story. Nonetheless, allowing the witness to provide their own account in their own words, is more likely to provide accurate investigative and evidential information compared to a non-transparent, ill-monitored, translational process such as statement-taking.

This preliminary project examined a small sample of cases and, although consistent with the case samples examined by previous researchers (e.g., McLean, 1995; Lamb et al., 2000; Hyman Gregory et al., 2011), further work is necessary to examine this issue across a larger case sample involving different case types. For instance, it may be the case that certain case types are more prone to some of the translational issues we observed in the current sample. Indeed, there are potentially a multitude of factors that could influence the statement taking process (such as training regimes, method trained for interviewing witnesses and so on). Notably, however, every statement in our sample contained errors—a finding that is also consistent with previous research (McLean, 1995). It is also important to note that the interview-statement pairings were from the 2001 evaluation study, however, there has been almost no research or practice change since that time with regard the production of witness statements, and thus the results are reflective of current practice. Many countries also do not electronically record their interviews/interrogations with suspects, instead a written report is produced (e.g., the Netherlands). A limited amount of work has started to look at the accuracy of this written report and has similarly found omission errors (e.g., Malsch et al., 2018). For example, Malsch et al. (2018) found that only 24% of all spoken words were accounted for in the reports, though this included interviewer and interviewee utterances. More research is urgently needed in this area.

To conclude, in the current study, a comprehensive coding protocol allowed us to determine that the errors identified in this sample in the form of omissions, distortions, and intrusions, pertained to forensically relevant details. In all fifteen statements there were errors, across all detail types, though there was a lot of variability across the statements. Omission errors were the most frequently observed error. Thus, due to cognitive demands of the multi-faceted interviewing task, errors will emerge. Perhaps it is time to acknowledge that, despite their importance within the criminal justice system, statements generated in this translational way are likely error-ridden as a result of imperfect human cognition and that technological solutions should take precedence.

DATA AVAILABILITY STATEMENT

Due to the private/sensitive nature of the material, the datasets presented in this article are not readily available. Requests to access the datasets should be directed to the lead author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Home Office funded research permission and Metropolitan Police permission, and complied with the University of Portsmouth Ethics procedure at that time. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

RM was PI on the project, developed coding system, attained the data, helped code and score, analyzed the data, and wrote the manuscript. JN coded and scored data set, analyzed the data, and helped write the manuscript. LH helped design the coding system and edited manuscript. JH coded and scored data. CC attained data set in the first place. All authors contributed to the article and approved the submitted version.

REFERENCES

Boon, R., Milne, R., Rosloot, E., and Heinsbroek, J. (2020). Demonstrating detail in investigative interviews: An examination of the DeMo technique. *Appl. Cogn. Psychol.* 34, 1133–1142. doi: 10.1002/acp.3700

Brown-Schmidt, S., and Benjamin, A. S. (2018). How we remember conversation: Implications in legal settings. *Policy. Insights. Behav. Brain Sci.* 5, 187–194. doi: 10.1177/2372732218786975

Clarke, C., and Milne, R. (2001). *National Evaluation of the PEACE Investigative Interviewing Course. Police Research Award Scheme: Report No. PRAS/149*. Portsmouth, UK: University of Portsmouth.

Davies, G., Bull, R., and Milne, R. (2016). "Analysing and improving the testimony of vulnerable witnesses interviewed under the 'Achieving Best Evidence' protocol," in *Witness Testimony in Sexual Cases: Investigation, Law and Procedure*, eds P. Radcliffe, A. Heaton-Armstrong, G. Gudjonsson, and D. Wolchover (Oxford: Oxford University Press), 207–220.

Fisher, R. P., Brewer, N., and Mitchell, G. (2009). "The relation between consistency and accuracy of eyewitness testimony: Legal versus cognitive explanations," in *Handbook of Psychology of Investigative Interviewing: Current Developments and Future Directions*, eds R. Bull, T. Valentine, and T. Williamson (Chichester: Wiley).

Fisher, R. P., Schreiber Compo, N., Rivard, J., and Hirn, D. (2014). "Interviewing witnesses," in *The SAGE Handbook of Applied Memory*, eds T. J. Perfect and D. S. Lindsay (Sage: Oxford University Press).

Frenda, S. J., Nichols, R. M., and Loftus, E. F. (2011). Current Issues and Advances in Misinformation Research. *Curr. Dir. Psychol. Sci.* 20, 20–23. doi: 10.1177/0963721410396620

Gabbert, F., Hope, L., Carter, E., Boon, R., and Fisher, R. (2016). "The role of initial witness accounts within the investigative process," in *Communication in Investigative and Legal Contexts: Integrated Approaches From Forensic Psychology, Linguistics and Law Enforcement*, eds G. Oxburgh, T. Myklebust, T. Grant, and R. Milne (Chichester, UK: Wiley-Blackwell). doi: 10.1002/9781118769133.ch6

Gabbert, F., Hope, L., and Fisher, R. (2009). Protecting eyewitness evidence: examining the efficacy of a self-administered interview tool. *Law Hum. Behav.* 33, 298–307. doi: 10.1007/s10979-008-9146-8

Hanway, P. (2021). *The Effects of Cognitive Load for Investigative Interviewers*. Ph.D thesis, Portsmouth, UK: University of Portsmouth.

Hanway, P., and Akehurst, L. (2018). Voices from the front line: Police officers' perceptions of real-world interviewing with vulnerable witnesses. *Investig. Interv. Res. Pract.* 9, 14–33.

Hanway, P., Akehurst, L., Vernham, Z., and Hope, L. (2021). The effects of cognitive load during an investigative interviewing task on mock interviewers' recall of information. *Legal Criminol. Psychol.* 26, 25–41. doi: 10.1111/lcrp.12182

Heaton-Armstrong, A., and Wolchover, D. (1992). Recording witness statements. *Criminal Law Rev.* 1992, 160–172.

Home Office (2011). "Achieving Best Evidence in Criminal," in *Proceedings Guidance on Interviewing Victims and Witnesses, and Guidance on Using Special Measures*. (London: Home Office).

Hyman Gregory, A., Schreiber Compo, N., Vertefeuille, L., and Zambruski, G. (2011). A comparison of US police interviewers' notes with their subsequent reports. *J. Investig. Psychol. Offender Profiling* 8, 203–215. doi: 10.1002/jip.139

Johnson, M. K., Hashtroudi, S., and Lindsay, D. S. (1993). Source monitoring. *Psychol. Bull.* 114, 3–28. doi: 10.1037/0033-2909.114.1.3

Kleider-Offutt, H. M., Cavrak, S. E., and Knuycky, L. R. (2015). Do police officers' beliefs about emotional witnesses influence the questions they ask? *Appl. Cogn. Psychol.* 29, 314–319. doi: 10.1002/acp.3111

Kohnken, G., Thurer, C., and Zoberbier, D. (1994). The cognitive interview: Are interviewers' memories enhanced too? *Appl. Cogn. Psychol.* 8, 13–24. doi: 10.1002/acp.2350080103

Kontogianni, F., Hope, L., Taylor, P. J., and Gabbert, F. (2020). "Tell me more about this. . ." An examination of the efficacy of follow-up open-ended questions following an initial account. *Appl. Cogn. Psychol.* 34, 972–983. doi: 10.1002/acp.3675

Lamb, M. E., Orbach, Y., Sternberg, K. J., Hershkowitz, I., and Horowitz, D. (2000). Accuracy of investigators' verbatim notes of their forensic interviews with alleged child abuse victims. *Law Hum. Behav.* 24, 699–708. doi: 10.1023/A:1005556404636

Malsch, M., Kranendonk, R. P., De Keijser, J. W., Komter, M. L., De Boer, M., and Elffers, H. (2018). Reporting on police interrogations: Selection effects and bias related to the use of text, video and audiotape. *Investig. Interv.* 9, 61–76.

McLean, M. (1995). Quality investigation? Police interviewing of witnesses. *Med. Sci. Law* 35, 116–122. doi: 10.1177/002580249503500205

Meissner, C. A. (2021). What works? Systematic reviews and meta-analyses of the investigative interviewing research literature. *Appl. Cogn. Psychol.* 35, 322–328. doi: 10.1002/acp.3808

Milne, R., and Bull, R. (2006). "Interviewing victims of crime, including children and people with intellectual difficulties," in *Practical Psychology for Forensic Investigations and Prosecutions* eds M. R. Kebbell and G. M. Davies (Chichester: Wiley), 7–24. doi: 10.1002/9780470713389.ch1

Milne, R., and Bull, R. (2002). Back to basics: A componential analysis of the original cognitive interview mnemonics with three age groups. *Appl. Cogn. Psychol.* 16, 743–753. doi: 10.1002/acp.825

Milne, R., and Bull, R. (2016). "Witness interviews and crime investigation," in *An Introduction to Applied Cognitive Psychology*, eds D. Groome and M. W. Eysenck (London: Routledge), 175–196.

Milne, R., Griffiths, A., Clarke, C., and Dando, C. (2019). "The Cognitive Interview – a tiered approach in the real world," in *Evidence-Based Investigative Interviewing: Applying Cognitive Principles*, eds B. Schwartz, J. Dickenson, N. Schreiber Compo, and M. McCauley (London: Routledge London), 56–73. doi: 10.4324/9781315160276-4

Milne, R., and Shaw, G. (1999). Obtaining witness statements: The psychology, best practice and proposals for innovation. *Med. Sci. Law* 39, 127–137. doi: 10.1177/002580249903900207

Oxburgh, G. E., Myklebust, T., and Grant, T. (2010). The question of question types in police interviews: A review of the literature from a psychological and linguistic perspective. *Int J. Speech Lang Law* 17, 45–66. doi: 10.1558/ijsll.v17i1.45

Rock, F. (2001). The genesis of a witness statement. *Forensic Linguis.* 8, 44–72. doi: 10.1558/sll.2001.8.2.44

Sagana, A., Sauerland, M., and Merckelbach, H. (2017). Witnesses' failure to detect covert manipulations in their written statements. *J. Investigat. Psychol. Offender Profiling* 3, 320–331. doi: 10.1002/jip.1479

Shepherd, E., and Milne, R. (1999). *Full and Faithful: Ensuring Quality Practice and Integrity of Outcome in Witness Interviews. Analysing Witness Testimony: A Guide for Legal Practitioners and Other Professionals*. London: Blackstone Press Limited.

Smith, K., and Milne, R. (2018). Witness interview strategy for critical incidents (WISCI). *J. Forensic Pract.* 20, 268–278. doi: 10.1108/JFP-03-2018-0007

Vrij, A., Hope, L., and Fisher, R. P. (2014). Eliciting reliable information in investigative interviews. *Policy Insights Behav. Brain Sci.* 1, 129–136. doi: 10.1177/2372732214548592

Westera, N., Milne, R., and Kebbell, M. (2011). Interviewing witnesses will investigative and evidential requirements ever concord? *Br. J. Forensic Pract.* 13, 103–113. doi: 10.1108/14636641111134341

Westera, N. J., Powell, M. B., and Milne, B. (2017). Lost in the detail: Prosecutors' perceptions of the utility of video recorded police interviews as rape complainant evidence. *Aust. N. Z J. Criminol.* 50, 252–268. doi: 10.1177/0004865815620705

Zander, M., and Henderson, P. (1993). *Crown Court Study*. London: HMSO.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

PERSPECTIVE
published: 08 March 2022
doi: 10.3389/fcomm.2022.779434

The Benefits of a Jeffersonian Transcript

Song Hee Park[1] *and Alexa Hepburn*[2]*

[1]College of Medicine, Chung-Ang University, Seoul, South Korea, [2]Department of Communication, Rutgers University, New Brunswick, NJ, United States

Over the past 6 decades, researchers in conversation analysis have repeatedly shown that everyday social activities such as inviting a friend over, interviewing a police suspect, teaching a class, or cross-questioning in a courtroom–are achieved in orderly and reproducible ways. Jeffersonian transcription has been refined to both capture and crystallize the interactionally relevant specifics of how such tasks get done. Conversation analytic work has shown that by leaving out features like the timing of turns, and changes in prosody, volume and other vocal and embodied specifics of delivery, a standard orthographic transcript bleaches out crucial components of how humans perform discursive actions, and how they continuously analyze one another across sequences of talk. This short paper will overview some of the benefits of investing time in the Jeffersonian system. Rather than simply describing the system, we will illustrate the analytic usefulness of its systematic and detailed transcription practices; we show how transcription facilitates a clearer picture of how things get done in interaction.

Keywords: transcription, conversation analysis (CA), jeffersonian transcription, social interaction, transcription conventions

OPEN ACCESS

Edited by:
Emma Richardson,
Aston University, United Kingdom

Reviewed by:
Amanda Bateman,
University of Waikato, New Zealand
Peter Muntigl,
Simon Fraser University, Canada
Jessica Nina Lester,
Indiana University, United States

***Correspondence:**
Alexa Hepburn
alexa.hepburn@rutgers.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *18 September 2021*
Accepted: *10 January 2022*
Published: *08 March 2022*

Citation:
Park SH and Hepburn A (2022) The Benefits of a Jeffersonian Transcript. Front. Commun. 7:779434. doi: 10.3389/fcomm.2022.779434

1 INTRODUCTION

This article overviews the benefits of working with a Jeffersonian transcript for researchers whose data comprises any kind of talk-in-interaction. Our argument will be that standard orthographic transcripts wipe out core elements that speakers themselves incorporate in order to construct activities of various kinds. Details of delivery such as timing, speed, emphasis, pitch, and volume, as well as embodied elements such as gaze direction, frowning etc., all affect how the action being built in the moment will be heard and responded to–this is what conversation analysis aims to tap into. Hepburn and Bolden (2017) wrote the definitive book on how to do Jeffersonian transcription, but here we rehearse some of the arguments for why one should invest the effort into representing these details of talk.

Conversation Analysis (CA) is a multi-disciplinary field first developed by Harvey Sacks, Emanuel Schegloff, and Gail Jefferson. It is dedicated to exploring the fundamental communication processes that underpin human interaction. Many of the transcription conventions originally developed by Gail Jefferson in the 1960's are still in use today and comprise largely intuitive conventions, such as up and down arrows representing pitch changes, underlining for emphasis, and capital letters for increased volume. Since it was first developed, Jeffersonian transcription has evolved to represent various embodied features of actions such as gaze, facial expressions, and body positions (e.g. Goodwin 1981;

Mondada 2007) as well as non-speech sounds such as laughter and crying (e.g. Hepburn 2004). Transcripts are designed specifically to represent interactionally relevant changes in delivery that we all use to ground our understandings about one another, for example that someone is having trouble responding or conveying difficult news, or that they are upset, disappointed, or angry about something.

2 CONVERSATION ANALYTIC PERSPECTIVE

Potter and Hepburn (2012) showed how the process of transforming spoken words into a verbatim, or orthographic, transcript skates over the activities being performed by speakers when dealing with a challenging question. Similarly, Hepburn and Bolden (2017) provided a simple illustration showing how a speaker's acknowledgement of having heard the question was misunderstood in journalistic outputs. Here we offer a similar illustration, taking a clip from a Senate Judiciary Committee hearing. This clip shows Rachel Mitchell, head of the Special Victims Division in Maricopa County, Arizona, and Brett Kavanaugh, Supreme Court nominee, who was providing testimony regarding allegations that he assaulted Christine Blasey Ford while the two were teenagers. Mitchell was hired by the Republicans to question Kavanaugh. First, we show the basic transcript as it appeared on various journalistic sites. Then we illustrate what more can be made of this piece of interaction by deploying a Jeffersonian transcript designed to facilitate a conversation analysis. The Jeffersonian transcript was created using original video footage from the Committee hearing recording.

1. C-SPAN Kavanagh-Blasey Ford 59.05–1.00.17 Orthogonal transcript

MITCHELL: Have you ever passed out from drinking?

KAVANAUGH: Passed out would be no but I've gone to sleep. I've never blacked out. That's the allegation and that's wrong.

MITCHELL: So let's talk about your time in high school. In high school, after drinking, did you ever wake up in a different location than you remembered passing out or going to sleep?

KAVANAUGH: No, no.

MITCHELL: Did you ever wake up with your clothes in a different condition, or fewer clothes on than you remembered when you went to sleep or passed out?

KAVANAUGH: No, no.

MITCHELL: Did you ever tell—did anyone ever tell you about something that happened in your presence that you didn't remember during a time that you had been drinking?

KAVANAUGH: No, the we drank beer, and you know, so did I think the vast majority of people our age at the time. But in any event, we drank beer, and still do. So whatever, yeah.

```
   2. C-SPAN Kavanagh-Blasey Ford 59.05-1.00.17 Jeffersonian transcript
01 MIT:     Have you e:ve:r passed out from drinking.
02          (.)
03 KAV:     Tch .hhh I- w- thwu- (0.2) >Passed out wou'be< (.) ↓neo:
04          but I've gone to slee:p,=<uh but (0.2) I've never blacked
05          ↓out,=That's the— (0.2) that's the— (0.2) the- (0.2)
06          allegation¿ (0.2) .hhh (.) uh:: (0.3) a:n:d (0.8)
07          uh: (0.3) thut— (0.2) that— (.) that's wrong,
08          (0.5)
09 MIT:     Tk .Hhh So let's talk about your time in high school. (0.5)
10          In high school after drinking, did you ever wake up in
11          a different location than you remembered
12          (0.5)/((hand gesture))[passing out or going to sleep.]
13 Kav:                           [     (( head shake ))         ]
14 KAV:     No.
15          (0.5)
16 KAV:     No.
17          (0.7)
18 MIT:     Did you ever wake up with (0.7)/((circular hand gesture))
19          your clothes in a different condition,=or fewer
20          clothes o:n: .hhh than you remembered
21          [when you went to sleep, or passed #out.]
22 Kav:     [     (( head shake ))                  ]
23          (1.5)/((Kav frowns, open hand gesture))
24 KAV:     ((frowning, gaze to MIT)) ↑No.=Ye- (0.4) No.
25          (0.6)/((Kav gaze down))
26 KAV:     ((Kav gaze to Mit, smile/grimace)) Hh-hh-hh-((breathy laugh))
27          (0.2)
28 MIT:     Did you e↑ver tell: (0.2) #uh (0.2) s:=uh: (0.3) Did (.)
29          >anyone ever< tell you about something that happened in
30          your presence, .hh [that you didn't re]#member. =#°in-=
31 Kav:                        [ ((disgust face)) ]
32 MIT:     =u-j- during a time that you had been drinking.
33 KAV:     Tch N- (0.4) No.
34          (.)
35 KAV:     The— (0.4) the— (0.5) We drank bee:r, a:n:d (2.7)
36          s- #°y'know (0.3) so— >so did I think< (0.3)
37          the vast majority of— (0.3) of people our age at
38          the ti:me=But in any event we drank bee:r, and (0.2)
39          .Hhh/((disgust face)) (0.5) and uh: (0.7) still ↑do.
40          So whatever: (0.4) .Hhnh y- yeah.
```

We can straight away see that the second transcript is both three times longer, and harder to read for non-conversation analysts. It has numbered lines, includes specifics of timing and delivery, some interactionally relevant visual details such as gestures, gaze direction, and facial expressions, and is given a non-proportional font (e.g., Courier). Why add in all this detail? While space does not permit a full answer to this question (see Hepburn and Bolden 2017, for more detail), below we show some of the more obvious elements that an orthographic transcript misses.

Examining Kavanaugh's answers using only the orthographic transcript makes him sound like he has no trouble with the questions put to him. However, this skates over the halting way that Kavanaugh delivers his responses. For example, in the Jeffersonian transcript, line three contains a tut particle "Tch," an inbreath ".hhh," several false starts "I-w-thwu-" and a timed pause (0.2). On lines 5-7, rather than "that's the allegation and that's wrong" we can see something that looks much less definitive: "That's the— (0.2) that's the— (0.2) the- (0.2) allegation¿ (0.2).hhh (.) uh:: (0.3) a:n:d (0.8) uh: (0.3) thut— (0.2) that— (.) that's wrong," Again there are many false starts, with a great deal of pausing between them, which are common occurrences in the doing of "hesitation" or "delicacy" (see Lerner 2013).

Closer attention to the detail of the question design and response also raises some important issues. Heritage and Raymond (2021), have argued that polar (or yes/no) questions like these unavoidably incorporate within their main proposition the un/likelihood of some state of affairs, thereby creating the conditions for (or setting up a "preference" for in conversation analytic terms) a positive or negative response. For example, "have you ever x" questions encode that there is little likelihood of "x" happening (note Mitchell was chosen by Republicans). It is interesting to note that there is emphasis (shown by underlining) and stretched delivery (shown by the colons) on "ever"–the 'negative polarity' item itself (Heritage 2002; see also; Raymond and Heritage, 2021)–perhaps adding further to the improbability of such an event.

This negative polarity design is continued in Mitchell's further question on lines 9–12: "did you ever wake up in a different location than you remembered (0.5)/((hand gesture))passing out or going to sleep." Again there is emphasis on "ever," shown by the underlining. It is interesting to note that although Kavanaugh has denied passing out, Mitchell has included it in the question, after some silence, accompanied by a kind of circling hand gesture. In response, Kavanaugh is unusually definitive. He is primed with a negative response, shown by his head shake in overlap with the end of Mitchell's turn (indicated by the lining up of square brackets across lines 12 and 13–one reason why a non-proportional font is important). Note that following the first "No." on line 14 (the turn final period showing falling intonation, one common way of indicating turn completion), there is a gap on line 15–such silences are shown on their own line to indicate that a new speaker could have taken a turn at this point. Kavanaugh repeats the "No." indicating that he has nothing more to add here. The presupposition in the question–that Kavanaugh might have passed out–goes unchallenged.

We can contrast these definitive responses with Kavanaugh's response following Mitchell's third question on lines 18–21: "Did you ever wake up with (0.7)/((circular hand gesture)) your clothes in a different condition, =or fewer clothes o:n:.hhh than you remembered when you went to sleep, or passed #out." Although once again the design of the question primes Kavanaugh to say 'no' (again he is shaking his head in overlap towards the end of Mitchell's turn), he nevertheless spends some time frowning and gazing to and from Mitchell before responding "↑No.=Ye- (0.4) No.," followed by gazing downwards, and gazing back to Mitchell doing breathy and tense clenched jaw laughter 'Hh-hh-hh-' (see Hepburn and Varney 2013 on different types of laughter). Jefferson's transcription conventions have given rise to some important research on the interactional role of laughter. Jefferson (1979) noted laughter's role in inviting recipient laughter, and as Shaw et al. (2013) noted, when laughter is in turn final position, as Kavanaugh's is, it can also have a proactive role in managing recipient responses, e.g., to encourage the overhearing audience to affiliate with and not take seriously his evident trouble with the question.

Rather than pursuing Kavanaugh's equivocal response to the question, Mitchell continues with her final question on lines 28–32, which gets a further element of negative polarity: "did anyone ever tell you." Leaving aside the oddness of a question about whether Kavanaugh remembers about "something that happened in your presence,.hh [that you didn't re]#member." we can note that Kavanaugh's response is again less definitive than it might appear on an orthographic transcript. After a delayed "No," he proceeds to account for his answer on lines 35–40–CA findings would predict that this is an unusual thing to do unless one's response is counter to the preference encoded in the question.

3 DISCUSSION

In sum, we have shown some of the interactional relevance of adding in elements of speech delivery and timing, as well as some basic visual information. The detailed elements of interaction included in the Jeffersonian transcript allow a more nuanced and comprehensive understanding of what participants actually say and do. In the example above, we saw how the emphatic delivery of particular words (e.g., negative polarity items) may prime the respondent in specific ways to answer. Furthermore, the disfluencies (e.g. pauses, false starts) displayed in the talk indicate that the speaker is having difficulty in conveying a clear position. These interactional features are something that can only be captured in the Jeffersonian transcript, being crucial resources for understanding what actually happened during the hearing. Our argument is that, in order to understand what is accomplished interactionally, we should transcribe not just what people say but how and when they say it.

Jeffersonian transcription not only helps conversation analysts examine the social world "as it is" but also allows a wide range of readers to see things that happened. The aim of Jeffersonian transcription has always been to make the transcripts "accessible to linguistically unsophisticated readers" (Sacks et al., 1974, p. 734), with the details added in for an accurate representation of the interactional process. While it is true that readers may find it difficult to follow complex and detailed transcripts (Hammersley 2010), adding in the relevant details is imperative because, as CA studies have convincingly demonstrated, they are what participants find consequential. The details captured in the Jeffersonian transcripts are those that are oriented to by participants themselves and are relevant to the ongoing interaction. The Jeffersonian system, therefore, can be found useful by social scientists, practitioners, clients, policy makers, professionals, and laypersons as it enables a close examination of how things are done in everyday social interaction.

Social institutions may benefit from consulting conversation analysts to determine important features of interactions. Training may be needed to focus on participants' orientations in the ongoing interaction. Especially when the subtle specifics of interactional display can change the meaning of what is being done in significant ways, as in our example above, producing detailed accurate transcripts is critical. Some exercises accompany Hepburn and Bolden's book on how to do Jeffersonian transcription, available via this link: https://rucal.rutgers.edu/transcription/.

Journalists should be careful in representing what was said and done in the interactional event they describe so as not to omit features that are fundamental to understanding what happened. Furthermore, a careful transcription can help institutions (e.g., helpline services) identify and promote good practice (e.g. Hepburn 2006; Hepburn et al., 2014). As practitioners engage with recordings and transcripts on their own, and attend to various features of talk, they can better understand the practices they use every day and what makes them "good" and "bad" practices.

REFERENCES

Goodwin, C. (1981). *Conversational Organization: Interaction between Speakers and Hearers*. New York: Academic Press.

Hammersley, M. (2010). Reproducing or Constructing? Some Questions about Transcription in Social Research. *Qual. Res.* 10 (5), 553–569. doi:10.1177/1468794110375230

Hepburn, A., and Bolden, G. B. (2017). *Transcribing for Social Research*. Thousand Oaks, CA: Sage.

Hepburn, A. (2004). Crying: Notes on Description, Transcription, and Interaction. *Res. Lang. Soc. Interaction* 37 (3), 251–290. doi:10.1207/s15327973rlsi3703_1

Hepburn, A. (2006). Getting Closer at a Distance: Theory and the Contingencies of Practice. *Theor. Psychol.* 16 (3), 325–342. doi:10.1177/0959354306064282

Hepburn, A., and Varney, S. (2013). "Beyond ((laughter)): Some Notes on Transcription," in *Studies in Laughter in Interaction*. Editors P. J. Glenn, and E. Holt (London: Bloomsbury), 25–38.

Hepburn, A., Wilkinson, S., and Butler, C. W. (2014). Intervening With Conversation Analysis in Telephone Helpline Services: Strategies to Improve Effectiveness. *Res. Lang. Soc. Interaction* 47 (3), 239–254. doi:10.1080/08351813.2014.925661

Heritage, J. (2002). "Designing Questions and Setting Agendas in the News Interview," in *Studies in Language and Social Interaction*. Editors P. Glenn, C. LeBaron, and J. Mandelbaum (Hillsdale: Lawrence Erlbaum Associates), 57–90.

Heritage, J., and Raymond, C. W. (2021). Preference and Polarity: Epistemic Stance in Question Design. *Res. Lang. Soc. Interaction* 54 (1), 39–59. doi:10.1080/08351813.2020.1864155

Jefferson, G. (1979). "A Technique for Inviting Laughter and its Subsequent Acceptance/declination," in *Everyday Language: Studies in Ethnomethodology*. Editor G. Psathas (New York: Irvington), 79–96.

Lerner, G. (2013). "On the Place of Hesitating in Delicate Formulations: A Turn-Constructional Infrastructure for Collaborative Indiscretion," in *Conversational Repair and Human Understanding*. Editors M. Hayashi, G. Raymond, and J. Sidnell (Cambridge: Cambridge University Press), 95–134.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

AH and SP contributed to conception and design of the study. AH organized the data and wrote the first draft of the manuscript. SP wrote the discussion section of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Mondada, L. (2007). Multimodal Resources for Turn-Taking. *Discourse Stud.* 9 (2), 194–225. doi:10.1177/1461445607075346

Potter, J., and Hepburn, A. (2012). "Eight Challenges for Interview Researchers," in *Handbook of Interview Research*. Editors J. F. Gubrium, and J. A. Holstein. 2nd ed. (London: Sage), 555–570.

Raymond, C. W., and Heritage, J. (2021). Probability and Valence: Two Preferences in the Design of Polar Questions and Their Management. *Res. Lang. Soc. Interaction* 54 (1), 60–79. doi:10.1080/08351813.2020.1864156

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language.* 50, 696–735. doi:10.1353/lan.1974.0010

Shaw, C., Hepburn, A., and Potter, J. (2013). Having the Last Laugh: On post-completion Laughter Particles. *Studies of Laughter in Interaction*. Editors P. Glenn and E. Holt (London: Bloomsbury), 91–106.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 10 May 2022
doi: 10.3389/fcomm.2022.816461

The Influence of Police Reporting Styles on the Processing of Crime Related Information

Anita Eerland[1] *and Tessa van Charldorp*[2*]

[1] *Department of Communication Science, Behavioural Science Institute, Radboud University, Nijmegen, Netherlands,*
[2] *Department of Languages, Literature, and Communication, Utrecht Institute of Linguistics OTS, Utrecht University, Utrecht, Netherlands*

Police records drawn up during or after a suspect's police interrogation play a crucial role in judicial systems and should therefore be factual representations of what occurred in the spoken interrogation. Within the judicial domain, however, little is known about how style of reporting (i.e., the specific language used) affects the interpretation of these facts. Furthermore, the relationship between police record 'quality' and variations in judgment of guilt, credibility or reliability has not been studied to date. In three studies, we investigated the influence of three commonly used recording styles (i.e., monolog, recontextualized and question-answer style) on judgments of guilt, credibility, and reliability in fictitious criminal cases. We hypothesized that participants would (1) find records in the question-answer style more credible and reliable than those in the monolog or recontextualized style, and (2) consider the recontextualized style to be the least credible and reliable. Experiment 1 showed that the Q&A style was perceived as more reliable than the other two styles. Experiment 2, a replication in which we also tested new hypotheses based on explorative analyses of Experiment 1, showed no effects of reporting style. To investigate whether the discrepancy in results was due to different scenarios, a third experiment that made use of multiple scenarios was conducted. We found effects of reporting style on perceived accuracy, imageability, and understandability. In sum, this study showed that factors as subtle as reporting style might impact the processing of information in contexts where only factual information should be taken into account.

Keywords: language comprehension, recording styles, judgments, police records, linguistic cues

OPEN ACCESS

Edited by:
Debbie Loakes,
The University of Melbourne, Australia

Reviewed by:
Steve Oswald,
Université de Fribourg, Switzerland
Anna Hatzidaki,
National and Kapodistrian University of Athens, Greece

****Correspondence:***
Tessa van Charldorp
t.c.vancharldorp@uu.nl

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *16 November 2021*
Accepted: *05 April 2022*
Published: *10 May 2022*

Citation:
Eerland A and van Charldorp T (2022) The Influence of Police Reporting Styles on the Processing of Crime Related Information. Front. Commun. 7:816461. doi: 10.3389/fcomm.2022.816461

INTRODUCTION

To understand language, people form mental representations of a described situation. This mental representation is known as a situation model (Johnson-Laird, 1983; Van Dijk and Kintsch, 1983; Morrow et al., 1987; Zwaan and Radvansky, 1998). Differences in the specific language used to convey a message are known to influence situation model construction. Perhaps the best-known example of research that supports this idea is the study by Loftus and Palmer (1974). In this study, participants first watched a video clip of a car crash and then answered questions about this clip. Some participants were asked to estimate the speed of the cars when they "hit" each other, others were asked to estimate the speed of the cars when they "smashed into" each other. Results showed, among other things, that the specific wording that was used to frame the question influenced the speed estimates of the participants. Participants that were asked the estimation question using the word 'hit' thought the cars drove significantly slower than those who were asked the same question using "smashed into." One could argue that the difference between "hit" and "smashed into" is not a very subtle one. As a result, it is hardly surprising that the wording chosen to formulate a question impacts the answer given (i.e., the estimated speed of the cars). Later research, however, demonstrated that even more subtle differences in language use, known as linguistic cues, impact situation model construction and as a result influence how people think of a described situation (Givón, 1992; Magliano and Schleich, 2000). While there is a large variety of linguistic cues that could impact situation model construction, we will discuss two of these cues in more detail.

The first linguistic cue we will discuss is grammatical aspect. Grammatical aspect, more specifically the difference between the imperfective (e.g., *He was shooting a gun*) and the perfect/perfective aspect (e.g., *He had shot a gun/He shot a gun*), primes semantic knowledge associated with the described event, like location information (Carreiras et al., 1997; Ferretti et al., 2007; Ferreti et al., 2009; Anderson et al., 2013). For example, Ferretti et al. (2007) showed that participants were faster at naming a location after having read a verb conveyed in the imperfective aspect (e.g., *was skating*) than the perfective aspect (e.g., *skated*). Grammatical aspect is not only known to influence the construction of a situation model but also cognitive processes that rely on situation model information, like memory (Carreiras et al., 1997; Magliano and Schleich, 2000), problem solving (Salomon et al., 2013), and voting behavior (Fausey and Matlock, 2011).

Secondly, speech type, more specifically the use of direct speech (e.g., *The witness said: 'I saw the attacker entering the building'*) vs. indirect speech (e.g., *The witness said that she saw the attacker entering the building*), is also known to impact situation model construction (Yao and Scheepers, 2011; Yao et al., 2011; Stites et al., 2013; Eerland and Zwaan, 2018). For example, the use of direct speech results in a more vivid situation model and is as more perceptually engaging than indirect speech. In sum, the ways in which we formulate events matters with regard to how this information is processed and remembered. That can be problematic in contexts in which language is used as an objective means to record—on paper—what has taken place during a spoken interaction, like a police interrogation.

During or after a police interrogation, police records are constructed. These records—or written statements—play an important role in the judicial system as they can be used as evidence in court if they adhere to certain criteria. Therefore, it is important that police records contain information that is accurate and of high quality. The need for accuracy and quality of these reports has been discussed in the judicial and police context (Malsch et al., 2010; Gregory et al., 2011; Jansen, 2011).

However, while the importance of accurate police records seems obvious, guidance on *how* to write up police records has only recently received more attention, for example in the form of handbooks and training. Often, police officers have relative freedom as to how they produce a written document that is supposed to reflect the spoken interrogation (see De Keijser et al., 2012). In different countries or judicial systems, we also see different recording systems: whereas in some judicial systems the recording of the police interrogation is typed up (or transcribed) verbatim afterwards (i.e., UK), other judicial systems require police officers to type up a police record while interrogating (i.e., the Netherlands).

Yu and Monas (2020) provide a brief overview of current literature on report writing in which they conclude that interview techniques and note taking are prioritized over actually *how to write* the police record. This finding can be confirmed by looking at various handbooks for police officers in which they are trained to interrogate and write a report (e.g., Schellingen and Scholten, 2014). There are exceptions where police officers are elaborately trained on how to write objectively, in a structured way, etc. (e.g., Reynolds, 2012; Miller and Whitehead, 2018). From the judicial and police training perspective there seems to be a focus on being accurate, objective and providing a step-by-step account (Reynolds, 2012), or on being accurate, objective, complete, concise and clear (Morley, 2008).

Although these instructions are helpful, in most cases they remain rather vague as to how to operationalize for example accuracy, conciseness, or objectivity in actual language. Furthermore, the impact of choosing certain linguistic characteristics over others is relatively unexplored and unattended to. For example, in an important study carried out by the Dutch police academy on judicial knowledge and police records, the author concluded with suggestions for improvement focusing on teaching police officers more judicial knowledge and judicial language (Jansen, 2011). Whereas, more judicial knowledge and language in a police record may be evaluated as qualitatively better according to judges and prosecutors—it tells us little about whether this actually improves the accuracy, comprehensibility, objectivity, or conciseness of the text. Furthermore, it seems to contradict the guidance from—for example—the Dutch law stating that a police record must be written as much as possible in the suspect's own words. Lastly, using more judicial language could possibly have other effects, such as on judgment.

Before we suggest how police records could be qualitatively improved, we need to have a much clearer understanding of how language use within the context of police records can affect judgments. In this study, rather than theorizing about what linguistic aspects could lead to what kinds of differences in quality or judgment, we take a bottom-up approach to see what kinds of linguistic aspects are prevalent in actual police records. Based on a corpus of 35 actual police records Van Charldorp (2011) found that there are three main linguistic styles used by Dutch police officers: the monolog style, the recontextualized style and the question-answer style. These styles make use of various linguistic constructions such as *perspective* and *visibility of the source*. Besides the restrictions that the computer format and the law provide, an officer is free to write up the suspect's story in either one of these three styles, or a combination thereof.

The *monolog style* is written from the perspective of the suspect (in first person, using direct speech). The questions asked or remarks made by the interrogator(s) do not appear in the record. This style is relatively informal, not too lengthy and comprehensible. An example is the following:

> "I have heard and understood that I am not obliged to answer. I am 14 years old. I live with my father. My father's name is Steven Pinas. My mother passed away."

In the *recontextualized style* the officer's question is told from the perspective of the suspect using indirect speech. In other words, the "interrogator reworks his own questions, remarks, or suggestions into the narrative" (Komter, 2013) while still using the first-person perspective. By doing so, the officers "ensure their visibility" in the police records (Komter, 2013). This style is often lengthy, formal and somewhat complicated. An example is the following:

> "You ask me which act I performed.
> I state to you that I kept the garage door closed and I also helped pull Mervellino into the garage.
> You ask me, if a knife was used.
> I state to you that no knife was used. At least, not that I know."

In the *question-answer (Q&A) style* the police officers' questions and the suspect's answers are written up as such, generally written as "Q" and "A" or "Question: ... " and "Answer: ... ". As a result, direct speech is used. The Q&A style is short, relatively informal and comprehensible. It most closely resembles the actual spoken interaction that occurred during the police interrogation. An example is the following:

> "Question: In the living room we also encountered drug-resembling goods, whose are these?
> Answer: I don't know, those are not mine.
> Question: Have you ever seen your uncle with a fire weapon?
> Answer: No."

As can be concluded from the description above, there are different dimensions that distinguish the reporting styles from each other. First of all, there is the number of *sources* that provide the information mentioned in the police record. Whereas, the recontextualized and the Q&A style include information from both the police officer and the suspect, the monolog style only states information provided by the suspect. Next, the styles differ when it comes to *perspective* taken in the record. The recontextualized and monolog style records only present the perspective of the suspect, whereas the Q&A style records represent the perspective of both the police officer and the suspect. One could derive a score from both dimensions as indication of the record's representativeness of the interrogation that took place (see **Table 1**).

TABLE 1 | Overview of the number of sources represented and the perspective offered by reports using various recording styles.

	Sources	Perspectives	Representativeness
Recontextualized	2 *(officer, suspect)*	1 *(suspect)*	3
Question-answer	2 *(officer, suspect)*	2 *(officer, suspect)*	4
Monolog	1 *(suspect)*	1 *(suspect)*	2

Representativeness is the sum of the number of sources and perspectives presented.

In this study we will explore how the above mentioned different linguistic reporting styles influence reader judgments concerning reliability (i.e., accuracy) and credibility (i.e., believability) of police records and the interrogated suspects. If linguistic style affects reader judgments, much clearer guidelines will be necessary for police officers on how to construct a written police record in order to most accurately institutionalize a suspect's spoken words.

Our predictions of how reporting style influences credibility and reliability judgments are based on (1) the representativeness score for each reporting style, and (2) how common each style is. Given that the Q&A style best represents the actual interrogation (see **Table 1**) and this format is commonly used in everyday discourse, we expected a Q&A style record to be perceived as more credible and reliable than records that use the recontextualized or monolog style. The recontextualized style has a higher representativeness score than the monolog style. However, because the recontextualized style is the most complex (and deviates the most from everyday language), we hypothesized that this style leads to the least credibility and reliability of the record. The credibility and reliability of the record reported in the monolog style is expected to be lower than that of the Q&A style but higher than that of the recontextualized style. Analyses regarding the credibility and reliability of the suspect will be exploratory. Our preregistration, materials, and data can be found on the following project page on the Open Science Framework: https://osf.io/fpgz5/.

EXPERIMENT 1

Methods

Sample

We aimed at 100 participants per condition. Therefore, we recruited 350 participants online through Amazon's Mechanical Turk (MTurk, http://www.mturk.com) and 352 completed the experiment (this can occur most likely due to technical issues involving the coordination of the platform we used for recruitment, MTurk, and the platform we used for running the experiment, Qualtrics). We excluded data from 28 participants because they had reading times <0.05 ms per word. This indicates that they could not have read the police report properly. We also excluded data from participants that did not report English as their native language ($n = 3$), indicated they found the report extremely difficult to understand ($n = 2$), found the report extremely difficult to understand and had reading times <0.05 ms per word ($n = 2$), or did not indicate their native language ($n = 1$). Because exclusion of these participants yielded unequal lists, we also removed the data from five last run participants in the recontextualized condition, and two last run participants in the Q&A condition. The remaining sample ($N = 309$; 103 participants per condition) had a mean age of 38.37 [$SD = 11.25$, range = 21–71, 172 (55.66%) females]. All participants were US residents and received $1 for their participation (that took ∼9 min).

Materials and Procedure

We selected an authentic police report of the interrogation of a man being suspected of stealing a motorcycle for this experiment. We considered this case to be useful for our study, as the crime involved is moderately severe and the evidence presented could be interpreted as incriminating as well as exculpatory. This was done to prevent any ceiling or floor effects for the guilty judgments as these would make it more difficult to investigate how these judgments might be impacted by reporting style. Importantly, the suspect does not confess to the crime. The report was originally recorded in the recontextualized style in Dutch. We translated the report to English and we created two additional versions of the original report; one using the monolog style and one using the Q&A style. All three reports were checked by two native speakers of English.

Participants were randomly assigned to one of three reporting style conditions (i.e., monolog, Q&A, and recontextualized). After participants carefully read the police report we asked them (1) how easy or difficult it was to understand the police record (7-point scale, 1 = *extremely easy*, 7 = *extremely difficult*), (2) if they thought the suspect was guilty of the crime (stealing a motorcycle; *yes/no*), and (3) how confident they were about their judgment (7-point scale, 1 = *not at all*, 7 = *extremely*). In addition, participants indicated how credible (7-point scale, 1 = *extremely credible*, 7 = *extremely uncredible*) and reliable (7-point scale, 1 = *extremely reliable*, 7 = *extremely unreliable*) they thought the record and the suspect were. Finally, participants stated what they thought this study was about and provided some demographic information. We recorded the time people spent reading the report. This task was presented online in the Qualtrics survey research suite (http://www.qualtrics.com).

Results

To test whether recording style influenced credibility and reliability judgments, we conducted a one-way ANOVA with recording style as between subjects factor and credibility and reliability scores for the police record as dependent variables (see **Table 2**)[1]. We found that recording style impacted the perceived reliability of the police record [$F_{(2,306)} = 3.480$, $p = 0.03$, $\eta^2 = 0.022$] but not its credibility [$F_{(2,306)} = 2.775$, $p = 0.06$, $\eta^2 = 0.018$]. *Post-hoc* comparisons, using a Bonferroni correction, showed that police records written in the Q&A style were perceived as more reliable (as indicated by a lower score for unreliability) than those written in the recontextualized style. The perceived reliability of police records in the monolog style did not differ from that of records in the other two styles.

In addition, we conducted an ANOVA with recording style as between subjects factor and (1) understandability of the report, (2) reliability and (3) credibility scores for the suspect, (4) judgments of the suspect, and (5) judgment confidence as dependent variables. These exploratory analyses seem to suggest that recording style might also impact reliability [$F_{(2,306)} = 3.490$, $p = 0.03$, $\eta^2 = 0.022$] and credibility [$F_{(2,306)} = 4.774$, $p < 0.01$, $\eta^2 = 0.030$] judgments of the suspect (see **Table 2**). *Post-hoc* comparisons, using a Bonferroni correction, revealed that participants thought the suspect to be more reliable after reading the police record written in the monolog style than the recontextualized style. Perceived reliability of the suspect in these two conditions did not differ significantly from that in the Q&A condition. Regarding the perceived credibility of the suspect, we found a similar pattern. Participants considered the suspect to be more credible when they read the police record in the monolog style than in the Q&A or recontextualized style. Recording style did not seem to impact the perceived understandability of the report [$F_{(2,306)} = 2.489$, $p = 0.085$, $\eta^2 = 0.016$], or the confidence of participants regarding their judgment of the suspect [$F_{(2,306)} = 0.474$, $p = 0.623$, $\eta^2 = 0.003$]. An exploratory Chi square analysis suggested that reporting style did not impact the likelihood of a guilty judgment of the suspect ($\chi^2 = 0.63$, $p = 0.73$, Cramer's $V = 0.05$; see **Table 3**).

TABLE 2 | Mean (SE) scores per recording style in experiment 1 ($N = 309$).

Measures	Monolog ($n = 103$)	Question-answer ($n = 103$)	Recontextualized ($n = 103$)
Report			
Understandability	1.84 (0.11)	1.54 (0.09)	1.81 (0.11)
Credibility	2.35 (0.12)	2.15 (0.09)	2.52 (0.13)
Reliability	2.51 (0.13)	2.20 (0.09)^a	2.65 (0.14)^b
Suspect			
Credibility	2.99 (0.14)^a	3.51 (0.16)^b	3.60 (0.15)^b
Reliability	3.13 (0.14)^a	3.58 (0.16)	3.62 (0.14)^b
Confidence	4.97 (0.14)	4.93 (0.12)	4.80 (0.13)

Confidence = reported confidence in judgements regarding suspect. Dependent variables were measured on a 7-point scale. Different superscripts indicate a significant (p < 0.05) difference.

TABLE 3 | Percentage of guilty judgments of the suspect per recording style in experiment 1 ($N = 309$).

	Recording Style			Total
	Monolog ($n = 103$)	Question-answer ($n = 103$)	Recontextualized ($n = 103$)	
Guilty judgments (%)	13.59	17.48	16.50	15.86

Discussion

Based on the number of sources represented and the number of perspectives presented we calculated a representativeness score for all three recording styles under investigation. We expected the recording style with the highest representativeness score, the Q&A style, to be perceived as more credible and reliable than police records that used either the recontextualized or the

[1] Please note that we preregistered the following: "For each participant we will use the two credibility scores (police record and suspect) and the two reliability scores (police record and suspect). We will perform a 2 (credibility) × 2 (reliability) ANOVA to examine if these scores differ across conditions." A 2 × 2 ANOVA, however, makes no sense, as we are interested in comparing these four outcome measures across the three reporting styles. Therefore, we deviated from our preregistered plan.

monolog recording style. We expected the recontextualized style to be perceived as the least credible and reliable as this style is the most complex (and deviates the most from everyday language). Our hypothesis was partly supported by our data. We found that a Q&A recording style was perceived as more reliable, but not credible, than a recontextualized recording style and not a monolog recording style.

Although we expected the recontextualized style to be the most complex, our data suggest that recording style does not impact understandability. In other words, participants did not seem to perceive the recontextualized style to be more difficult to understand than the Q&A or monolog style. Recording style also does not seem to impact the perceived guilt of the suspect. Interestingly, our results seem to suggest that recording style impacts the perceived reliability and credibility of the suspect. As these analyses were exploratory in nature, we conducted a second experiment to test our newly generated hypotheses.

EXPERIMENT 2

Experiment 2 served as a conceptual replication study of Experiment 1. We used a comparable case (i.e., a robbery, no confession by the suspect, original report in the recontextualized style). Based on the results of Experiment 1, we hypothesized that recording style would impact the reliability of the police record with the record in the Q&A style perceived as more reliable than the record in the recontextualized style. We also expected the recording style to impact the reliability and credibility of the suspect, with the suspect being perceived as most reliable and credible when the police record was written in the monolog style. We did not expect recording style to impact understandability of the record, guilty judgments of the suspect, or confidence ratings with respect to this judgment.

Methods

Sample

We used the program G*Power (Faul et al., 2007) to conduct a power analysis based on the effect sizes found in Experiment 1. According to this power analysis we needed at least 495 participants (i.e., 165 per condition) to obtain statistical power at the recommended 0.80 level (Cohen, 1988). Therefore, we recruited 600 participants online through MTurk. Again, most likely to technical issues, 608 participants completed the experiment. We excluded data from 38 participants because they had reading times <0.05 ms per word. This indicates that they could not have read the police report properly. We also excluded data from participants that indicated they found the report extremely hard to understand ($n = 10$), or did not report English as their native language ($n = 12$). Because exclusion of these participants yielded unequal lists, we also removed the data from 13 last run participants in the recontextualized condition, and one last run participant in the Q&A condition. The remaining sample ($N = 534$; 178 per condition) had a mean age of 37.90 ($SD = 11.65$, range = 19–74, 279 [52.25%] females). All participants were US residents and received $1 for their participation (that took ~7 min).

Materials and Procedure

For this conceptual replication we selected another real police report of the interrogation of a suspect. This time, we selected a case in which a man was suspected of a robbery. Again, the suspect did not confess, and the information presented could be perceived as incriminating as well as exculpatory. The original report was written in the recontextualized style. We translated the original report from Dutch to English, and also created a monolog and Q&A style version of the translated original report. All versions were checked by two native speakers of English. We then followed the same procedure as in Experiment 1.

Results

To test whether recording style influenced credibility and reliability judgments, we conducted a one-way ANOVA with recording style as between subjects factor and credibility and reliability scores for the police record as well as the suspect as dependent variables (see **Table 4**). Contrary to our hypothesis and the results of Experiment 1, we found no impact of recording style on the perceived reliability of the police record [$F_{(2,531)} = 0.771$, $p = 0.46$, $\eta^2 = 0.003$], its perceived credibility [$F_{(2,531)} = 0.867$, $p = 0.42$, $\eta^2 = 0.003$], the perceived reliability of the suspect [$F_{(2,531)} = 0.468$, $p = 0.63$, $\eta^2 = 0.002$], and his perceived credibility [$F_{(2,531)} = 0.028$, $p = 0.97$, $\eta^2 < 0.001$]. As in Experiment 1, we did not find support for the idea that recording style influences the understandability of the record [$F_{(2,531)} = 0.637$, $p = 0.529$, $\eta^2 = 0.002$], guilty judgments ($\chi^2 = 0.76$, $p = 0.68$, Cramer's $V = 0.04$; see **Table 5**), or confidence regarding these judgments [$F_{(2,531)} = 1.230$, $p = 0.293$, $\eta^2 = 0.005$].

Discussion

Our conceptual replication of Experiment 1 yielded some interesting findings. Contrary to our expectations, we found no effects of recording style on reliability and credibility judgments of the police record or the suspect. The finding that recording style did not influence the understandability of the record, judgments regarding the guilt of the suspect, and the confidence with which these judgments were made confirmed the hypotheses generated through exploratory analyses of the data collected in Experiment 1.

In an attempt to explain why our results of Experiment 1 regarding the credibility and reliability of the police record and the suspect did not replicate, we looked more closely at the materials that we used. After all, we only used one scenario in each experiment. Although we controlled for some factors (e.g., type of crime, whether the suspect confessed or not, the style of the original police record), it might be that the scenarios we used differed in other ways. Any difference between our two scenarios might therefore (partly) explain why our experiments show different results. We looked specifically at the understandability of the case and the percentage of guilty judgments per experiment and over conditions (i.e., our results did not indicate a difference between conditions regarding the understandability and guilty judgments within experiments). It seems like the case used in Experiment 1 was easier to understand ($M = 1.73$, $SD = 1.04$) than the case used in Experiment 2 ($M = 2.96$, $SD = 1.48$). Also, we found far more guilty judgments in Experiment 2 (84.27%)

TABLE 4 | Mean (SE) scores per recording style in experiment 2 (*N* = 534).

Measures	Monolog (*n* = 178)	Question-answer (*n* = 178)	Recontextualized (*n* = 178)
Report			
Understandability	3.06 (0.12)	2.92 (0.11)	2.89 (0.11)
Credibility	2.66 (0.09)	2.50 (0.09)	2.52 (0.09)
Reliability	2.71 (0.10)	2.56 (0.09)	2.59 (0.09)
Suspect			
Credibility	3.50 (0.12)	3.53 (0.12)	3.53 (0.11)
Reliability	3.51 (0.11)	3.61 (0.12)	3.67 (0.12)
Confidence	5.77 (0.09)	5.58 (0.10)	5.61 (0.09)

Confidence = reported confidence in judgements regarding suspect. Dependent variables were measured on a 7-point scale.

TABLE 5 | Percentage of guilty judgments of the suspect per recording style in experiment 2 (*N* = 534).

	Recording Style			Total
	Monolog (*n* = 178)	Question-answer (*n* = 178)	Recontextualized (*n* = 178)	
Guilty judgments (%)	85.96	84.27	82.58	84.27

than in Experiment 1 (15.86%). It might thus be the case that recording style influences judgments (Experiment 1) but not when the case is somewhat more difficult to understand or when people are convinced the suspect is guilty (Experiment 2). Other studies have also shown that language effects may be overruled by other effects (e.g., order effects can overrule linguistic effects as was shown by Sherrill et al., 2015).

We considered our set of two experiments with mixed findings not strong enough to draw conclusions about the impact of reporting style on how people perceive a police record and a suspect. In addition, the fact that we only used one scenario in each experiment makes it difficult to generalize any result. Finally, in the two experiments so far, we found a strong correlation between credibility and reliability judgments for the record (0.90 in Experiment 1, 0.86 in Experiment 2) as well as the suspect (0.92 and 0.85, respectively). This raises the question whether we measured the same or different constructs. To address these issues, we conducted a third experiment.

EXPERIMENT 3

We conducted Experiment 3 to get a better understanding of *if* and *under what conditions* police reporting style impacts how people perceive a police record and a suspect. Our procedure for Experiment 3 deviated from that in Experiment 1 and 2 in several aspects. First, in Experiment 3 we used multiple scenarios instead of a single scenario (as was the case in Experiment 1 and 2). Second, we felt that—in retrospect—the questions regarding the reliability of the suspect and the credibility of the police record might have been semantically odd. After all, participants could only judge whether they thought the suspect came across as believable (i.e., hence the question about credibility), and whether they thought the police record accurately reflected the interrogation (i.e., hence the question about reliability). Judging the believability of the record and/or the accuracy of the suspect seems odd and provided us with information that is difficult to interpret. Therefore, we decided to only include a credibility question for the suspect, and a reliability question for the record. Third, with the question about the reliability of the police record, we were interested in learning how well people thought the police record reflected the interrogation. We considered a question relating to the accuracy rather than the reliability of the police record to be more intuitive. Therefore, we decided to ask participants to judge the accuracy rather than the reliability of the police record. Asking about accuracy instead of reliability might make the difference with credibility more salient. We also asked participants about the likability of the suspect because judgments of credibility are known to be influenced by the likability of a person (e.g., Ohanian, 1990). Finally, we decided to also ask participants to rate the imageability of the described events. Imageability is known to be influenced by subtle linguistic differences (e.g., Carreiras et al., 1997; Magliano and Schleich, 2000; Yao and Scheepers, 2011; Yao et al., 2011; Stites et al., 2013) and might be one of the mechanisms through which language impacts cognitive processes. For example, information that is perceived as more vivid is remembered better and easier to retrieve from memory (Reyes et al., 1980). Adding a question about the imageability of events might be informative to the question *if* and *under what conditions* reporting style impacts information processing and guilty judgments.

Methods

Sample

According to an a priori power analysis (Faul et al., 2007) we needed at least 288 participants (i.e., 96 per condition) to obtain statistical power at the recommended 0.80 level (Cohen, 1988). In total, we recruited 375 participants online through MTurk (in several batches) and 376 completed the experiment. We excluded data from 68 participants because they had reading times <0.05 ms per word for at least one of the eight police reports. This indicates that they could not have read all reports properly. We also excluded data from participants that participated twice (due to the release of several batches, *n* = 4), did not report English as their native language (*n* = 3), or a combination of both (*n* = 2). The final sample (*N* = 299) involved 105 participants in the monolog condition, 98 in the recontextualized condition, and 96 in the Q&A condition. One participant did not provide any demographic information. The mean age of the remaining 298 participants was 37.35 [*SD* = 10.85, range = 19–71, 120 (40.27%) females]. All participants were US residents and received $3 for their participation (that took ∼30 min).

Materials and Procedure

We selected eight real police reports concerning various crimes of comparable severity (i.e., shoplifting (2×), street robbery, counterfeit money/robbery, domestic violence, threatening with knife, stealing, attempted theft). In all cases a male suspect was brought to the police station for questioning where he actively denied being guilty of the crime. The reports of this interrogation contained information that could be perceived as incriminating as well as exculpatory. Some police records were based on authentic records. As in Experiment 1 and 2, we translated the original reports to English and we created two additional versions of each original report. Some police records were fictitious cases. All 24 reports were checked by two native speakers of English.

Experiment 3 had a mixed within-between subjects design with scenario as within subjects factor and reporting style as between subjects factor. That means that all participants were presented with the eight different scenarios, but that reporting style was consistent. We chose to present participants with eight scenarios in the same reporting style to make sure participants were not aware of the different reporting styles (and our interest in them).

Participants were randomly assigned to one of three reporting style conditions (i.e., monolog, Q&A, or recontextualized). Within each condition the eight scenarios were presented in random order to account for order effect. After participants carefully read a police report we asked them (1) how easy or difficult is was to understand the police record (7-point scale, 1 = *extremely easy*, 7 = *extremely difficult*), (2) how easy or difficult is was to imagine what happened (7-point scale, 1 = *extremely easy*, 7 = *extremely difficult*), (3) if they thought the suspect was guilty of the crime (*yes*/*no*), (4) how confident they were about their judgment (7-point scale, 1 = *not at all*, 7 = *extremely*), (5) how accurate they thought the report was (7-point scale, 1 = *extremely accurate*, 7 = *not accurate at all*), (6) how credible (believable) they thought the suspect was (7-point scale, 1 = *extremely credible*, 7 = *extremely uncredible*), and (7) how likable they thought the suspect was (7-point scale, 1 = *extremely likable*, 7 = *extremely unlikable*). Finally, participants stated what they thought this study was about and provided some demographic information. We recorded the time people spent reading the report. Again, this task was presented online in the Qualtrics survey research suite (http://www.qualtrics.com).

TABLE 6 | Linear mixed model results for all dependent measures in experiment 3.

Measures	Model 0		Model 1		Change in model fit
	−2LL	Parameters	−2LL	Parameters	*p*
Imageability	7,767.34	2	7760.57	4	0.034*
Report					
Understandability	7,536.35	2	7528.18	4	0.017*
Accuracy	8,076.94	2	8060.95	4	<0.001*
Suspect					
Credibility	9,114.85	2	9114.32	4	0.767
Likability	7,878.11	2	7877.69	4	0.814
Confidence	8,347.44	2	8347.03	4	0.812

Model 0 is the intercept only model. For Model 1 we added condition as fixed factor to the intercept only model.
**Significant at 0.05 level.*

TABLE 7 | Estimated mean (SE) scores per recording style for experiment 3 ($N = 299$).

Measures	Recording style		
	Monolog	Question-answer	Recontextualized
Imageability	2.47 (0.19)^a	2.16 (0.19)^b	2.32 (0.19)
Report			
Understandability	2.35 (0.20)^a	2.01 (0.20)^b	2.22 (0.20)
Accuracy	3.21 (0.13)^a	2.63 (0.13)^b	2.78 (0.13)^b
Suspect			
Credibility	4.41 (0.25)	4.50 (0.26)	4.47 (0.26)
Likability	4.74 (0.21)	4.70 (0.21)	4.67 (0.21)
Confidence	4.82 (0.17)	4.90 (0.17)	4.83 (0.17)
Guilty judgments*	0.62 (0.12)	0.65 (0.11)	0.65 (0.11)

Dependent variables were measured on a 7-point scale. Different superscripts in a row indicate a significant ($p < 0.05$) difference.
**Reported as the estimated proportion of guilty judgments.*

Results

To test whether reporting style influenced the imageability of the described crime, the understandability and accuracy of the police report, the credibility and likability of the suspect, or confidence regarding guilty judgments we used linear mixed models generated with SPSS (version 27). Compared to a repeated measures ANOVA, a linear mixed model is thought to reduce the chance of a Type I error (Quené and Van den Bergh, 2008). For all dependent measures we first estimated an intercept only model with a random intercept for *participant* and *scenario*. These models indicated that there was significant variance between participants regarding the imagebility of the described crime [Var[u_{oj}] = 0.59, $p < 0.001$], the understandability [Var[u_{oj}] = 0.58, $p < 0.001$] and accuracy [Var[u_{oj}] = 1.04, $p < 0.001$] of the police report, the credibility [Var[u_{oj}] = 0.63, $p < 0.001$] and likability [Var[u_{oj}] = 0.44, $p < 0.001$] of the suspect, and confidence regarding guilty judgments [Var[u_{oj}] = 0.56, $p < 0.001$]. There was also significant variance between scenarios regarding the imageability of the described crime [Var[u_{1j}] = 0.23, $p = 0.049$], the understandability of the police report [Var[u_{1j}] = 0.26, $p = 0.049$] and the credibility [Var[u_{1j}] = 0.45, $p = 0.049$] and likability [Var[u_{1j}] = 0.29, $p = 0.049$] of the suspect, but not regarding accuracy of the police report [Var[u_{1j}] = 0.04, $p = 0.076$] and confidence regarding guilty judgments [Var[u_{1j}] = 0.16, $p = 0.053$]. Following Barr et al. (2013) we decided to keep *participant* and *scenario* as random intercepts for all variables. We then added *condition* as fixed effect and compared for each variable separately the −2LL of this model that includes a fixed factor with the −2LL of the intercept only model. A decrease in −2LL indicates an increase in model fit. A significant increase in model fit suggests an effect of condition, and thus reporting style.

Table 6 shows how well the intercept only models fit our data and whether adding *condition* as a fixed effect significantly increases the fit of this model for each dependent measure. As can be seen, adding *condition* as a fixed effect did not improve the intercept only model for the credibility and the likability of the suspect, or confidence regarding guilty judgments. Reporting style did thus not influence these variables. The intercept only model did, however, improve significantly by adding *condition* as a fixed effect for the remaining three dependent measures. Reporting style had a significant effect on the imageability of the described crime [$F_{(2,298.15)} = 3.423$, $p = 0.034$], and the accuracy [$F_{(2,298.68)} = 8.216$, $p < 0.001$] and understandability [$F_{(2,298.13)} = 4.144$, $p = 0.017$] of the police report.

Pairwise comparisons, for which we used the Šidák correction to correct for multiple comparisons, showed that participants indicated that it was easier for them to imagine the described crime when having read a report in the Q&A style than in the monolog style ($p = 0.028$). Also, participants indicated that reports written in the Q&A style were easier to understand than those written in the monolog style ($p = 0.013$). Finally, participants considered police reports written in the monolog style to be less accurate than those written in the Q&A ($p < 0.001$) or recontextualized style ($p = 0.011$). We found no other significant effects of reporting style (see **Table 7**).

In addition, we conducted a generalized linear mixed model (i.e., because our outcome variable is measured on a dichotomous instead of continuous scale) to test whether reporting style influenced judgments of guilt. Our intercept only model with a random intercept for *participant* and *scenario* correctly estimated 78.5% of the observations in our sample. This model indicated that there was significant variance between participants [$\text{Var}[u_{0j}] = 0.64$, $p < 0.001$] but not between scenarios [$\text{Var}[u_{1j}] = 1.87$, $p = 0.066$]. Adding *condition* as fixed effect resulted in a model with a predictive value of 78.5% which did not differ from that of the intercept only model. Comparing−2LL of the intercept only model (11049.85) and that of our model that included *condition* as fixed effect (11058.79) even suggests that adding *condition* decreased the model fit. Reporting style did thus not influence guilty judgments [$F_{(2,2.389)} = 0.485$, $p = 0.62$].

Discussion

We conducted Experiment 3 to gain a better understanding of the discrepancy in results between our first two experiments. To rule out that this discrepancy was caused by unintended differences between the scenarios that we used, we decided to use multiple scenarios in our third study. Our results showed that reporting style did influence the perceived accuracy of the report with the monolog style being perceived as less accurate than the Q&A style and recontextualized style. In our previous experiments, we did not ask participants to rate the accuracy of the report. Instead, we asked for its reliability which was impacted by recording style in Experiment 1 but not in Experiment 2. The results of Experiment 1 showed that a police report written in the recontextualized style was perceived as less reliable than that in the Q&A style. If accuracy and reliability tapped into the same construct, the results of Experiment 3 and Experiment 1 both suggest that the Q&A style is considered the most accurate/reliable. This is congruent with our hypothesis that a police report written in the Q&A style represents the actual interview better (i.e., reflected in a representativeness score) than a police report written in the monolog or recontextualized style.

In contrast to the findings of Experiment 1 and 2, we found a significant effect for reporting style on understandability. Police reports written in the Q&A style were easier to understand than those written in the monolog style. This finding is consistent with Van Charldorp (2011) who concluded that out of the three main reporting styles the Q&A style is relatively informal and comprehensible. The fact that we found no significant difference in understandability between the recontextualized style and the monolog or Q&A style was surprising, given the complexity and rarity of the recontextualized style. After all, information that is presented in a way that deviates from our expectation (i.e., which is the case with presentations that we encounter less often) is more difficult to process (Zwaan, 1994).

Our finding that a described crime was easier to imagine after reading a police report in the Q&A style than in the monolog style fits with the result regarding understandability and supports the theory that mental model construction lies at the heart of language comprehension (Johnson-Laird, 1983; Van Dijk and Kintsch, 1983; Morrow et al., 1987; Zwaan and Radvansky, 1998). Information that is easier to imagine, is easier to understand, and also more likely to be remembered better (Reyes et al., 1980). Therefore, it is important that future research focuses on the impact of reporting style on memory and cognitive processes that rely on memory function (e.g., decision making).

Future research might also want to use alternative methods to measure the variables of interest. We were interested in very subtle effects of language use, yet our dependent variables were measured using a 7-point scale or a dichotomous scale. It might be that our method was not subtle enough to pick up on such subtle effects. This could also explain why we found no evidence that recording style influenced guilty judgments in all three experiments. An alternative measure of interest might be a think-aloud protocol. A think-aloud protocol—in which participants share their thoughts while reading police reports and answering questions—will provide useful information about how people process information.

CONCLUSION

So far, linguistic studies show that the written police record is often a selection of the actual interrogation that preceded it (Jönsson and Linell, 1991; Van Charldorp, 2011), and that transformations take place such that the written document becomes a structured, logical, chronological and neutrally told story of what happened (Van Charldorp, 2020). Processes of entextualisation, recontextualisation and decontextualisation across legal contexts have been elaborately discussed elsewhere (Heffer et al., 2013) showing, amongst many other things, that legal texts not only travel physically, but also across discursive spaces creating new contexts, interpretations and meaning. These types of transformations are not only relevant in the legal domain, but across many institutional settings where spoken interaction forms the basis of written documents. Such transformations, however, seem to be taken for granted in many studies. What the consequences are of very specific elements within this transformation process, has received very little attention (however, see De Keijser et al., 2012). In this study we took a closer look at how different linguistic reporting styles influence reader judgments concerning reliability and credibility of police records and the interrogated suspects. We found that reporting style indeed influenced the processing of information. More specifically, reporting style impacted the perceived accuracy of the report, as well as the understandability and imageability of the described event.

In sum, our study showed that language is important and that subtle differences in language use might have unintended effects. Clearly more research is needed. Only when we better understand the impact of subtle differences in language use and the mechanisms through which language operates, we can design better guidelines for police officers on how to construct a written police record that does not—unintentionally—influence the course of justice.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in an online repository and can be found below: https://osf.io/fpgz5/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Faculty Ethics Assessment Committee – Humanities, Utrecht University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AE and TC contributed to the planning of the studies, preparation of the materials, and wrote the manuscript. AE collected and analyzed the data. Both authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by (1) a VENI grant from the Netherlands Organization for Scientific Research (NWO) [grant number 275-89-026] awarded to AE, (2) funds generated from advisory work by TC, and (3) UiL-OTS at Utrecht University.

ACKNOWLEDGMENTS

The authors would like to thank Tanya Stevens and Isabel Zwaan for their help in translating the materials for our study and Kirsten Schutter for her help in analyzing the data for Experiment 3.

REFERENCES

Anderson, S. E., Matlock, T., and Spivey, M. (2013). Grammatical aspect and temporal distance in motion descriptions. *Front. Psychol.* 4, 337. doi: 10.3389/fpsyg.2013.00337

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Memory Lang.* 68, 255–278. doi: 10.1016/j/jml.2012.11.001

Carreiras, M., Carriedo, N., Alonso, M. A., and Fernández, A. (1997). The role of verb tense and verb aspect in the foregrounding of information during reading. *Memory Cogn.* 25, 438–446. doi: 10.3758/BF03201120

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edn.* Hillsdale, NJ: Lawrence Erlbaum.

De Keijser, J., and Malsch, M., and Kranendonk, R., and de Gruijter, M. (2012). Written records of police interrogation: differential registration as determinant of statement credibility and interrogation quality. *Psychol. Crime Law* 7, 613–629. doi: 10.1080/1068316X.2010.526119

Eerland, A., and Zwaan, R. A. (2018). The influence of direct and indirect speech on source memory. *Collabra Psychol.* 4, 5. doi: 10.1525/collabra.123

Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 175–191. doi: 10.3758/BF03193146

Fausey, C. M., and Matlock, T. (2011). Can grammar win elections? *Polit. Psychol.* 32, 563–574. doi: 10.1111/j.1467-9221.2010.00802.x

Ferreti, T. R., Rohde, H., Kehler, A., and Crutchley, M. (2009). Verb Aspect, event structure, and coreferential processing. *J. Memory Lang.* 61, 191–205. doi: 10.1016/j.jml.2009.04.001

Ferretti, T. R., Kutas, M., and McRae, K. (2007). Verb aspect and the activation of event knowledge. *J. Exp. Psychol. Learn. Memory Cogn.* 33, 182–196. doi: 10.1037/0278-7393.33.1.182

Givón, T. (1992). The grammar of referential coherence as mental processing instructions. *Linguistics* 30, 5–55. doi: 10.1515/ling.1992.30.1.5

Gregory, A. H., Compo, N. S., Vertefeuille, L., and Zambruski, G. (2011). A comparison of US police interviewers' notes with their subsequent reports. *J. Investig. Psychol. Offender Profiling* 8, 203–2015. doi: 10.1002/jip.139

Heffer, C., Rock, F., and Conley, J. (2013). *Legal-Lay Communication: Textual Travels in the Law*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199746842.001.0001

Jansen, T. G. (2011). *Juridische kennis en het proces-verbaal. Een onderzoek naar de kwaliteit en aandachtspunten voor verbetering*. Uitgave van de Politieonderwijsraad, Den Haag. Available online at: https://www.politieacademie.nl/kennisenonderzoek/kennis/mediatheek/PDF/81033.PDF. (accessed November 01, 2021).

Johnson-Laird, P. N. (1983). *Mental Models*. Cambridge, MA: Harvard University Press.

Jönsson, L., and Linell, P. (1991). Story generations: from dialogical interviews to written reports in police interrogations. *Text* 11, 419–440. doi: 10.1515/text.1.1991.11.3.419

Komter, M. (2013). "Travels of a suspect's statement," in *Legal-Lay Communication: Textual Travels in the Law*, eds C. Heffer, F. Rock, and J. Conley (Oxford: Oxford University Press), 126–146.

Loftus, E. F., and Palmer, J. C. (1974). Reconstruction of automobile destruction: an example of the interaction between language and memory. *J. Verbal Learn. Verbal Behav.* 13, 585–589. doi: 10.1016/S0022-5371(74)80011-3

Magliano, J. P., and Schleich, M. C. (2000). Verb aspect and situation models. *Discourse Process.* 29, 83–112. doi: 10.1207/S15326950dp2902_1

Malsch, M., De Keijser, J., Kranendonk, P. R., and de Gruijter, M. (2010). Het verhoor op schrift of op band? De gevolgen van het 'verbaliseren' van verhoren voor het oordeel van de jurist. *Nederlands Juristenblad* 37, 1931–2407.

Miller, L. S., and Whitehead, J. T. (2018). *Report Writing for Criminal Justice Professionals*, *6th Edn*. New York, NY: Routledge. doi: 10.4324/9781315267579

Morley, P. (2008). *Report Writing for Criminal Justice Professionals: Learn to Write and Interpret Police Reports*. West Berkshire: Kaplan Publishing.

Morrow, D. G., Greenspan, S. L., and Bower, G. H. (1987). Accessibility and situation models in narrative comprehension. *J. Memory Lang.* 26, 165–187. doi: 10.1016/0749-596X(87)90122-7

Ohanian, R. (1990). Construction and validation of a scale to measure celebrity endorsers' perceived expertise, trustworthiness, and attractiveness. *J. Advertising* 19, 39–52. doi: 10.1080/00913367.1990.10673191

Quené, H., and Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *J. Memory Lang.* 59, 413–425. doi: 10.1016/j.jml.2008.02.002

Reyes, R. M., Thompson, W. C., and Bower, G. H. (1980). Judgmental biases resulting from differing availabilities of arguments. *J. Pers. Soc. Psychol.* 39, 2–12. doi: 10.1037/0022-3514.39.1.2

Reynolds, J. (2012). *Criminal Justice Report Writing.* Scotts Valley, CA: CreateSpace Independent Publishing Platform.

Salomon, M. M., Magliano, J. P., and Radvansky, G. A. (2013). Verb aspect and problem solving. *Cognition* 128, 134–139. doi: 10.1016/j.cognition.2013.03.012

Schellingen, R., and Scholten, N. (2014). *Verdachtenverhoor. Meer dan het stellen van vragen*. Mechelen: Kluwer Belgium

Sherrill, A. M., Eerland, A., Zwaan, R. A., and Magliano, J. P. (2015). Understanding how grammatical aspect influences legal judgment. *PLoS ONE*. 10:e0141181. doi: 10.1371/journal.pone.0141181

Stites, M. C., Luke, S. G., and Christianson, K. (2013). The psychologist said quickly, "Dialogue descriptions modulate reading speed!". *Memory Cogn.* 41, 137–151. doi: 10.3758/s13421-012-0248-7

Van Charldorp, T. C. (2011). *From police interrogation to police record.* Oisterwijk: Uitgeverij BOXPress.

Van Charldorp, T. C. (2020). "Reconstructing suspects' stories in various police record styles," in *The Discourse of Police Interviews*, ed M. Mason and F. Rock (Chicago, IL: The University of Chicago press), 329–348.

Van Dijk, T. A., and Kintsch, W. (1983). *Strategies of Discourse Comprehension*. New York, NY: Academic Press.

Yao, B., Belin, P., and Scheepers, C. (2011). Silent reading of direct versus indirect speech activates voice-selective areas in the auditory cortex. *J. Cogn. Neurosci.* 23, 3146–3152. doi: 10.1162/jocn_a_00022

Yao, B., and Scheepers, C. (2011). Contextual modulation of reading rate for direct versus indirect speech quotations. *Cognition* 121, 447–453. doi: 10.1016/j.cognition.2011.08.007

Yu, H., and Monas, N. (2020). Recreating the scene: an investigation of police report writing. *J. Techn. Writ. Communic.* 50, 35–55. doi: 10.1177/0047281618812441

Zwaan, R. A. (1994). Effect of genre expectations on text comprehension. *J. Exp. Psychol.Learn. Memory Cogn.* 20, 920–933. doi: 10.1037/0278-7393.20.4.920

Zwaan, R. A., and Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychol. Bull.* 123, 162–185. doi: 10.1037/0033-2909.123.2.162

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 14 June 2022
doi: 10.3389/fcomm.2022.803452

Does Automatic Speech Recognition (ASR) Have a Role in the Transcription of Indistinct Covert Recordings for Forensic Purposes?

Debbie Loakes[1,2]*

[1] Research Hub for Language in Forensic Evidence, School of Languages and Linguistics, The University of Melbourne, Parkville, VIC, Australia, [2] ARC Centre of Excellence for the Dynamics of Language, Parkville, VIC, Australia

The transcription of covert recordings used as evidence in court is a huge issue for forensic linguistics. Covert recordings are typically made under conditions in which the device needs to be hidden, and so the resulting speech is generally indistinct, with overlapping voices and background noise, and in many cases the acoustic record cannot be analyzed via conventional phonetic techniques (i.e. phonetic segments are unclear, or there are no cues at all present acoustically). In the case of indistinct audio, the resulting transcripts that are produced, often by police working on the case, are often questionable and despite their unreliable nature can be provided as evidence in court. Injustices can, and have, occurred. Given the growing performance of automatic speech recognition (ASR) technologies, and growing reliance on such technologies in everyday life, a common question asked, especially by lawyers and other legal professionals, is whether ASR can solve the problem of what was said in indistinct forensic audio, and this is the main focus of the current paper. The paper also looks at forced alignment, a way of automatically aligning an existing transcriptions to audio. This is an area that needs to be explored in the context of forensic linguistics because transcripts can technically be "aligned" with any audio, making it seem as if it is "correct" even if it is not. The aim of this research is to demonstrate how automatic transcription systems fare using forensic-like audio, and with more than one system. Forensic-like audio is most appropriate for research, because there is greater certainty with what the speech material consists of (unlike in forensic situations where it cannot be verified). Examples of how various ASR systems cope with indistinct audio are shown, highlighting that when a good-quality recording is used ASR systems cope well, with the resulting transcript being usable and, for the most part, accurate. When a poor-quality, forensic-like recording is used, on the other hand, the resulting transcript is effectively unusable, with numerous errors and very few words recognized (and in some cases, no words recognized). The paper also demonstrates some of the problems that arise when forced-alignment is used with indistinct forensic-like audio—the transcript is simply "forced" onto an audio signal giving completely wrong alignment. This research shows that the way things currently stand, computational methods are not suitable for solving the issue of

OPEN ACCESS

Edited by:
Dominic Watt, University of York, United Kingdom

Reviewed by:
Vincent Hughes, University of York, United Kingdom
Michael Jessen, Bundeskriminalamt, Germany

***Correspondence:**
Debbie Loakes
dloakes@unimelb.edu.au

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *28 October 2021*
Accepted: *28 April 2022*
Published: *14 June 2022*

Citation:
Loakes D (2022) Does Automatic Speech Recognition (ASR) Have a Role in the Transcription of Indistinct Covert Recordings for Forensic Purposes? Front. Commun. 7:803452. doi: 10.3389/fcomm.2022.803452

transcription of indistinct forensic audio for a range of reasons. Such systems cannot transcribe what was said in indistinct covert recordings, nor can they determine who uttered the words and phrases in such recordings, nor prove that a transcript is "right" (or wrong). These systems can indeed be used advantageously in research, and for various other purposes, and the reasons they do not work for forensic transcription stems from the nature of the recording conditions, as well as the nature of the forensic context.

Keywords: forensic linguistics, transcription, automatic speech recognition (ASR), phonetics, forced-alignment

INTRODUCTION

Covert recordings are "conversations recorded electronically without the knowledge of the speakers" — these are crucial records because "legally obtained covert recordings can potentially yield powerful evidence in criminal trials, allowing the court to hear speakers making admissions or giving information they would not have been willing to provide in person, or in an overt recording" (Fraser, 2014, p. 6). However, indistinct forensic audio is generally captured by hidden recording devices, with uncontrolled variables such as overlapping speech, background noise and distance from the microphone to name a few. As such, resulting audio is especially unclear, to the extent that a transcript is often needed to assist in determining what was said. While there are some moves toward improving the process of creating a transcription of indistinct forensic audio, especially by the Research Hub for Language in Forensic Evidence at The University of Melbourne (see e.g., Fraser, 2020), misconceptions abound in terms of what is possible as far as this type of audio is concerned.

A common question asked of people working with indistinct forensic audio, especially by lawyers and other legal professionals, is how the problem of what is said in indistinct forensic audio can be solved automatically, with artificial intelligence (AI) and specifically automatic speech recognition (ASR). This is a fair question, because automatic methods are useful for many real-world issues, but it is a question that needs to be explored experimentally to understand what the problem involves, the mechanisms of ASR, and also what happens when one attempts to solve the problem computationally — this will all be addressed in the current paper. In the paper, forced alignment is also analyzed because it is a way in which an existing transcript can be "overlaid" onto an audio file, effectively segmenting and aligning words (and even individual phonemes) to audio, yet there are many aspects of this which need to be properly understood to use forced alignment effectively and appropriately.

A working definition of AI is that it is intelligence demonstrated by machines instead of humans, and importantly, as noted by McCarthy (2007) "computer programs have plenty of speed and memory but their abilities correspond to the intellectual mechanisms that program designers understand". ASR specifically involves the recognition of speech, generally segmented orthographically into words. The following definition of ASR (from O'Shaugnessy, 2008, p. 2965) gives a good general introduction to what systems are attempting to do when faced with speech signals:

> As in any PR [pattern recognition] task, ASR seeks to understand patterns or "information" in an input (speech) waveform. For such tasks, an algorithm designer must estimate the nature of what "patterns" are sought. The target patterns in image PR, for example, vary widely: people, objects, lighting, etc. When processing audio signals such as speech, target information is perhaps less varied than video, but there is nonetheless a wide range of interesting patterns to distill from speech signals. The most common objective of ASR is a textual translation of the speech signal. . .

In their review of ASR systems, Malik et al. (2021, p. 9419–9420) describe that ASR performance architecture of ASR systems falls into four "modules". These are:

1) A pre-processing module–this is a stage in the process in which the signal-to-noise ratio is reduced (various methods are used such as end-point detection and pre-emphasis). While it makes sense that this would work to possibly enhance or make speech clearer, any pre-processing of a file in forensic situations needs to be considered extremely carefully (see e.g., Fraser, 2019).
2) A feature extraction module. Malik et al. (2021, p. 9421) describe how the most used methods for this are Mel frequency cepstral coefficients, linear predictive coding, and discrete wavelet transform.
3) A classification module, which outputs the predicted text. Malik et al. (2021, p. 9421) note that different methods can be used to do this, either using joint probability distribution (a generative approach), or a method that calculates predictions based on input and output vectors (a discriminative approach). Importantly, both make use of training data.
4) A language module — this contains language dependent rules about syntax and phonology. Malik et al. (2021, p. 9421) explain that many ASR systems now work without a language module, but they also note the improved performance that comes with using the language module.

Writing this research paper as a phonetician who has worked with forensic speech evidence, it seems obvious that there will be problems with an automatic approach, and that it is unrealistic to assume it would work, but what are these problems specifically? Using the definitions of both AI and ASR above from McCarthy (2007) and O'Shaugnessy (2008), who mention programme/algorithm designers respectively, it is evident that humans are also decision-makers — there are a whole host of decisions and assumptions built in to the systems *by* humans. So it needs to be noted from the outset that these approaches are certainly not devoid of human intervention, and are thus not objective, despite common belief. Some biases in training data, for example, are discussed in research (e.g., Koenecke et al., 2020; Malik et al., 2021; Wassink et al., 2022) and this is expanded upon further in the next section of the paper. Additionally, O'Shaugnessy (2008), describes the fact that systems are taught to recognize "patterns", so perhaps one of the most obvious barriers expected in this research will be what kind of patterns (if any) are actually available in a noisy signal where speech can be less of an obvious feature than the noise. This issue will be explored in the current paper, which seeks to show what actually happens when ASR and forced alignment systems are used to help solve the problem of transcription of indistinct audio.

BACKGROUND

AI is particularly useful in various domains of our everyday lives, with cars that can center the vehicle in a laneway or brake before a collision can occur, facial recognition software that enables access to mobile phones, even spam filters on email systems that save time by automatically filtering emails that are not directly relevant. When it comes to speech, voice activated software is relatively commonplace–in smart phones, smart watches and in cars and homes to improve efficiency–for example people can ask their devices to turn on light switches, tell them the weather report, to find a location and direct them to that location, and so on.

In research, ASR, and forced alignment, have already proven extremely useful in the field of phonetics, sociophonetics and speech science more generally (some examples are Gonzalez et al., 2017; Mackenzie and Turton, 2020; Villarreal et al., 2020; Gittelson et al., 2021). Kisler et al. (2017) describe the "paradigm shift" that has occurred over recent years due to internet speed and connections being vastly improved, now allowing web-based platforms to be accessed and used easily by researchers. Automatic methods have also become very useful for language documentation purposes (e.g., Jones et al., 2019) and community members can also become involved due to accessibility (Bird, 2020). Such tools are also used very effectively in creating automatic subtitles, which can be done at very low cost, and even freely, with specific types of software. As many researchers have noted, the benefit of such tools lies in their efficiency, combined with the ability to analyse large amounts of data in order to better understand patterns in language. For example, one paper showed that it is possible to do 30 times the amount of analysis using automatic compared to manual methods (Labov et al., 2013), while another showed that depending on the task, automatic methods can improve efficiency of speech analysis by up to five times when compared with manual methods (for segmenting speech into utterances), or up to 800 times (for phonetic segmentation) (Schiel et al., 2012). This efficiency in processing, however, can also come hand in hand with a loss of precision. As noted by Coto-Solano et al. (2021, p. 17), for example, "in any scientific endeavor, there is a tradeoff between accuracy and speed, and each research project can determine what type of approach is appropriate". In forensics, however, there is no point at which speed is valued over accuracy due to the high-stakes nature of what is being analyzed.

This issue of efficiency also comes to the fore with forced alignment, which is a way of automatically aligning audio to a transcript (i.e., Jones et al., 2019), and is said to be "…highly reliable and improving continuously [yet] human confirmation is needed to correct errors which can displace entire stretches of speech" (Mackenzie and Turton, 2020, p. 1), and this is when clear recordings are used. In this paper, the analysis also focuses on how forced alignment fares with poor-quality recordings. This is of interest in the forensic domain, because a transcript can be created and then "matched" with an audio file—but there are various problems with this approach that need to be considered. Still on the topic of precision, in research contexts it has been convincingly argued that errors can be a risk worth taking. For example Evanini et al. (2009, p. 1658) state that "when very large corpora are used, errors in individual tokens and even individual speakers will not harm the analysis". Again, the same cannot be said for forensic situations, where what the speakers are saying is generally unknown and there is no definitive transcript to check the automatic version against. It is also often unclear who the speakers are, and even how many speakers there are (unlike in research situations). This is especially true in light of the fact that the success of systems comes with underlying assumptions which are explained well in the following quote "[i]n the cases of forced phonetic alignment and automated transcription … the technique rests on the assumption that there is some learnable, predictable pattern in the input that can be used to predict new cases" (Villarreal et al., 2020, p. 1); in forensic audio this condition is unlikely to be satisfied.

Before moving on further, it should be noted that most ASR systems work with HTK (Hidden Markov Model Toolkit) or Kaldi. HTK was developed at The University of Cambridge in 1993, and is described as "a toolkit for research in automatic speech recognition [which] has been used in many commercial and academic research groups for many years" (see e.g., Cambridge, 2021), while Kaldi is a more recently designed toolkit used for similar purposes (see e.g., Povey et al., 2011). MAUS, one of the systems used in this paper, uses HTK. Malik et al. (2021, p. 9417), explain that most ASR systems in use now also tend to use "long-short term memory (LSTM) … a type of recurrent neural network in combination with different deep learning techniques". Researchers are in agreement that ASR systems have shown vast improvements in a relatively short amount of time. For example Coto-Solano et al. (2021) explain the fact that this is due to the availability of training data, and deep learning algorithms, resulting in "important reductions in transcription errors". It is also important to note that ASR systems work differently due to "different feature extraction techniques and language models", yet this information is not always readily available to users seeking to understand and compare how the systems operate (see e.g., Malik et al., 2021). Even in "ideal conditions", then, ASR systems are certainly not error-free, and they are generally evaluated based on accuracy and/or speed, with "word error rate" and "word recognition rate" being metrics used to determine accuracy (Malik et al., 2021).

Even the developers of automatic systems report that "transcriptions and annotations should undergo a final correction step"–internal validity is needed to keep improving system performance and ensure consistency–in other words, it is not expected to be error-free. Schiel et al. (2012, p. 118), reporting on internal validity of systems with human analysts, note that around 99% accuracy between humans performing orthographic transcription (of clear speech) has been observed, 97% for clear spontaneous speech, 95% accuracy for phoneme boundaries on read speech with a window of 20 ms, 85% accuracy for phonemic boundaries on spontaneous speech with a window of 20 ms accuracy, and quite poor agreement at 66% accuracy with prosodic labeling. This itself shows actually making decisions about language is not categorical due to the continuous stream of acoustic information that makes up the speech stream (see further Fraser and Loakes, 2020).

Another issue with respect to ASR performance is inherent biases that filter in at various stages. This is covered well in a paper by Wassink et al. (2022), who note that male speech is recognized better than female speech and also that effects on signal quality are different depending on gender, and that when dialectal differences are included in training data, dramatic improvements in performance can ensue. Racial biases are also shown to exist; in their "cross-ethnicity study" comparing white and non-white voices, Wassink et al. (2022) show that sociophonetic differences in ASR are involved in 20% of system errors. They note that if dialect forms were included in the language module, better performance would ensue. Aside from just the issues with accuracy, Wassink et al. (2022) note that "…it is, of course, clear that unevenness in the accuracy of ASR systems primarily occurs to the disservice of everyday people in these social dialect communities, who use voice assistants to accomplish a wide range of tasks, from interacting with mobile devices to paying bills, and many others". Their results support findings of a related study, which showed a word-error rate of 0.19 for white speakers, and 0.35 for black speakers, when comparing performance of five popular and widely-used ASR systems (Koenecke et al., 2020). Another broader issue to consider is, as pointed out by Malik et al. (2021, p. 9412) that "training models are available only for a handful of languages out of a total of ∼6,500 world languages".

So, errors with ASR are not unexpected due to the variable nature of the systems, the speech that is fed into such systems, and bias in training data. Forced aligners, too, have differing levels of accuracy. A research paper by Jones et al. (2019) compared the performance of two automatic forced-alignment systems using one transcription and one audio recording, and showed some of the issues that arise when using automatic methods not completely set up for the problem at hand, as well as some of the inherent merits of the systems. It is interesting because it shows that "tweaking" by humans can achieve some improvements in performance, but only because humans are aware of the source of the data and thus what it is possible to achieve. It also shows that performance will not be ideal. The speech data analyzed in Jones et al. (2019) is produced by five young adults conversing in Kriol, an Australian English-based lexifier creole. Jones et al. (2019) used two options within MAUS (a programme also used in the current paper). They used a language-independent model (i.e., one in which the system learns "from scratch" on the available data) as well as a language-specific model (one in which the system was trained on a major world language), noting that there are advantages and disadvantages of both approaches. For the language-independent model, the steps were relatively straightforward given that no assumptions are made by the system about which language the data (input) is in. The authors note that "[t]he more different the "small" language is from the world language, the more errors in orthography, phonology, and phonetics" in the resulting output. For testing with a language-specific model, Jones et al. (2019), on the suggestion of MAUS developers, tried Italian because like Kriol it has a transparent orthography, a similar number of vowels in the inventory, and relatively comparable data (i.e. spontaneous speech data was used in the Italian training model).

Comparing to a "gold-standard" human segmentation of the data, Jones et al. (2019) show that, for forced alignment, the language-specific model (using Italian) had an overall better accuracy than the language-independent model. Looking at the alignment boundaries for vowel onset and vowel offset, they showed that the language-dependent model was 41.4% accurate within 10 ms of a boundary, and 85.9% accurate within 50 ms; it should be noted, however, that in the context of a speech segment 50 ms is quite wide and so "accuracy" does not mean an exact match, simply that the system was in the vicinity of marking the correct segment. For the language-independent model, results showed accuracy of 31.8% within 10 ms of the vowel, and 75.4% within 50 ms of the vowel. They also noted that the system was better at determining vowel boundaries at the onset rather than offset.

The results in the Jones et al. (2019) study show that with relatively good audio, but mismatched modeling (i.e., the wrong language input), forced alignment systems can assist in analysis but errors occur, and this is when the system is fed a transcript to assist in the task. The benefits of automatic systems are said to be their increased efficiency as discussed above, but as noted by Jones et al. (2019, p. 296) the errors are "concerning because they tend to take even longer to manually edit the alignment" — in other words, efficiency is reduced.

Of interest for the current paper, Jones et al. (2019, p. 294) reflecting on some specific parts of their attempts to use AI for coding Kriol, note that:

> …neither MAUS Italian system nor MAUS language independent mode is originally designed for the forced alignment of north Australian Kriol. Unavoidably, there are missing, extra, and wrong phonetic labels … and misaligned segments. In this study, the tokens with missing labels were excluded before further analysis. In some extreme cases, the onset and offset time can be off for a few seconds compared with the manually-edited data [which occurs for other automated aligners as well (Mackenzie and Turton, 2020)]. In our dataset we noticed that completely misaligned tokens tended to involve long stretches of sonorous segments (e.g., vowels, nasals, liquids, and glides) where presumably MAUS lacked strong acoustic landmarks like stop-vowel boundaries to assist in the alignment.

Other papers have also compared how systems perform under various conditions. Kisler et al. (2017, p. 333) look at system validation, reporting that when the MAUS system is tested on forced alignment, there is a 97% "MAUS-to-ground-truth agreement" with three human labellers when spontaneous German speech is used, and accuracy with segmental boundaries is around 90% when compared with humans. Kisler et al. (2017, p. 333) also report on accuracy rates when an existing language model (Standard Southern British English) is used for a variety that the system has not been trained on (Scots English) finding in this case that "MAUS had an error rate twice that of human experts", which highlights the importance of using systems with inputs they have been trained on.

In a paper comparing the performance of forced aligners with Australian English, as well as a second human coder, Gonzalez et al. (2020) showed that the human coders were most alike and accurate in their performance, at around 80% agreement in this paper compared to between 65and 53% for the ASR systems. They also showed the ASR systems made errors depending on particular phonetic environments, whereas crucially, human coders were not prone to such errors. Gonzalez et al. (2020, p. 9) note that their "study lends empirical support to the common wisdom that humans are far more consistent in creating alignments than are forced aligners, indicating that regardless of the aligner used, alignment accuracy will be enhanced by manual correction".

The research discussed here highlights some important issues relating to good-quality audio, which need to be considered before exploring the usefulness of ASR with indistinct forensic audio. Coming from a position of knowing what the material involves in the first place (who recorded it, who the speakers are and what language/dialect they are speaking) is one of the key factors in effectively using these tools to recognize speech and perform a transcription. In other words, the ground truth needs to be accessible from the outset, which is not the case in forensic situations. In forensic cases, the stakes are high and errors are not a trivial matter.

The question addressed in this paper is how automatic transcription might assist in indistinct forensic transcription, whether via ASR or using a transcript and forced alignment. A common query in both academic and non-academic circles is whether this can be done — in Australia, automatic transcription is indeed sometimes used to assist with summarizing lengthy recordings collected for investigative purposes, while police in Australia and elsewhere are also actively looking at extending this technology for indistinct audio used as evidence. In recent years researchers have also been investigating the application of automatic methods in the forensic context, such as alignment of telephone tapped speech with an already existing orthographic transcription (i.e., Lindh, 2007). It is feasible that aside from simply making analysis easier, a transcript (whether correct or not) could be fed into to a forced alignment system — again while it may be intuitive that this is inappropriate, it does not take away the possibility that this method could be used.

AIM

This study has a specific aim of demonstrating how automatic systems work with forensic-like audio, in comparison with good-quality audio. As pointed out by Lindh (2017, p. 36) "if only limited work has been done on the combination of auditory and automatic methods in comparing voices and speakers, even less work has been done on combining automatic speech recognition and forensic phonetic transcription". In other words, relatively little is known about the best ways forward, or even if there *should* be a way forward.

The aim of this research is thus to analyse, experimentally, how two ASR systems perform when tasked with the transcription of indistinct forensic-like audio. It also aims to assess what happens when a transcript is fed into a system with indistinct forensic audio (i.e., a forced alignment system). Potential issues in forensic transcription which result from these demonstrations will be discussed.

METHODS

Data

This project used two recordings to test two ASR systems, and compare their performance. The number of recordings is minimal so that broad issues can be demonstrated[1]. The recordings are purposely different to replicate the forensic context where "mismatched conditions" are par for the course (e.g., Jessen, 2008, p. 700).

The recordings used are:

Audio

1. "*poor-quality" audio*. This is a 44.2 second stretch of audio from a recorded rehearsal by a singer and some musicians. This stretch of audio includes speech and instrument noise, and is forensic-like in that there are varying background noises, there are multiple speakers who are at a distance from the microphone, there is overlapping speech, and there are also people present who were not recorded (but this was not recorded in the context of crime). This audio was recorded by one of the speakers via an iPhone and streamed to Facebook live, where it was retrieved with permission. We are in a fortunate position with the audio, because the speakers are known, access to an associated video was granted, nouns used have been checked, and the transcript has been verified with one of the speakers who organised and streamed this event. The recording used has one female voice and three male voices, and all speakers are using Australian English. In this case the speakers knew they were being recorded, but were focused on the task at hand and not attempting to be clear to the audience; they were sharing the file so fans could see what a rehearsal looked like, and so the audience could experience the music (in those parts of the file, microphones were being used). The content of the speech produced in between the songs was focused on planning the live music event, as well as general

[1] Another research project is currently underway using more data - real forensic audio, "fake" transcripts and recordings made on different channels (including telephone recordings).

conversation, and it is one section of speech in between songs used in this research[2].

For the poor-quality recording in the current experiment, a reliable transcript is as follows. Here we make no attempt to attribute the utterances to particular speakers.

Yeah so just slowly building energy and nnnn and then I yeah
What about what about another big drum fill will you let us know when you
Yeah
Alright
Nah nah
You gonna give us a hand signal or tell us what you do
I I can't [laughter] ok
From the from the top are we fine to go there
Mel you don't need to do it so you know
I mean this song I think is OK no it's relatively OK I I mean from the top of the set just marking it out what do you think yea nay care
Sorry my brain just
What song are we practicing?
Run through
From the top
yeah

2. Unlike the poor-quality recording, the second audio file is termed *"good-quality" audio*. This was also recorded on an iPhone. In this case, there was a single speaker, the microphone was close to her mouth, there was little background noise, and the speaker was mindful of being understood. She was producing an utterance for a summer school for students learning about the programme MAUS, which is used in the current research paper (and described in the next section). This audio file is 8.4 s, and is spoken by an Irish English speaker recorded in Australia. The speaker has given permission to use this recording. The transcript for this file, separated into intonation units, is:

Hello
my name's Chloé
I live in Melbourne
I'm from Ireland
I moved from Galway
two and a half years ago
and I love MAUS.

It should be noted that these recordings, aside from being recorded on iPhones, are extremely divergent in nature — choosing divergent recordings is purposeful because it attempts to replicate forensic situations with their mismatched conditions. In the forensic domain, so-called "questioned samples" are compared with non-forensic "suspect" samples, and they are generally from extremely divergent sources — because forensic samples contain important speech evidence, it is often necessary for some kind of analysis to go ahead (i.e., simply discarding the samples due to these differences is not appropriate). This is discussed by, for example Rose (2002), and also see Jessen (2008, p. 685–686), who review some common technical differences across such samples, citing that forensic samples may be shorter, contain echo, have a mismatched sampling frequency compared to the suspect sample, be recorded via telephone, or have overlapping speech and/or background noises. The forensic sample in this recording is actually longer than the good-quality recording, but does indeed contain overlapping speech and background noise, with speech also at a distance from the microphone.

Software

There are three programmes used for the task of recognizing speech in the good-quality and poor-quality recordings respectively.

BAS SERVICES (Bavarian Archive for Speech Services)—ASR and WebMINNI

There is "a set of web services" at the Bavarian Archive for Speech Signals (BAS) in Munich that were developed for the processing of speech signals" (Kisler et al., 2017, p. 327). These include ASR, forced alignment, voice activity detection, speech synthesis and an online "labeller" which can be used to mark boundaries between linguistic events (syllables, intonation units) called EMU – these can all work together[3]. In this paper the focus is on two of these services.

Firstly, MAUS is used, and specifically "WebMINNI" because, as stated on the website, it "computes a phonetic segmentation and labeling based solely on the speech signal and without any text/phonological input". In this case, the result needs to be read back by reconstructing phonemes as there is no resulting orthographic transcription as such. This is effectively a forced-alignment tool which, in the words of Kisler et al. (2017, p. 331), uses

> [a] two-step modeling approach: prediction of pronunciation and signal alignment .... In the first step, MAUS calculates a probabilistic model of all possible pronunciation variants for a given canonical pronunciation. This is achieved by applying statistically weighted re-write rules to a string of phonological symbols. The language-specific set of re-write rules is learned automatically from a large transcribed speech corpus. The pronunciation variants, together with their conditional probabilities are then transformed into a Markov process, in which the nodes represent phonetic segments and the arcs between them represent transition probabilities.... In the second step, this Markov model is passed together with the (pre-processed) speech signal to a Viterbi coder ... which calculates the most likely path through the model, and – by means of backtracking this path – the most likely alignment of nodes to segments in the signal.

The WebMINNI service does not have an Irish English model, so a UK model was used. It is acknowledged that this model probably included a majority of non-rhotic speakers, unlike the Irish English used by the speaker, but as the results will show this is not an issue for what is being focused on in the current study.

[2] Other sections of the audio which contain speech are being used for a separate experiment on the transcription of indistinct audio with human transcribers.

[3] https://www.bas.uni-muenchen.de/Bas/BasMAUS.html

The BAS services ASR system was also used, which requires only audio and returns an orthographic output[4]. For the ASR service there are many language models that can be selected, including both an Australian English and Irish English model which are used for the poor-quality and good-quality recordings respectively. As noted on the website for the BAS services, third party services are used for this service, including Google Cloud and IBM.

Descript

Descript is another programme used in this research[5]. It is described as "all in one video and audio editing" and has functions to assist with podcasting, screen recording, video editing and transcription (used in this research). It is freely available (up to 300 h per month) and has an ASR component, which works using "Google Cloud's Speech-to-Text technology" (Opiah, 2021), and in this way has some similarity with BAS Services (which uses Google Cloud, but other technology as well[6]). The mechanisms of Descript are less well-described, presumably as it is not normally a research tool in the way BAS services are, and is available for use to anyone without the need for explicit training.

RESULTS

BAS SERVICES: ASR

Firstly focusing on how the MAUS fared with the poor-quality recording, the ASR option was used within the BAS Webservices. The number of speakers was selected (four) and an Australian English model was used. Once we uploaded the file, this was unable to be read at all, the system returned the following error

> *StdErr: ERROR: callGoogleASR: can't find a transcript in server response; this means either a bad signal quality or empty signal–exiting*

Because we know it was not an empty signal, we can be confident that there was a bad signal, which is unsurprising. So in this case, the ASR failed for this recording.

When we tried the ASR service with the good-quality recording, and chose one speaker as well as an Irish English model, we had a successful result (with some errors, underlined).

> *Hello, my name is Chloe I live in Melbourne are from Ireland I met from Galway to 1/2 years ago and I love maths.*

This is a successful output, although there are some minor errors in the form of introduced sounds or wrong words, which are underlined. These are:

1. *name is* should be *name's*,
2. the word *are* should be *I'm*
3. *to 1/2* is almost correct (even though two *1/2* is technically more correct) but the words *and a* is missing, i.e. the speaker said two *and a 1/2;*
4. *maths* should be *MAUS*

The free "WebMINNI" service was also tried, which has the component allowing recognition of phonemes without any transcription. For the poor-quality recording, we found that almost no speech (no phonemes) were recognized at all–although the system did very well at finding silence intervals. To give some examples, **Figure 1** shows a screenshot of the waveform as well as the resulting phoneme tier which was the output from the WebMINNI system[7]:

As seen in the image, there are some sections that are labeled "<p:>" which means *silence interval*, and some labeled "<nib>" which means *non-human noise*. This image does not show the whole file. It is certainly not the case that the <nib> sections were non-human noise, in fact this is where the human speech was located in the file in many cases. The silence intervals, however, were relatively well captured.

As another example, and to be more specific about the kinds of errors observed, **Figure 2** shows some of the output from WebMINNI, which occurred later in the file after **Figure 1**. There is a small amount of overlap between the end of **Figure 1** and start of **Figure 2**.

Figure 2 shows more activity on the phoneme tier compared to **Figure 1**. Here it can be seen that the system attempted to find some speech segments, and while this is the case the actual identification of sounds was not successful.

Some specific examples are:

> <nib> at the left of **Figure 2** is an entire section of speech produced by the female speaker in which she says *are we fine to go there*, but is analyzed by the system as non-human noise.

For the first section marked "h" the female speaker is in fact saying "Mel" (so there are three segments, not just one, and the marked segment is wrong). The remaining four are trumpet noises (trumpet noise is also occurring in other sections).

In the section marked V (which technically represents an open vowel) the female speaker is saying the phrase *is OK no*.

Additionally, the first <p:> in **Figure 2** is in fact marked correctly as a silence interval–and while some activity can be seen on the waveform, this is background noise which is almost inaudible. The second <p:> (at the end of the **Figure 2**) is the speaker saying *it's relatively OK I I mean from the top of the s-* (the remainder of the word *set* is not shown). In this case, the <p:> is wrong.

WebMINNI then, has not been able to segment speech sounds in the poor-quality recording. It has identified some sections of speech as "non-human noise" and has incorrectly identified whole words and phrases as one speech segment.

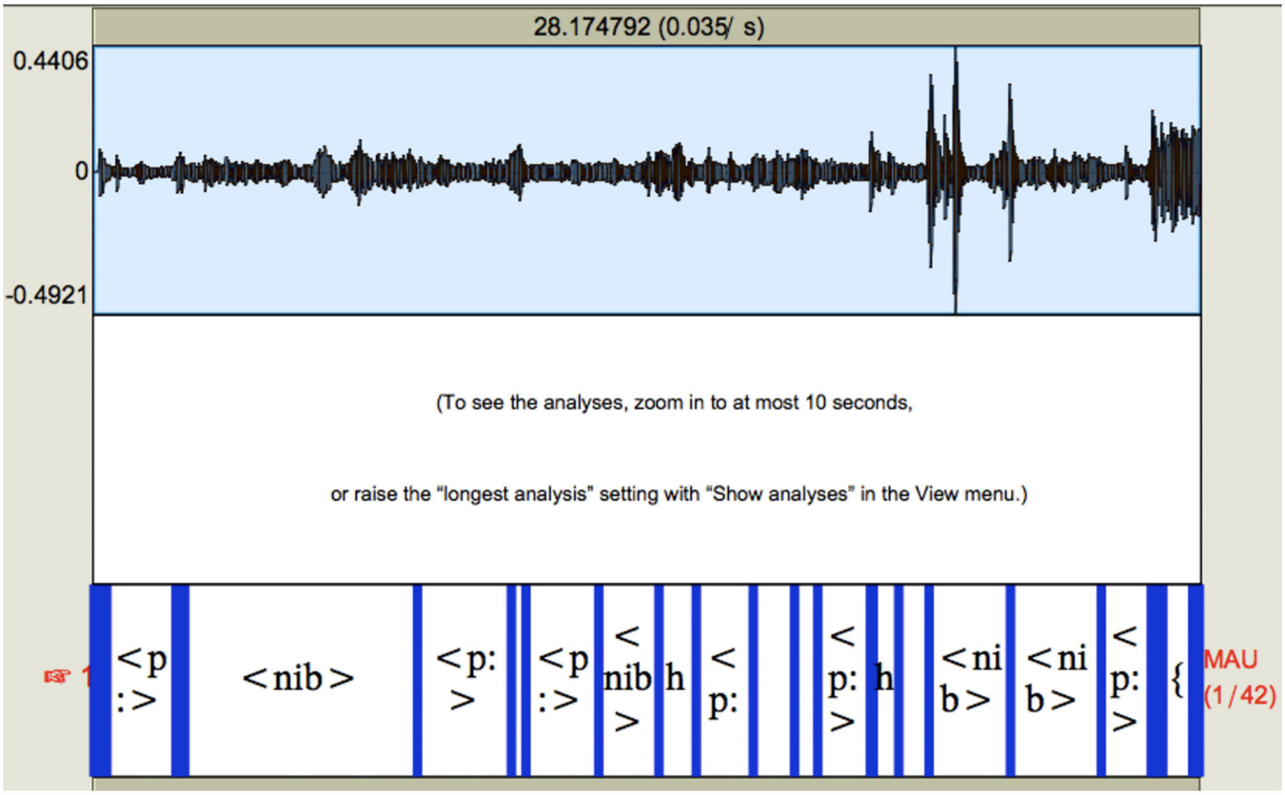
On the other hand, the good-quality recording fared relatively well (but better when the ASR option was chosen). WebMINNI

[4] This requires a login via a Clarin account which can be accessed through education institutions.

[5] https://www.descript.com/

[6] While there is thus some similarity with BAS services and Descript, their differences lie in the specific language modules they use as well as different ways of applying feature extraction and prediction.

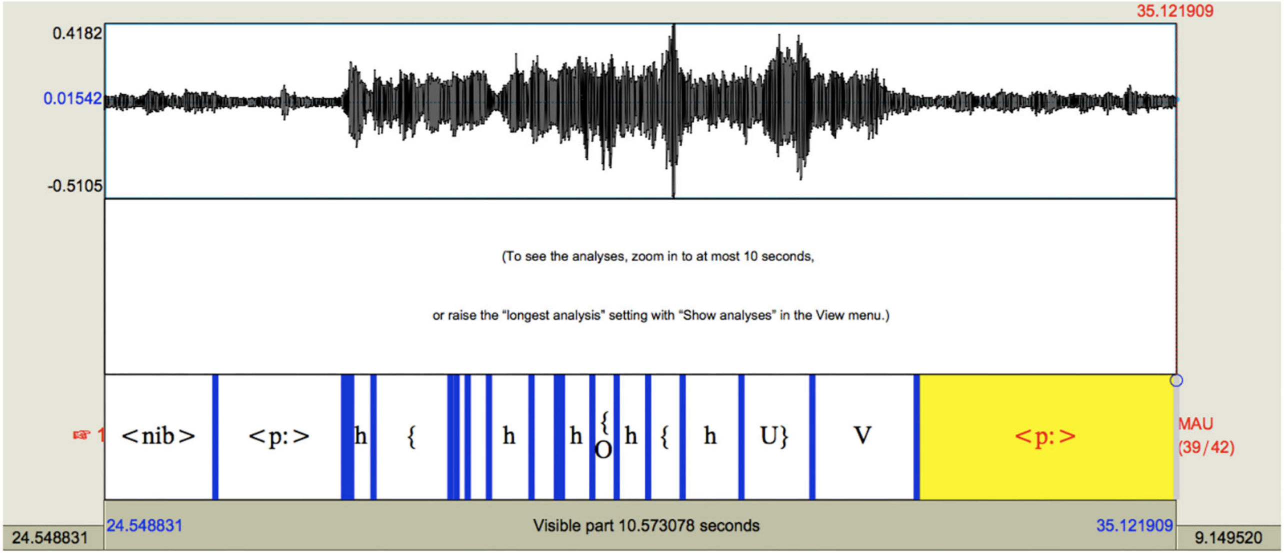
[7] The spectrogram is not visible in this Figure, nor in **Figure 2**, as the aim is to show "non-speech" category labels on the phoneme tier.



(To see the analyses, zoom in to at most 10 seconds,

or raise the "longest analysis" setting with "Show analyses" in the View menu.)

FIGURE 1 | Example 1 of system output with the poor quality recording using WEBMINNI, ASR.



(To see the analyses, zoom in to at most 10 seconds,

or raise the "longest analysis" setting with "Show analyses" in the View menu.)

FIGURE 2 | Example 2 of system output with the poor quality recording using WEBMINNI, ASR.

was able to segment the speech segments but with some errors, and so it is possible from that to reconstruct what the speaker was saying. Using names as examples, some errors in the good-quality recording are:

Chloé is rendered /koƱaɪ/
Galway is /kaoɪeɪ/

This indicates there is some inability for the system to pick up the /l/ sound in the speaker's voice. Interestingly, the system appears to have been making predictions about /l/ vocalization (replacing the speaker's relatively dark /l/ with back vowels), which may be because we are using the British English model, so anything /l/-like may be being converted to a back vowel for this reason. The best pattern recognition that the system could do in this case was a back vowel; in other words the system is interpolating from the available data and the assumptions being made about it. Across the file there are also some other minor errors, with some nasal sounds confused – i.e. /m/ sounds written as /n/. So, in this case, for the good-quality recording the ASR system worked better than WebMINNI, likely promoted by the Irish English model in the former – it is known that suitable training data, when it comes to sociophonetic and linguistic factors, boosts performance (i.e., Wassink et al., 2022). For the poor-quality recording, neither the ASR or WebMINNI was successful.

BAS SERVICES: Forced Alignment

Within the BAS services, the forced-alignment option was used, with an orthographic transcript. The important thing to note is that this was a reliable transcript — the subject matter is known, and the speakers are known, so the speech matter has been verified. This would not be possible to do in a forensic situation where there is no way of verifying anything that could be fed into the machine.

When the transcript was used with the poor-quality recording, WebMINNI was able to correctly segment (force-align) some of the words, although there were more errors than correct segmentations. The background noise and overlapping speech made the task difficult for the system because the noisy signal does not allow acoustic landmarks to be recognized. As an example, **Figure 3** shows a section of speech in which the speaker is saying *Just slowly building ener-* (not all of the word *energy* is visible in the figure shown). However, the system has force-aligned only the word *just* correctly, and none of the other words are correctly aligned. In fact the whole word *energy* is shown, as well as the word *and*, despite the fact that they are not present in this exact stretch of audio. Additionally, the poor-quality of the spectrogram is evident in this example.

As another example of WebMINNI's performance, in the following example shown in **Figure 4** the phrase (*From) the—from the top* is force-aligned onto a section of the recording that is actually drumming noise and laughter, but this was recognized as speech. This can be likened to what happens when software which is designed to recognize faces "believes" that clouds and trees are people. The system has attempted to match boundaries, or qualities observed in the signal, with phonemes / words—which it is designed to do but of course the trouble here is that there are no phonemes or words in this section.

In contrast, using a transcript with the good-quality recording is very successful as seen in **Figure 5**, although there are some errors which should be addressed. Because a non-rhotic model was used, the transcription of *Melbourne* and *Ireland* (of which the output does not contain /r/) are incorrect in this respect– in other words the system failed to recognize the rhotic in the speaker's pronunciation of these names because it is effectively trained to ignore them in the UK English model–presumably if we had tried an American English model the transcription would have been more reflective of the actual pronunciation of these items. Also, the second syllable of *Melbourne* is not transcribed with a schwa vowel (in the transcription system, schwa is the @ symbol) so the "O:" symbol, a long back vowel, is also technically wrong. Here, the system has inferred the statistically most likely pronunciation based on the "-ourne" spelling in this word. The remainder of the file, not shown here, was also relatively successfully transcribed.

Regarding alignment, the only errors visible in **Figure 5** are the boundaries between *Chloé I live in,* which are misaligned. The word *Chloe,* for example, is force-aligned onto just the onset segments of the /kl/ portion of the word. There are also alignment errors in the following words, but from *Melbourne* the alignment becomes accurate again.

DESCRIPT: ASR

Descript is a system which is designed for the general public, and so is very straightforward in terms of having an audio input and an orthographic output. When Descript was tried with the poor-quality recording, only three words were recognized by the system, the words *yes*, *yeah* and *okay*. While three words were identified, the word *yes* was not exactly correct (the speaker was actually producing another repetition of *yeah*). These words were recognized (or partially recognized) likely because they were somewhat louder, and so potentially "stood out" from the background noise. The Descript system did not recognize any other words. The total number of words uttered by the four speakers was 116, so this means the recognition rate was only 1.7%.

When Descript was tried with the good-quality recording, the output was almost entirely correct aside from the spelling of Galway (which was spelt with *Gallway*, but this is effectively inconsequential) and the very last word in the phrase *I love MAUS* which was recognized instead as *I love my house*. This recording was of course much shorter, but even if we say *Galway* is incorrect due to its spelling, and say that the error in *MAUS* is two errors, the recognition rate is 22/25 and effectively 88%. If we are more generous and say that *Galway* is correct, and *MAUS* is only one error (being an incorrect noun phrase) the recognition rate is 96%. Whichever way we decide to judge these errors, the performance of Descript is clearly superior when we use the good-quality recording. Mistakes are explainable due to predictability, which is especially low for the software *MAUS.*

DISCUSSION

This research shows that if we have clear, non-overlapping speech in a language variety that the system is familiar with, then ASR systems work very well. This is not surprising, as this is what the systems are designed to handle. However, if we have indistinct forensic-like audio, where speakers are not positioned near a microphone, or have overlapping speech with multiple sources of background noise, the systems perform badly. As shown with WebMINNI, even with a transcript, performance is far from ideal–forced-alignment does not accurately recognize word boundaries in most cases. However, this is not surprising, and not a criticism of developers of these systems, who have

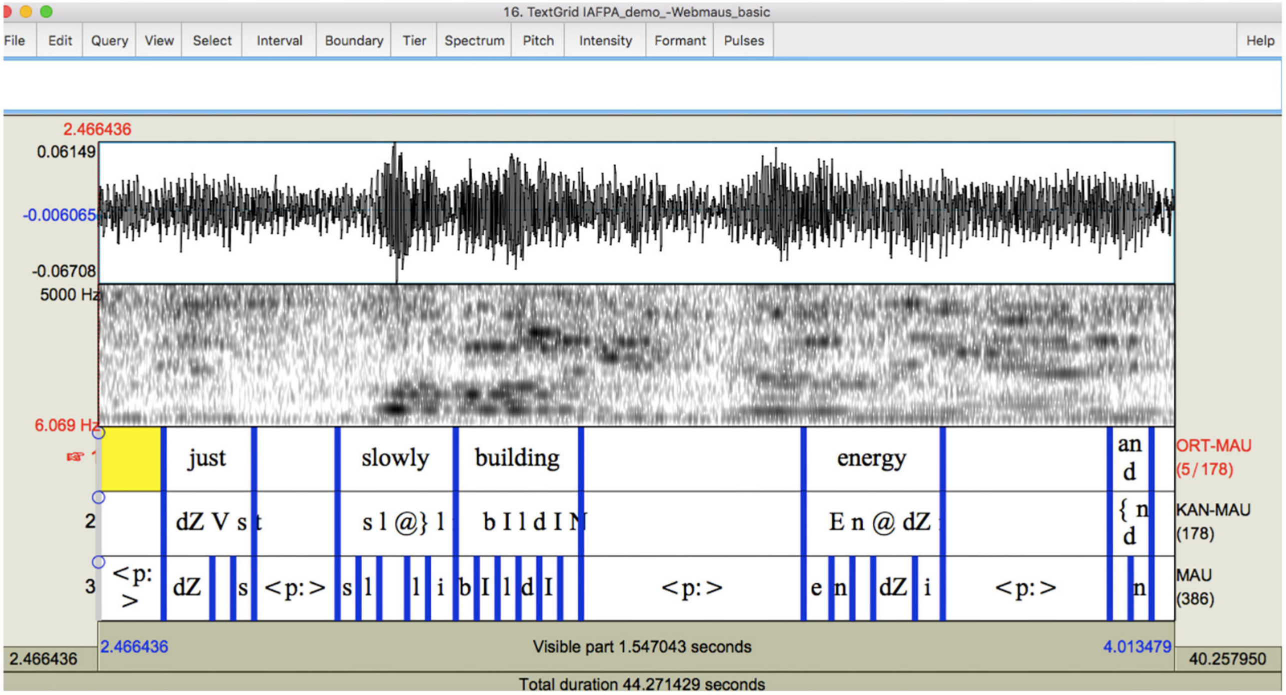


FIGURE 3 | Example 1 of system output with the poor quality recording using WEBMINNI, forced-alignment.

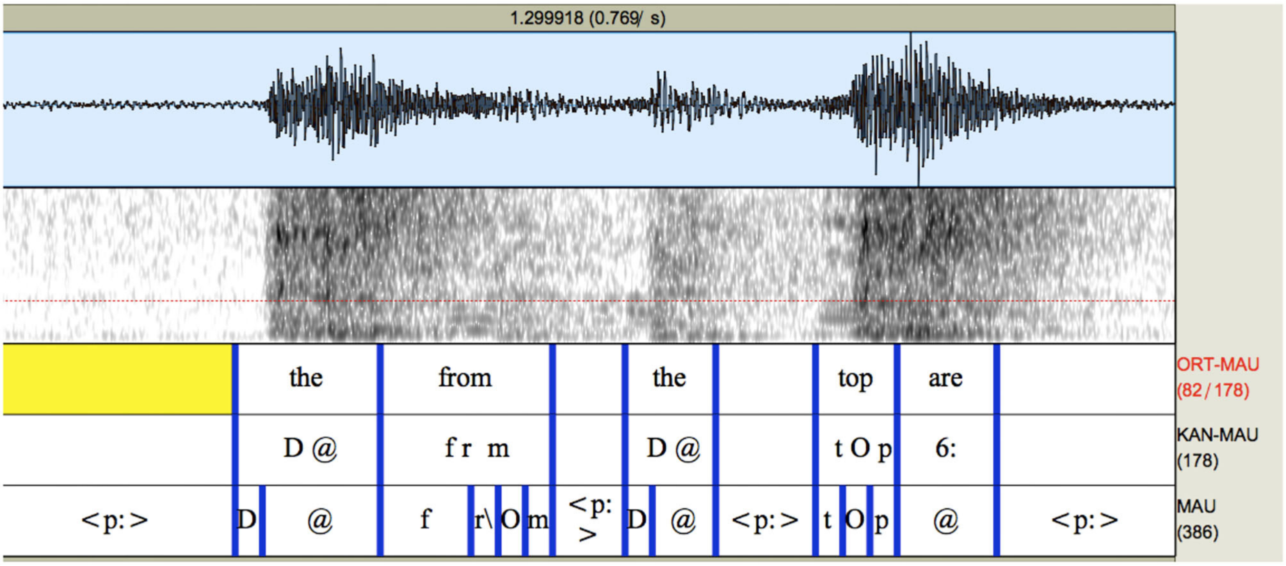


FIGURE 4 | Example 2 of system output with the poor quality recording using WEBMINNI, forced-alignment.

not advertised their systems as being made for the transcription of indistinct audio. It does, however, make clear why people working in the area of transcription of indistinct audio do not turn to computational methods to solve the problem.

It must also be acknowledged that automatic methods can be used to solve some issues in forensics–for example they can cut down significantly on manual work by an analyst, making tasks more efficient. One example is the segmentation of speech from non-speech, even if the recordings are very poor quality, as shown here with the poor-quality recording when it was run through WebMINNI.

Given the results of the research shown here, the cautions and concerns raised about automatic transcription in sociophonetic and sociolinguistic literature, where fine detail and "a

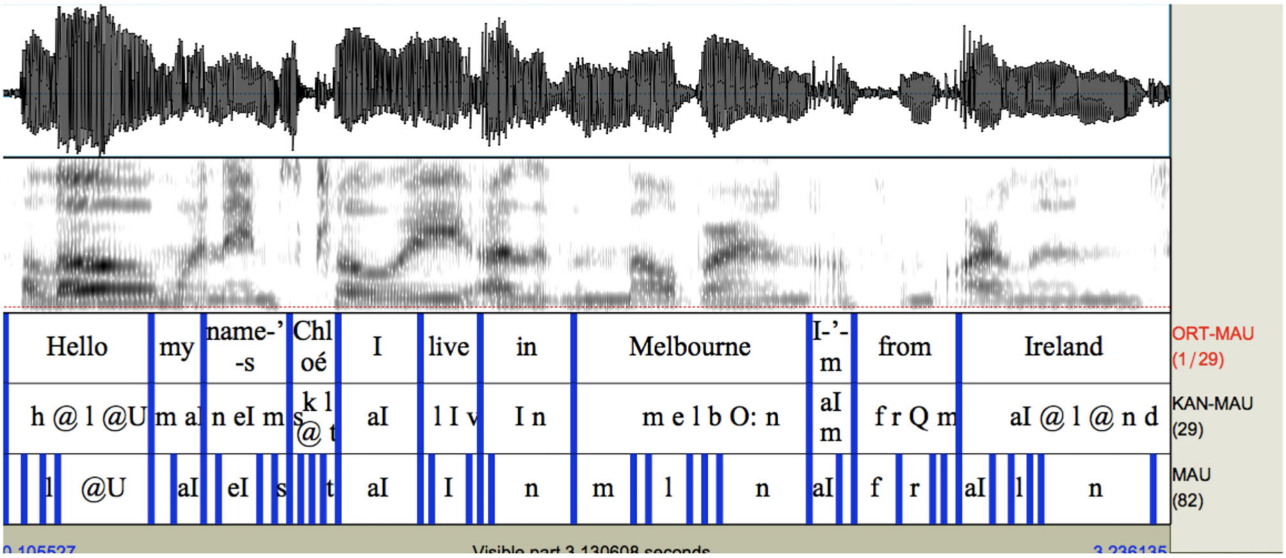


FIGURE 5 | Example of system output with the good quality recording using WEBMINNI, forced-alignment.

constellation of acoustic cues" are important and should not be factored out (Villarreal et al., 2020, p. 2), are even more pertinent for forensic purposes where the stakes are far higher. Returning to the quote from Mackenzie and Turton (2020, p. 1) though "…although forced alignment software is highly reliable and improving continuously, human confirmation is needed to correct errors which can displace entire stretches of speech." This human intervention raises the question of bias and priming which is unproblematic in research and language documentation situations, where acoustic cues are also clear and the ground truth can reasonably be established, and mistakes would regardless be occurring in a relatively low stakes environment. It is of course a concern for transcription of indistinct audio for use in court situations where stakes are far higher, and just like Lindh (2017, p. 58) reports for automatic speaker recognition contexts, it "would be unwise to presume that one can be a completely ultra-objective bystander feeding a system with the necessary inputs to decide the strength of the evidence".

As noted by Jones et al. (2019, p. 284), however, when evaluating whether to use a language-independent or language-specific model for Kriol within MAUS "the choice is always dataset-specific". This holds for indistinct forensic audio, but the very fact that the contents of the file are generally unknown (unlike in research) this means that any choices made about how to deal with the data effectively are simply guesswork, which is unsatisfactory.

Even though some people may expect better performance when computational methods are used, the requirement for human intervention can be *greater* when we use systems not designed for the task at hand (e.g., Jones et al., 2019). This is also clear in the current analysis, where using automatic methods offered arguably no benefit in assisting with the transcription of the poor-quality recording, with a refusal to read the signal when the BAS ASR service was used, nothing correct when using MAUS without a transcript, two words correct with Descript, and quite poor performance when forcing segmentation onto a transcription which we know to be a "gold standard" transcription. The good-quality recording, however, produced a useable transcript in the BAS ASR service and in Descript, although as shown there were some errors, especially where predictability was low, i.e., the word *MAUS* and some other cases in which small words were added or not recognized. However, when these automatically-produced transcripts are fed into MAUS, very little manipulation would be required at all. In other words, even though some manual intervention would be required for checking and correcting (especially for low-predictability items, as we saw), using ASR systems with data such as our good-quality recording is clearly more efficient than a fully manual method of analysis, as has been reported by other researchers.

CONCLUSION

As things currently stand, when recordings are poor quality and there is no definitive transcript (typical for forensic contexts), this research has demonstrated that automatic methods cannot solve the problem of what was said in indistinct forensic audio. The issue of what material ASR systems are trained on is unresolvable for many forensic contexts–the noisy conditions are problematic, as is the fact that speakers are often contested–therefore guesswork is needed to apply automatic methods and this is entirely unsatisfactory. It is also problematic that a transcript can be fed on to any audio and possibly *look* correct. Systems can appear to work on transcription data that is simply wrong, and just because a system error does not occur, it does not mean that an output is correct. These main points of the paper may perhaps be obvious to linguists and phoneticians, but the issues need to be demonstrated, explored and acknowledged for a broader audience as has been achieved here. The demonstration in this paper has used data which is extremely mismatched to replicate common forensic situations, and has shown marked breakdowns in performance. Other experimental work that is planned on automatic methods will investigate the deterioration of ASR performance in a more stepwise manner, to better understand where these breakdowns in performance occur and why (focusing first on signal quality reductions and keeping speaker numbers equal, for example)[8].

In the new Research Hub for Language and Forensic Evidence at The University of Melbourne, we hope to work with others to find "solutions that allow maximal value of the intelligence contained in covert recordings, while reducing the risk of injustice through biased perception of indistinct audio" (Fraser, 2014, p. 5). This means taking a cautious and measured approach when it comes to the use of ASR (and forced alignment) in forensic phonetics, without discounting their effectiveness in every domain. We are engaged in experimental work which aims to better understand how well human transcribers (with an aptitude for transcription of indistinct forensic audio) handle forensic-like audio when producing transcripts. As mentioned in the background, and as can be deduced from comparing the research discussed here, we should expect that humans will perform better than machines, but also that it will take them longer (i.e., Schiel et al., 2012). This matter of efficiency should be subject to a risk-benefit analysis, and we argue that in forensics the risk of losing accuracy is too great, and that human intervention is entirely appropriate for this task – however, the specifics of how to do this in the best way is still an open question.

As noted by Watt and Brown (2020, p. 411) in their discussion of the role of automatic methods in speaker recognition, there is a clear need to "[develop] initiatives to stimulate broader and deeper dialogue among practitioners in … closely related fields" so that all parties understand the nature of indistinct covert recordings, as well as the capabilities of automatic systems– what they have been developed for, and their extension outside that realm.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the author upon request.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the University of Melbourne, Project ID 21285. Participants provided written consent to participate in this study, and written informed consent was obtained from the individual whose potentially identifiable data is included in this article.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

The author receives funding support from the School of Languages and Linguistics and the Faculty of Arts, at the University of Melbourne. She also receives support from the ARC Centre of Excellence for the Dynamics of Language, Grant ID CE140100041.

ACKNOWLEDGMENTS

Thanks to Helen Fraser for assistance with ideas in this manuscript and discussion of issues surrounding the main themes within. Thanks are also due to Hywel Stoakes for discussion around ASR techniques.

[8] Thank you to reviewer 2 for explicitly pointing out this research focus.

REFERENCES

Bird, S. (2020). Sparse transcription. *Comput. Linguist*. 46, 713–744. doi: 10.1162/coli_a_00387

Cambridge (2021). *HTK–Hidden Markov Model Toolkit - Speech Recognition Toolkit*. Available online at: http://htk.eng.cam.ac.uk/HTK (accessed Sept. 21, 2021).

Coto-Solano,. R., Stanford, J., and, Reddy, S. (2021). Advances in completely automated vowel analysis for sociophonetics: using end-to-end speech recognition systems with DARLA. *Front. Artif. Intell.* 4, 1–19. doi: 10.3389/frai.2021.662097

Evanini, K., Isard, S., and Liberman, M. (2009). *Automatic Formant Extraction for Sociolinguistic Analysis of Large Corpora*. Brighton, UK: Interspeech. p. 1655–1658. Available online at: http://www.evanini.com/papers/evanini_INTERSPEECH09b.pdf (accessed April 28, 2022).

Fraser, H. (2014). Transcription of indistinct forensic recordings: problems and solutions from the perspective of phonetic science. *Linguagem e Direito*. 1, 5–21. Retrieved from: https://ojs.letras.up.pt/index.php/LLLD/article/view/2429

Fraser, H. (2019). Enhancing' forensic audio: what if all that really gets enhanced is the credibility of a misleading transcript? *Aust. J. Forensic Sci.* 52, 465–476. doi: 10.1080/00450618.2018.1561948

Fraser, H. (2020). Introducing the research hub for language in forensic evidence. *Judicial Offic. Bull.* 32, 117–118.

Fraser, H., and Loakes, D. (2020). "Acoustic injustice: the experience of listening to indistinct covert recordings presented as evidence in court", in *Law, Text, Culture (special issue "The Acoustics of Justice: Law, Listening, Sound")*, eds M. San Roque, S. Ramshaw, and J. Parker (Wollongong: The University of Wollongong). p. 405–429.

Gittelson, B., Leeman, A., and Tomaschek, F. (2021). Using crowd-sourced speech data to study socially constrained variation in nonmodal phonation. *Front. Artif. Intell.* 3, 1–7. Article. *No*. 565682. doi: 10.3389/frai.2020.565682

Gonzalez, S., Grama, J., and Travis, C. (2020). Comparing the performance of forced aligners used in sociophonetic research. *Linguistics Vanguard.* 6, 1–13. doi: 10.1515/lingvan-2019-0058

Gonzalez, S., Travis, C., Grama, J., Barth, D., and Ananthanarayan, S. (2017). "Recursive forced alignment: a test on a minority language," in *Proceedings* *of the 17th Australasian International Conference on Speech Science and Technology*, Epps, J., Wolfe, J., Smith, J., and Jones, C. (eds). *ASSTA Inc: Sydney*. p. 145–148.

Jessen, M. (2008). Forensic phonetics. *Language and Linguistic Compass.* 2, 671–711. doi: 10.1111/j.1749-818X.2008.00066.x

Jones, C., Li, W., Almeida, A., and German, A. (2019). Evaluating cross-linguistic forced alignment of conversational data in north Australian Kriol, an under-resourced language. *Lang. Doc. Conserv.* 13, 281–299. Available online at: http://hdl.handle.net/10125/24869

Kisler, T., Reichel, U., and Schiel, F. (2017). Multilingual processing of speech via web services. *Comput. Speech Lang.* 45, 326–347. doi: 10.1016/j.csl.2017.01.005

Koenecke, A., Nam, A., and Lake, E. (2020). Racial disparities in automated speech recognition. *PNAS.* 17, 7684–7689. doi: 10.1073/pnas.1915768117

Labov, W., Rosenfelder, I., and Fruehwald, J. (2013). One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and reanalysis. *Language.* 89, 30–65. doi: 10.1353/lan.2013.0015

Lindh, J. (2007). Semi-automatic aligning of swedish forensic phonetic phone speech in praat using viterbi recognition and HMM. *Proceed. IAFPA. 2007*. Plymouth, UK: The College of St Mark and St John.

Lindh, J. (2017). *Forensic Comparison of Voices, Speech and Speakers: Tools and Methods in Forensic Phonetics*. PhD dissertation. University of Gothenburg.

Mackenzie, L., and Turton, D. (2020). Assessing the accuracy of existing forced alignment software on varieties of British English. *Linguistics Vanguard.* 6. doi: 10.1515/lingvan-2018-0061

Malik, M., Malik, M. K., Mehmood, K., and Makhdoom, I. (2021). Automatic speech recognition: a survey. *Multimed. Tools. Appl.* 80, 9411–9457. doi: 10.1007/s11042-020-10073-7

McCarthy, J. (2007). *What is Artificial Intelligence?* Available online at: http://www-formal.stanford.edu/jmc/whatisai/whatisai.html (accessed Sept 14, 2021).

Opiah, A. (2021). *Descript Audio and Podcast Platform Review TechRadar Pro.* Available online at: https://www.techradar.com/au/reviews/descript (accessed October 11, 2021).

O'Shaugnessy, D. (2008). Automatic speech recognition: history, methods and challenges. *Pattern Recognit.* 41, 2965–2979. doi: 10.1016/j.patcog.2008.05.008

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., et al. (2011). *The Kaldi Speech Recognition Toolkit.* Hawaii: Paper presented at the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.

Rose, P. (2002). *Forensic Speaker Identification*. London, UK: Taylor and Francis. doi: 10.1201/9780203166369

Schiel, F., Draxler, C., Baumann, A., Elbogen T., and Steen, A. (2012). *The Production of Speech Corpora*. Available online at: www.bas.uni-muenchen.de/Forschung/BITS/TP1/Cookbook (accessed 21 Sept, 2021).

Villarreal, D., Clark, L., Hay, J., and, Watson K. (2020). From categories to gradience: Auto-coding sociophonetic variation with random forests *Laboratory Phonology* 11, 1–31. doi: 10.5334/labphon.216

Wassink, A.B., Gansen, C., and Bartholomew, I. (2022). Uneven success: automatic speech recognition and ethnicity-related dialects. *Speech Commun.* 140, 50–70. doi: 10.1016/j.specom.2022.03.009

Watt, D., and Brown, G. (2020). *Forensic Phonetics and Automatic Speaker Recognition. The Routledge Handbook of Forensic Linguistics.* London: Routledge. p. 400–415. doi: 10.4324/9780429030581-32

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

SYSTEMATIC REVIEW
published: 22 July 2022
doi: 10.3389/fcomm.2022.898410

A Framework for Deciding How to Create and Evaluate Transcripts for Forensic and Other Purposes

*Helen Fraser**

Research Hub for Language in Forensic Evidence, School of Languages and Linguistics, The University of Melbourne, Melbourne, VIC, Australia

Transcripts are used successfully in many areas of contemporary society. However, some uses of transcripts show systemic problems, with significant negative consequences. The key to finding effective solutions in these areas is to determine which factors contribute most strongly to the problems – which may be different from those to which they are commonly ascribed. This systematic review offers a conceptual framework for understanding the nature of transcripts in general, and the factors that contribute to a transcript's reliability and suitability for purpose. It then demonstrates how the framework can explain the (mostly) successful use of transcripts in two domains: court proceedings and linguistics research. Next, it uses the framework to examine two problematic cases: transcripts of forensic audio used as evidence in criminal trials, and transcripts of police interviews with suspects. A crucial observation is that, while it is common, and understandable, to focus on the transcriber as the source of problems with transcripts, transcription is actually a complex process involving practitioners in multiple roles, of which the transcriber role is not always the most important. Solving problems thus requires coordination of a range of factors. The analysis ends with practical suggestions for how to seek solutions for both the problematic areas reviewed, with attention to the role that linguistic science needs to play. The conclusion amplifies recent calls to consolidate transcription as a dedicated field of study within linguistics.

Keywords: transcription, transcript reliability, forensic, legal, verbatim reporting, covert recordings, police interviews, linguistic analysis

OPEN ACCESS

Edited by:
Dominic Watt,
University of York, United Kingdom

Reviewed by:
Alison May,
University of Leeds, United Kingdom
James Tompkinson,
Aston University, United Kingdom

Correspondence:
Helen Fraser
helen.fraser@unimelb.edu.au

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *17 March 2022*
Accepted: *16 June 2022*
Published: *22 July 2022*

Citation:
Fraser H (2022) A Framework for Deciding How to Create and Evaluate Transcripts for Forensic and Other Purposes. Front. Commun. 7:898410. doi: 10.3389/fcomm.2022.898410

1. INTRODUCTION

Transcripts are an essential part of our literate culture, providing a convenient and lasting record of otherwise ephemeral spoken language (Olson, 1994). Their ubiquity and familiarity make transcription seem like a simple and unproblematic process. However, it has many hidden complexities which not only cause problems, but make those problems hard to identify and solve.

The focus of the present paper is on transcripts used in legal contexts – specifically on transcripts of court proceedings, police interviews and covert recordings, as used in Australian and UK jurisdictions. As will be seen, while transcripts of court proceedings are mostly handled well (though with important exceptions), transcripts of interviews and covert recordings show systemic problems known to create a threat to justice (see Bucholtz, 2009; French and Fraser, 2018; Haworth, 2018).

Transcripts are also used in many branches of linguistic research, such as phonetics (e.g., Heselwood, 2013), language description (e.g., Himmelmann, 2018), conversation analysis (e.g., Hepburn and Bolden, 2012), discourse analysis (e.g., Edwards, 2008) – and indeed in studies of language used in the legal process (see Coulthard et al., 2020). However, with some notable exceptions (see Jenks, 2013), transcription is usually discussed in relation to specific branches of linguistic research, rather than as a general topic in its own right. This is unfortunate, as it means scholars may lack awareness of relevant issues from other branches, making it more difficult to determine the best solution for problems such as those mentioned above.

This systematic review aims to consolidate transcription as a dedicated field of research spanning multiple branches of linguistic science (cf. Fraser, 2020b). It starts by drawing together research findings about transcription, some of which, though well established, are subject to substantial misconceptions outside their own specialised areas. It then outlines a general framework for thinking about the stages involved in creating and using a transcript, and the factors that need to be managed at each stage to ensure a reliable product suitable for its purpose. Next it shows how consideration of the factors can help explain the successful use of transcripts in two very different contexts: court proceedings and linguistic research. Finally it uses the factors to identify the causes of systemic problems with transcripts of forensic audio and of police interviews, and to offer suggestions for effective solutions. A strong theme is that developing effective solutions for these serious problems requires the linguistic sciences not just to apply existing knowledge but to generate new knowledge.

It is natural for linguists to focus on solving problems by improving the actual transcripts used. However, the framework offered here shows that the quality of the transcript may not be the only, or even the main, cause of problems. Further, where improved transcripts are needed, emulating the kinds of transcript used in linguistics may not be the best approach. As discussed in detail throughout this paper (especially Sections 2.4–5 and 5.2), a major finding traversing all branches of linguistics is that no transcript is universally valid: each must be tailored for its context. Legal contexts differ substantially from the contexts of traditional linguistics research. For example, in many legal contexts, even if the transcript is created by a linguist, it is used by a third party who interprets it under conditions not controlled by the linguist.

Transcription in legal contexts, then, requires accountable, evidence-based methods designed to ensure reliable interpretation in relation to their specific purposes and the specific conditions under which they will ultimately be used. Achieving this requires "end-to-end" research, that considers all the factors affecting the system as a whole. This poses new challenges for linguistics – and the high stakes of the criminal justice system means failure to meet them fully has serious consequences. Success in meeting the challenges, however, has value beyond legal contexts. Improved understanding of transcription as a general process promises benefits for the many other branches of linguistic science whose research depends on transcripts.

2. WHAT IS A TRANSCRIPT?

2.1. Transcription vs. Writing

A transcript is a representation of spoken language using the symbols of written language. It is important to distinguish transcription from writing, which itself is often taken to be a representation of spoken language. However, while this view is fostered by (and, arguably, needed for) primary literacy acquisition, it is not technically correct (Daniels and Bright, 1996). Writing and speaking are completely different ways of representing linguistic meaning (Ong, 1982). It is true that, to count as writing (as opposed to a picture, for example) a representation must have a systematic relationship to the sound system of the particular language it represents (DeFrancis, 1989). However, that relationship is indirect and partial – nothing like the direct representation of individual "sounds" with letters that many assume it to be on the basis of literacy education (Linell, 1988; Gillon, 2007).

A transcript, then, is unlike writing precisely in that it *does* aim to create a direct representation of the words (and sometimes the sounds, gestures or other elements) that were actually used by a speaker during a specific speech event – after that event has taken place. Interestingly, however, as discussed in detail below, no transcript can fully achieve this aim. A transcript gives a valuable way to recall and refer to spoken language, but can never substitute for the speech itself. A useful analogy (see Fraser and Loakes, 2020) is that a transcript is like a map. No map can ever give a full account of the territory is represents, and any map is valuable only to the extent it helps its end-users fulfil their needs. The same, this paper will argue, is true of transcripts.

2.2. Verbatim Reporting

While there are many forms of transcript, we can introduce some key concepts by starting with the simplest: the verbatim report. Verbatim reports aim to represent each speaker's utterances, word by word, in ordinary spelling. They are now typically made from audio recordings. However, it is worthwhile to start by considering the traditional process: transcribing from live speech.

Writing down spoken language word by word seems simple in principle, but in practice it can be very hard. The most obvious difficulty is the speed at which spoken language is produced. No one can write quickly enough to capture all the words in real time – unless the speaker artificially slows down production, as in a schoolroom spelling exercise. At normal speaking rates, though a listener may recall the gist of what was said, the actual words are usually forgotten faster than they can be spelled out (Gurevich et al., 2010).

Transcription therefore requires an intermediate stage: creation of a temporary "record" of what was said, which can then be "written across" (the etymological meaning of "trans-scribe") to create the "verbatim transcript". The simplest way to make an intermediate record is by taking rough notes to use as an "aide memoire" (aid to memory). However, even with the aid of notes, it is hard to reconstruct the exact words the speakers used. Further, to the extent it can be done, there is no way to check for accuracy, except by comparing the memories – or notes – of other participants. The resulting "transcript" has, at best, the character more of meeting minutes than of a verbatim record.

The need for accountable verbatim transcripts of official events led to development of special ways of capturing the intermediate record quickly and accurately: stenography ("narrow writing") or shorthand. The skill of taking shorthand, and the techniques and procedures needed to transcribe shorthand into a text suitable for the readers who will eventually use it, were perfected over centuries, and professional stenographers have been in regular use in English courts, and other institutions, since the 1700s (Scharf, 1989). Since then, verbatim reporting has grown into the major world-wide industry our society relies upon today (e.g., intersteno.org). However, the increasing availability of practical audio recording techniques has seen reliance on stenographers gradually giving way to transcription from audio. Among other effects, this has highlighted some misconceptions about the nature of transcription.

2.3. Verbatim Transcripts From Audio

Those who have never tried transcribing from audio often assume it is easy, at least for a clear recording. After all, it solves the problem of speed faced by "live" transcribers. The audio captures a full record of exactly what was said, which can be paused and replayed at will, making transcription seem like a basic task, requiring little more than ability to spell.

The interesting thing is, however, that end-users often complain that the quality of transcription from audio is lower, not higher, than that of the apparently more difficult live transcription. The reason is that, on the assumption that "having the audio" makes transcription easy, managers tend to hire transcribers with lower qualifications than professional stenographers, and seek to increase output by farming work out to available transcribers, so that each transcribes short sections of multiple unrelated recordings.

The point is that, though the speed of speech may be the most obvious difficulty of transcription, it is not the only difficulty (Fraser, 2021a). So while the change to audio solves one problem, it creates others, especially by taking the speech out of its original context. The reasons are summarised in the next section; for extended discussion, see Fraser and Loakes (2020).

2.4. Transcription Is Not Transduction

The expectation that transcription should be easy reflects the everyday misconception that it is a mere transduction, in which words are mechanically copied from spoken to written form, and back again. This "transduction misconception" is incorrect, but nevertheless retains a powerful hold on common knowledge.

In this, it is similar to the widespread misconception that translating or interpreting from one language to another is a mechanical substitution of words in the source text with equivalent words of the target language. Actually, of course, translating and interpreting are complex skills, requiring many expert choices to be made in light of detailed understanding of the content and context of the material being translated (cf. Munday, 2016). That is why a translation is never "the" translation but always "a" translation – as demonstrated by the fact that back-translation (translating a translation back into the original language) typically creates a text quite different from the original.

What is less commonly noted, though on reflection it is perfectly evident, is that reading a transcript aloud (a process that could reasonably be called "back transcription") creates a speech event quite different from the original. This highlights the fact that a transcript, too, is never "the" transcript, but always "a" transcript. Speech is a massively complex signal, and it is impossible to represent it in its totality, even with specialised phonetic symbols (Heselwood, 2013). Transcribing speech into written text (like mapping a territory) requires many choices to be made regarding which elements to include, and how to represent them. Consider, for some simple examples: whether to include or omit false starts, self-corrections or hesitation markers; whether to represent colloquial or dialectal expressions with standard spelling or special symbols.

The effect is that any speech event can be represented in multiple ways, each with its own flavour. In fact, it is rare for two transcripts of the same material to be exactly the same. This gives linguists who teach transcription a handy way to detect cheating, as identical transcripts are likely to indicate that one has been copied from the other, despite student protests that they both independently "got it right". Similar reasoning, in a far more serious context, is discussed by Coulthard et al. (2017) p. 116–120.

These and other considerations demonstrate that transcription from audio, far from being a simple transduction, is an especially complex form of symbolic representation, well named as "entextualisation".

2.5. Entextualisation

The term "entextualisation" is relatively new (Urban, 1996; Park and Bucholtz, 2009), but the process has been researched for many decades (Ochs, 1979; Jefferson, 2004). One of the major findings is that producing verbatim transcripts requires context-sensitive interpretation by practitioners who are necessarily deeply embedded in specific social, cultural and political situations.

Much entextualisation research has focused on demonstrating that, despite this context-dependence, transcripts of official proceedings are often presented as "the" transcript – a manifestation of the transduction misconception that serves the interests of politically dominant elites, by treating the official transcript as objective, factual and neutral when really it reflects a particular point of view (Green et al., 1997; Roberts, 1997; Bucholtz, 2000).

This is important work – but the transduction misconception has other effects too. Erasing the role of the transcriber (Eugeni, 2020) diminishes respect for the many skills that professional transcribers bring to their task, meaning they may not receive the training and conditions they need to do an excellent job, as discussed above.

Another issue becomes particularly significant with transcription from audio. It is not only conscious choices that affect how words are represented. Context-sensitive interpretation, operating below the level of consciousness, plays a far larger role in speech perception than most people realise. For a famous example, the same stretch of speech can be heard as "recognise speech" or "wreck a nice beach", depending on the listener's contextual understanding (see Fraser and Loakes, 2020). This is one of the factors that limited computer speech recognition in early decades. Development of practical systems had to await the technical ability to build contextual prediction into the programming (Pieraccini, 2012). Even now, automatic transcription, while valuable as a labour-saving measure, is typically only useful for relatively clear speech with well-separated turns (Loakes, 2022), and even then, accuracy requires careful editing by a human who understands the context and intended content (Love, 2020).

However, while the role of contextual information is by now well established in speech perception research, the ubiquity of the transduction misconception means that transcripts are often produced with inadequate control over the conditions that affect their quality. We have seen, for example, that working hour by hour on recordings from different trials simply does not allow a transcriber to build up sufficient contextual understanding. Similar issues are a major cause of the systemic problems that this paper seeks to address. Identifying and solving such problems requires recognising transcription as a skilled practice which takes place as part of a complex process involving context-sensitive interpretation at multiple levels, by practitioners in multiple roles. The next sections aim to contribute to this recognition, by suggesting a framework that sets out the main components of the complex process of transcription, and examining the factors that affect the quality of the resulting transcript.

3. A FRAMEWORK FOR UNDERSTANDING THE FACTORS THAT AFFECT THE QUALITY OF A TRANSCRIPT

The framework suggested here is based on the understanding, discussed above, that transcription requires three stages, which may be performed by different practitioners, or by one practitioner taking different roles:

- Stage 1: capturing an intermediate record;
- Stage 2: producing a transcript; and
- Stage 3: interpreting and using the transcript.

The reliability of a transcript is often attributed directly to the accuracy of the transcriber at Stage 2. However, it is important to pay explicit attention to all stages, each of which, as we will see, is subject to substantial misconceptions. In particular, each tends to be treated as transduction, when in fact all of them require context-sensitive, and often content-aware, interpretation.

Stages 1 and 2 require practitioners to "abstract" the information that seems relevant, in light of their understanding of the purpose of the transcript, from the overall context. This results in the "decontextualisation" of the transcript often emphasised in the entextualisation literature. One effect is that only information abstracted at earlier stages is available at later stages, making it easy for errors to propagate from one stage to the next. In order to understand and use the decontextualised transcript, at Stage 3, the end-user has to "recontextualise" it, relying on knowledge, or assumptions, from various sources. Komter (2019) gives an especially clear account of these processes and their effects.

It is sometimes suggested that this reliance on context means transcription is necessarily "subjective" or even "biased". However, these terms have multiple meanings, some with negative connotations which are not always appropriate for transcription. For example, "bias", in its primary sense, suggests a conscious or unconscious intention to privilege interpretations that suit the practitioner's interests. Bias in that sense can certainly affect any stage of transcription, with seriously undesirable consequences. That makes it essential to manage the transcription process so as to minimise opportunities for self-interest to be served. (The fictional account in Hannelore Cayre's 2019 novella "The Godmother" gives an entertaining and not entirely implausible insight into the advantage an individual can take of a system with lax control.)

Managing bias has traditionally relied on security clearances and quality control. More recently, however, there has been a tendency to believe that it requires withholding contextual information from practitioners. This may be due to popularisation of the term "cognitive bias" for a range of psychological effects that do not necessarily involve self-interest (Kahneman, 2011). This usage has led some to believe that any context-awareness is necessarily biasing, and should therefore be eliminated. This is unfortunate. For reliable transcription, as for most other aspects of linguistic analysis, relevant, reliable contextual information is essential. Attempting to withhold all contextual information from practitioners can actually introduce biases of different kinds, which are even more difficult to manage effectively. The important thing, rather, is to ensure practitioners receive relevant and reliable contextual information, in a managed process, without exposure to potentially misleading information (cf. Dror et al., 2015).

Similar ambiguity surrounds use of the term "subjective". Here the primary sense suggests personal preference influenced by an individual's feelings or tastes – which is clearly not appropriate in scientific analysis. Avoiding subjectivity in this sense is often thought to require "objectivity". The problem is that this term, too, has different interpretations. Often it is understood in the sense of requiring only context-independent measurement of observable physical features. However, by now it is well established that, even in the so-called "hard" sciences, observations and measurements are rarely fully "objective" in this strong sense (Hoffman, 2019; Ritchie, 2020). Almost all require human judgment (Kara, 2022). Trying to pretend they do not merely allows hidden biases to have uncontrolled and potentially damaging effects (D'Ignazio and Klein, 2020; Fry, 2021).

Striving for "objectivity" in that unrealistic – and outdated – sense, then, may be counterproductive for some sciences, especially for human sciences involving analysis of language. The important thing for scientific reliability in such fields is not to deny the role of human judgment, but to ensure that important judgments are made by a disinterested expert in relevant disciplines, who has full possession of relevant reliable contextual information, carefully managed to preclude potentially misleading expectations, and can explain and justify their opinion in a transparent and accountable manner. To use the term "subjective" for the view of such an expert fails to distinguish it appropriately from a casual expression of personal preference. Perhaps some updated terminology is required in this area.

With these general remarks, we turn now to consideration of the factors that affect the overall enterprise of transcribing from audio, at each of its stages.

4. FACTORS AFFECTING THE CREATION AND USE OF TRANSCRIPTS

This section aims to set out some of the factors that affect the creation and use of transcripts of various kinds, with the focus on transcribing from an audio recording. The intention here is to present an overview for convenient reference, with examples and details in later sections. Of course, while it is useful to set the factors out separately, as this allows them to be considered methodically, they all interact extensively. The particular way they have been categorised here is influenced by the current focus on specific types of transcripts used in the legal process, and there are certainly other ways of conceptualising them (cf. Richardson et al., 2022). Indeed the present framework differs from, and supersedes, my own previous account (Fraser, 2014).

One key point that will be emphasised is that each factor involves expertise in a specialised field. Currently, few in linguistics have full expertise in all relevant fields, with a particular gulf between phonetics and other branches. Thus the discussion below does not claim to give definitive coverage of every factor, merely to indicate relevant considerations for each. Another key point is that all factors are heavily influenced by practitioners' practical understanding of the purpose and context of their work at that stage – which can be influenced by knowledge or assumptions they may not be consciously aware of. In short, the output of each stage is never "the" output but only "an" output. However, though specialists in each factor are well aware of this fact, others have a strong tendency to over-simplify, with the transduction misconception being a particular problem through all stages.

4.1. Stage 1: Capturing the Audio Record

4.1.1. Audio Factors

Audio factors affect how the speech is abstracted from its context, and preserved for later listeners in an audio recording (with or without video). It is important to recognise that no audio is ever neutral. Like a photograph, a recording necessarily reflects the viewpoint of the one making it. So an essential overarching factor is the recording practitioner's understanding of the purpose and context of the recording – which influences many decisions that affect the ultimate nature of the audio.

There are also numerous factors that affect the technical quality of the audio. These include the type of equipment being used, as well as the practitioner's knowledge of how to use it, and ability to control how it is deployed. It is also important to take account of any processing applied to the audio, whether at the time of recording, or later. For example, it is often assumed that "enhancing" indistinct audio makes it "clearer", but this is not always true, and, again, the misconception can have negative consequences (Fraser, 2020a). For example, reducing background noise can have the undesirable effect of making listeners more, not less, likely to accept an inaccurate transcript (for a quick and compelling demonstration see Fraser, 2019).

4.1.2. Speech (and Speaker) Factors

Speech factors include the language, variety, register and style of the speech captured in the recording – all reflecting the speakers' purpose, which, in almost all situations, is to make their meaning intelligible to intended or expected listeners. For "overt" (open) recordings, speakers may have awareness not just of listeners who are present at the time of the recording, but also of potential future listeners to the audio (cf. Haworth, 2013). In "covert" (secret) recordings speakers are typically aware only of the immediate listeners – though sophisticated criminals may consider possible hidden listeners, and attempt to disguise their meaning or identity.

An especially important factor is the location of the speech on the spectrum of formality. Informal conversation typically features overlapping and incomplete utterances, and is often highly elliptical, since listeners present at the time can rely for comprehension on implicit reference to aspects of the immediate context. However such references will be unavailable to those listening later to the decontextualised recording, potentially making the speech difficult to understand (video may help to some extent, assuming it is of good quality and designed to capture all relevant contextual information).

Since formal speech typically makes less reference to the immediate context, and is more likely to feature speakers taking separate turns, it may be intelligible even when technical quality is poor. Less formal conversation, however, may be heard inaccurately even with a good quality recording (Fraser and Loakes, 2020). A related factor is the pragmatic nature of the speech. For example, speech used for basic information exchange may be more readily represented in a verbatim transcript than nuanced social or emotional functions requiring subtle use of intonation and voice quality.

4.2. Stage 2: Producing the Transcript

4.2.1. Transcriber Factors

As we have seen, a recording is already an abstraction of the speech from its original context. Transcription involves further abstraction of the information needed to construct words and other linguistic entities from the recorded speech, and represent them in written form.

Perhaps the most obvious factor here is the practitioner's level of training and testing in the technicalities of the specific style of transcript required. Equally important, though harder to test, is the practitioner's personal aptitude for transcription. No transcript is ever "one and done". All require significant concentration for repeated listening, with or without feedback from an evaluator (Section 4.2.3), and continual reviewing and updating of their work to reach a point of personal satisfaction that it is of appropriate accuracy for the context. Another crucial factor, as always, is the transcriber's understanding of the purpose of the transcript, which affects many decisions about what aspects of the speech to include, and how to represent them.

4.2.2. Listener Factors

The "listener" here is not the listener to the original speech, but the listener to the recording. This is, of course, the same person as the transcriber, but in a different role. Indeed the listener role is arguably the most important role of all stages: after all, transcribers can only transcribe what they hear. Nevertheless it is one of the most overlooked roles of the entire transcription process, subject to many misconceptions.

One obvious factor is the listener's knowledge of the language, variety and register used by the speakers in the recording. Important as this is, however, it is only one factor – we cannot assume that anyone who knows a particular variety will automatically be good at transcribing any recording in that variety, especially if they have not been independently tested for aptitude under relevant conditions.

Another set of factors includes the listener's knowledge and expectations about the content and context of the recording, which, as outlined in Section 2.5 above, can have a large but typically unnoticed effect on perception, especially of audio with any degree of indistinctness. Again, however, while reliable contextual expectations can be helpful in understanding difficult audio, we cannot assume that those with reliable contextual knowledge will automatically create a reliable transcript – as this factor interacts strongly with aptitude and other factors.

A further important but little-recognised danger is that *unreliable* contextual expectations can be highly misleading, resulting in confident but inaccurate perception. Burridge (2017) gives a quick and accessible introduction to this concept, with entertaining examples showing just how easy it is for listeners to "hear" words that are not really there. Unfortunately, while examples like these are well known for their humour, their serious implications for transcription are not always fully recognised outside the specialised field of speech perception. This means that transcribers' contextual expectations are not always managed as diligently as they should be – a source of the problems discussed in Section 6.

4.2.3. Evaluator Factors

As mentioned above, a certain amount of personal evaluation is undertaken as part of the transcriber role. Some transcription situations also require external evaluation of the transcript, e.g., via a test used for accreditation or quality control. In such cases, there are additional factors to consider. One, clearly, is the evaluator's independence, understanding of their role, and knowledge of the factors that might influence their judgement.

Appropriate decisions about details of the test are also crucial. For example, it matters what the transcript is evaluated against – e.g., a known correct transcript, the evaluator's memory of what was said, or the audio itself. Particularly difficult issues arise in the last situation, since the very act of viewing the transcript in order to check it can affect the listener's interpretation of the audio (Section 6.1.1). Unfortunately, however, while the role of such decisions is well understood in language testing (e.g., Knoch and Macqueen, 2020), transcript evaluation has not yet developed a sophisticated methodology.

4.3. Stage 3: Using the Transcript

4.3.1. End-User Factors

Another often-overlooked consideration is how the eventual transcript is actually used in practice by its end-user (the linguist, lawyer, jury, etc., who ultimately interprets its content). After all, even the best transcript can be used wrongly or inappropriately (just as an excellent map can fail if the end-user does not understand its capabilities and limitations – see Section 2.1).

The first factor to consider, as always, is the end-user's intention and purpose in using the transcript – which may or may not be the same as the intention and purpose of practitioners at other stages. Another is the end-user's understanding of the nature of transcription in general. Are they simply picking up "a" transcript and treating it as "the" transcript? Or are they considering appropriately whether this particular transcript is suitable for their purpose? If the latter, do they have sufficient knowledge of the transcript's provenance to be able to assess its suitability, and take account of its (inevitable) limitations? Finally, the end-user's ability to interpret any specific transcription conventions is important.

4.3.2. Overall System-Design Factors

Considering end-user factors raises the need to consider the transcription process as a whole, by evaluating the factors that affect each stage, and assessing the extent to which the overall system is working as intended. Ideally this would be done as part of the design and management of a system created in pursuit of a unified overall purpose, with appropriate consultation of those with expertise relevant to each stage. Alternatively, it could be done "post hoc", by retrospectively reviewing the factors that have contributed to the quality of the transcript and the end-user's ability to use it appropriately. Either way, it should be undertaken with full understanding of the expertise that is required of practitioners at each stage, and all the factors that contribute to the output.

However it can happen that neither of these kinds of system evaluation are undertaken effectively – or at all. Section 6 considers two such situations: transcripts of police interviews and forensic audio, and their propensity to induce errors with far-reaching negative implications for our criminal justice system. First, however, we consider two situations where the transcription process is (with important exceptions) designed, evaluated and used well: court transcripts and research transcripts. This will help in determining the key factors that contribute to successful creation and use of transcripts.

5. USING THE FRAMEWORK: TWO (GENERALLY) SUCCESSFUL EXAMPLES

This section demonstrates use of the framework by looking at two kinds of transcripts that serve very different purposes: transcripts of court proceedings, and transcripts used in linguistics research. In each case, the transcripts are generally successful in serving their purpose – though, as we will see, both are subject to serious failings if particular factors are not managed appropriately. Discussion will demonstrate that success arises not from any single factor, but from pursuit of the transcript's overall purpose in light of well-informed, context-aware management of all relevant factors, along with careful, ongoing system evaluation.

5.1. Transcripts of Court Proceedings

The overall purpose of court transcripts is to create an official record of trial proceedings that can be used by anyone, and is trusted by all. Here we briefly consider the factors that affect the outcome, focusing first on the traditionally monolingual situation of Australia and the UK.

Most of the key speakers in a trial use relatively standard English, though individual witnesses may have a range of different dialects (witnesses who speak languages other than English are provided with an interpreter – at least in principle, if not always in practice: e.g., Cooke, 2009). Most speakers also use relatively formal language, monitored by the judge to ensure that everyone talks in turn, and all speak up clearly "for the tape". Much of the speech involves basic information exchange – with departures from this usually evident from subsequent turns.

The audio quality is typically fair. Together these factors mean the recording is mostly easily intelligible by transcribers familiar with the courtroom genre, though listeners may have difficulty in making out unfamiliar names or technical terms.

Court transcribers are accredited to ensure they have the necessary skills for accurate verbatim transcription, and undergo security clearance to ensure their independence in relation to trial outcomes. They are also highly trained in the use of specific conventions appropriate to court transcripts, including how to "tidy up" the representation of spoken language (e.g., by eliminating hesitation markers or false starts) to make it easier for end-users to read, and to give a respectful impression of court-room discourse (cf. Voutilainen, 2018).

The transcriber in the role of listener typically knows the language, variety and register of the court (though not necessarily those of all witnesses, as noted below), and is provided with names and technical terms, as well as general contextual information, to assist in perception of unpredictable content. Evaluation of individual transcripts is undertaken by the lawyers and judges who took part in the trial – in light of their memory of what took place, and their understanding of what information court transcripts should capture. The end users are readers who understand the transcription conventions and the courtroom context. As mentioned earlier, the overall system has been designed over centuries with ongoing evaluation and development aimed at ensuring that court transcripts meet the needs of society, or at least of its dominant sectors (cf. Section 2.5).

Not surprisingly, given all these circumstances, courtroom transcripts are, in general, well suited to their purpose, and mostly of high quality – at least in the monolingual scenario for which the factors have been optimised. The fact that substantial problems have been demonstrated in representing the speech of witnesses with non-standard dialects (Walsh, 1995; Jones et al., 2019) shows that court transcription processes, despite their long history, have been designed without full understanding of all relevant factors.

What is interesting to note now is that their general suitability for their own purpose does not imply that court transcripts are universally suitable for every purpose. In particular, they have substantial limitations when used as the basis of linguistic research on courtroom interaction, as discussed next.

5.2. Transcripts for Linguistic Research

Transcripts are used in many branches of linguistic research (some mentioned in Section 1 above). One that is of relevance here, and will enable exemplification of some general issues, is research on spoken interaction in court – aiming, for example, to demonstrate and theorise practices that create systematic disadvantage for certain categories of defendants (e.g., Eades, 2010; Mariottini, 2017).

The interesting thing is that court transcripts are generally not useful for this kind of research – precisely because they are not, in fact, strictly "verbatim" in the sense of representing each word as it was spoken (Eades, 1996). The "tidying up" undertaken by court reporters, though useful to intended end-users, can alter the very detail needed for the research. For this reason, researchers often choose to make their own transcripts – which of course are affected by their own set of factors.

Some factors are the same as for court transcripts. Research on courtroom interaction typically uses the courtroom recording, and the transcriber in the role of listener almost always knows the content with considerable certainty – as is true for almost all linguistic research.

Where the two differ sharply, however, is in the overall purpose of the transcript. Research transcripts aim, not to preserve the informational content of the speech for use by a generalised third party, but to represent and operationalise features of the spoken language for use by the transcriber (or close associates) in exploring whatever theoretical issues are under consideration. Thus while court transcripts are an end in themselves, linguistic transcripts are a means to an end: after peer review and publication, the transcripts themselves are rarely referred to again, unless to critique the research.

The transcriber is trained to focus on aspects of spoken language relevant to the research, and to annotate them via special formatting and technical symbols whose meaning and use must be learned via advanced education. Very importantly, however, these technicalities are an addition to, not a substitute for, reliable representation of the verbatim content. While technical symbols may impress outsiders, they can mask errors that reduce the overall reliability of the transcript. Also importantly, use of technical symbols does not imply the transcript is "objective" in the sense of being unbiased or neutral. It has long been known that research transcripts can display self-interested bias (Wald, 1995). For this reason, transcripts used in high-stakes research are usually subject to external evaluation, typically via inter-rater reliability checks, which compare transcripts from several transcribers, each with relevant expertise and knowledge of the overall purpose of the research – but "blinded" as to context that might engender bias.

5.3. Discussion

Both court and research transcripts are highly successful in their own domains – though not infallible, as we have seen. Indeed, the success of each comes precisely from its recognition of the potential for error, which motivates management of known risk factors, and commitment to ongoing independent evaluation and improvement of the system.

However, while these two types of transcript are successful in their own domains, they are very different – and not interchangeable. We have seen that court transcripts are generally not useful for linguistic research. Less obviously, perhaps, research transcripts are not useful as court transcripts. Importantly, this is not only because court transcribers and end-users lack the skills needed to produce and understand technical linguistic representations. Linguistic transcripts, like any others, require choices to be made, in light of context-aware understanding of their overall purpose, about what detail to include, and how to represent it. That is why linguists' transcripts can rarely be transferred from one research project to another (Jenks, 2013) – further reinforcement of the key insight, discussed above, that no transcript is a neutral representation.

This is important to emphasise here in light of the persistent misconception that certain kinds of technical transcripts can somehow capture the "objective truth" of what was said via "bottom-up" analysis. Such claims are sometimes made, for example, in relation to conversation analysis (CA). It may well be true that CA practitioners pursue data-focused analysis more diligently than some more "theory-driven" branches of linguistics. But this does not mean that CA transcripts are "neutral", or "objective" in the strong and outdated sense discussed in Section 3 – as CA experts themselves are at pains to acknowledge (Edwards, 2008; Hepburn and Bolden, 2012).

Even stronger claims of "objectivity" in the outdated sense are made for phonetic transcription. Again, however, experts are clear that such claims are overblown (Heselwood, 2013; Himmelmann, 2018). Indeed one of the best established findings of speech perception research is that "bottom up" word recognition is impossible. That is why, for example, expert phoneticians acknowledge that they have limited ability to transcribe languages they do not know, or to "read" spectrograms with unknown content (see Fraser, 2022 for extended discussion).

Of course, this is not to suggest that either of these kinds of transcription are "subjective" in the soft sense of reflecting mere personal preference. Nor does it suggest that not being "objective" in the outdated sense diminishes the value of CA or phonetic transcripts. To the contrary – both are highly valuable in the contexts for which they are developed. What is essential, however, is to acknowledge that valid use of their specialised symbols depends crucially on valid understanding, both of the context and content of the audio, and of the purpose of the transcript, being shared by both creator and interpreter of the transcript.

What makes a transcript reliable and useful, then, is expert judgment, exercised across all three stages, in a system designed to manage the complex intertwined factors that affect the suitability of the final product to the end-user's needs. It is this type of management that makes both linguistic and court transcripts successful – and it is in being the product of this kind of management that these two types of transcripts are similar, despite their many differences of style, content, layout, etc.

6. USING THE FRAMEWORK: TWO PROBLEMATIC EXAMPLES

With the insights of Section 5 in mind, it is now time to consider our two examples of transcripts being used in more problematic ways. Both forensic audio and police interviews start life as part of a criminal investigation, during which transcripts are used, if at all, in relatively unproblematic ways. Both, however, sometimes go on to serve as evidence in court, where transcripts can be used in ways that have been shown to create major problems for justice. This section aims to describe these problems, identify the factors that cause them, in light of the insights developed above, and discuss potential solutions.

The key observation will be that, while there has been an understandable tendency to focus on the transcriber as the main source of the problems, actually transcriber factors are only one part of the problem, and not necessarily the most important. So while expertise in linguistic science is essential to developing a better system for transcribing forensic audio, the expertise needed is not simply the ability to create technical linguistic transcripts. Rather expertise is needed to develop and manage an overall system that emulates, at a deep level, the practices that create successful transcripts – paying attention to all the factors, not just the superficial factor of being able to use technical symbols and terminology (Fraser, 2020c).

6.1. Transcripts of Indistinct Forensic Audio

Forensic audio is speech that has been captured, typically in a covert (secret) recording obtained as part of a criminal investigation, and is later used as evidence in a trial. Such recordings provide powerful evidence, allowing the court to hear speakers making admissions they would not make openly. One problem, however, is that the audio is often extremely indistinct, to the extent of being unintelligible without the assistance of a transcript.

Transcripts used to give this assistance are typically provided by police investigating the case, who, in court, are given the status of "ad hoc expert" on the grounds that they have listened to the audio many times. This is often found alarming by linguists, who suggest it would be better to have the transcripts produced by real experts. Surprisingly, however, insisting on expert transcripts, though surely an improvement, is not a fool-proof solution (Fraser, 2020b, 2021b). To gain an impression of the reasons, and to consider directions to look for better solutions, it is worth reviewing the factors that cause problems with police transcripts.

6.1.1. Factors Affecting the Reliability of Police Transcripts of Forensic Audio

The combination of very poor technical quality, and unmonitored, highly contextualised conversation means many covert recordings are essentially unintelligible to general listeners. The purpose of the transcript is to assist the court in perceiving the content, and thus in better understanding the context (i.e. the crime, and who is responsible for it).

Ad hoc experts have no training in transcription, and are not required to demonstrate skill. The reason they are asked to provide transcripts has to do with their role, not as transcriber, but as listener: they can often make out more of the content of indistinct audio related to their cases than other listeners can. Though the law attributes this ability to their having listened many times, the real reason is their access to contextual information – and it is important to acknowledge that reliable contextual information can sometimes help police understand specific utterances. As discussed in Section 4.2.2, however, mere access to contextual information cannot guarantee a reliable transcript. A particularly serious limitation on police transcripts is that not all contextual information available to investigators is reliable (that is why we need the trial). The powerful effect of contextual expectations on perception means that unreliable contextual information can easily mislead perception, without conscious awareness. For these reasons, police transcripts are rarely fully accurate, and often egregiously wrong (French and Fraser, 2018).

The end-user is the jury, who are instructed by the judge to listen carefully to the audio and form their own opinion as to its content, using the transcript only as assistance. Unfortunately, however, this is an unrealistic instruction. It is well known that an inaccurate transcript can easily "assist" listeners to hear words that are not there (Section 4.2.2). Indeed, the law is aware that police transcripts might be wrong, and a transcript is not provided as assistance to the jury until it has been evaluated. The problem is that the evaluation is carried out by lawyers checking the transcript against the indistinct audio, without realising that this very process inevitably subjects their own perception to the influence of a potentially misleading transcript (Fraser, 2018; Fraser and Kinoshita, 2021).

Finally, the overall system has been designed by judges, on the basis of their experience with court transcripts, with insufficient understanding of the factors that influence understanding of indistinct forensic audio. No system evaluation is undertaken. The whole process is driven, not by scientific values, but by legal precedent (Fraser, 2021b).

6.1.2. Discussion

Unsurprisingly, this process gives rise to serious problems, and numerous instances of injustice have emerged (for a quick introduction with an interesting connection to Section 6.2, see Fraser, 2013). However setting out the factors methodically has shown that the main cause of these problems is not the fact that transcripts are provided by investigators (though this is far from ideal). The problems are created by the system as a whole, with the most important factor being the fact that transcripts of indistinct forensic audio are evaluated by lawyers involved in the trial. Even transcripts provided by experts are evaluated by lawyers and judges, creating substantial problems (Fraser, 2021b). So the first step towards improvement must be to change the legal procedures that give so much credence to inexpert and unaccountable evaluation of transcripts (Fraser, 2020c).

The next step is to introduce processes for providing courts with reliable transcripts. Many have assumed that this can be achieved by individual experts evaluating police transcripts - as I did myself until casework experience led me to argue this it is not suitable, for a range of reasons (Fraser, 2020b). These reasons have recently been amplified by a ground-breaking study (Love and Wright, 2021) in which eight different (expert) transcribers of indistinct audio created eight transcripts that differ in substantial ways. The point is that the experts were operating under uncertainty regarding the true content of the audio. This of course is the standard situation with forensic audio – but very different from any kind of linguistic research (Section 5.2). Further, while acoustic analysis might confirm some parts as more or less likely to be right, the true content is unlikely to be established purely by "bottom up" analysis (Section 5.3). These differences clearly indicate a need for specialised system design.

Producing a reliable transcript of indistinct audio of unknown content needs methods beyond standard linguistic or acoustic analysis. To date, however, very little research has been directed explicitly towards developing such methods (see Fraser, 2022). New projects are needed to design an evidence-based process that can ensure all forensic audio used in court is provided with a reliable transcript (or certified as incapable of reliable transcription). Such projects need to take an end-to-end approach, to ensure the transcripts are suitable for the purpose of assisting a jury to understand the content under courtroom conditions (recognising there can be a major difference between the information an expert puts into a transcript, and the information end-users take from it).

We cannot leave this section without mentioning that indistinct covert recordings frequently feature languages other than English, which require not only reliable transcription, but also reliable translation. Unfortunately both of these tasks are carried out according to procedures developed with poor understanding of relevant aspects of linguistic science (Fraser, 2021b). Even more unfortunately, valuable efforts of experts to document the resulting problems (Capus and Griebel, 2021; Gilbert and Heydon, 2021) and suggest viable solutions (Gonzáles et al., 2012; NAJIT, 2019) are so far having limited impact on general practice.

6.2. Transcripts of Police Interviews With Suspects

We turn now to our second problematic example: transcripts of police interviews with suspects. Traditionally, these were created on the basis of an intermediate record made by officers taking notes about what the suspect said (cf. Section 2.2 above). This famously gave opportunities for "verballing" – police falsely claiming that suspects had made "verbal admissions" during the interview (Eades, 2010; Grant, 2022). In both Australia and the UK, Royal Commissions in the 1980s and 1990s sought to curtail opportunities for such "fabricated confessions", by instituting requirements that all police interviews with suspects should be audio/video recorded (Baldwin, 1985; Dixon, 2008). This is now gradually being extended to an expectation that police will use body-worn recording devices while interviewing witnesses or engaged in other duties (Roberts and Ormerod, 2021).

Electronically recorded interviews have many benefits. One disadvantage, however, is that recordings are not convenient to access or refer to. This makes it necessary to provide a transcript of each interview. Upon institution of compulsory recording, the large workforce needed for transcription was mobilised hastily and under severe cost constraints, often co-opting practitioners whose primary skills and responsibilities lay elsewhere. Unfortunately it was not till decades later that it was discovered that their transcripts sometimes contained egregious but undetected errors, with potential to affect justice (Haworth, 2018; Komter, 2019; Richardson et al., 2022).

Again, before considering solutions to this problem, it is useful to review the factors methodically, so as to ensure its key causes are identified properly.

6.2.1. Factors Affecting the Reliability of Police Interviews With Suspects

The audio quality of recorded police interviews is usually fair, and the style of speech is usually relatively formal and relatively well monitored. This means that the audio is usually reasonably intelligible – though typically well below the standard of recordings of court proceedings, making the task of interview transcribers harder than that of court transcribers. The audio quality of body-worn recordings can be particularly poor.

Despite the harder task they face, interview transcribers are rarely as well-qualified, nor as well-resourced, as court reporters. The fact that they are typically employed by police departments, or by agencies that undertake extensive police work, means they usually have contextual understanding of police and legal processes in general, and sometimes of specific cases. Nevertheless, various kinds of error are common, as well documented by Haworth (2018) and Komter (2019) – confirming that difficulties in understanding recorded speech are not limited to poor quality audio (Section 4.1.2).

Evaluation of interview transcripts is effectively non-existent. In principle, it is intended to be undertaken by lawyers, with the defence considered especially responsible for reviewing the transcript, as shown by the following advice for defence lawyers:

> It is important to watch the [video] or listen to audio tapes of records of interview. It will not only help you work out whether the transcript is accurate, but it may also indicate important aspects of the questioning and your client's manner and condition at the time of questioning which may be relevant in your case (for example, being intoxicated or not in a fit mental state) (NSW Young Lawyers Criminal Law Committee, 2004: 172).

Evaluation of transcripts by lawyers is not ideal, since they have neither the expertise nor the independence to undertake the task rigorously, making it unlikely that they would detect all relevant errors. Worse still, even this less-than-ideal evaluation is often skipped. Time pressures mean the advice below is not always followed – making it common for the transcript to be used as the definitive account of the interview, with the audio never being accessed at all, let alone used for careful evaluation of the transcript.

> Copies of your client's [recording] will not usually be included in the prosecution brief. You will generally be served only with a transcript of what was said in the [interview]. You should get a copy of your client's [recording] (NSW Young Lawyers Criminal Law Committee, 2004 p.284).

The end-user is the most complex factor in this situation. Typically, multiple parties use the transcript (cf. Haworth, 2013) – each with different needs. First, the police themselves may use it to aid their memory of what happened in the interview (though they may prefer their own notes). Then prosecution and defence solicitors use it, in preparing their cases, as a record of the information obtained during the interview. Next, if the interview is used as evidence in court, barristers quote from the transcript, using their own intonation and speaking style (Haworth, 2018). The final, and arguably most important, end-user, is the jury, who use the content of the interview, in combination with other evidence, to reach a verdict of guilty or not guilty. As is clear from the above account, however, they may understand the content only through a barrister's "back-transcription" (Section 2.4). Unlike the situation with forensic audio, there is no expectation or requirement that interview audio be played in court.

System design and evaluation are close to non-existent. Developed in haste, and with no input from relevant experts, the whole process was subject to little scrutiny until researchers like Haworth and Komter exposed some of its serious weaknesses:

> [I]n stark contrast to the strict principles of preservation applied to physical evidence, interview data go through significant transformation between their creation in the interview room and their presentation in the courtroom, especially through changes in format between written and spoken text (Haworth, 2018: 428).

6.2.2. Discussion

As with forensic audio, it is common for the failings of interview transcripts to be blamed on the transcriber. Again, however, it is clear from the above analysis that the problems lie in the system as a whole, which is designed and managed with insufficient attention to crucial factors. This means that the problems cannot be solved purely by seeking ways to ensure more reliable transcripts (though this is certainly an important part of the solution, as discussed shortly). After all, even an excellent transcript risks giving a misleading impression of the audio if it is read out by a barrister, selectively using intonation, pausing, etc., designed to persuade a jury to accept a particular version of what happened in the interview. Preventing this would seem to require working with the judiciary to reform practices for presenting interviews as evidence in courts – by demonstrating how essential it is for the court to listen to the actual audio.

Further, as discussed above, interview transcripts are not always excellent. It is really essential to ensure they are always of high quality. The question is how to achieve this. One common suggestion is to train interview transcribers to include more detail in their transcripts, perhaps creating a simplified version of the style of transcript used in branches of linguistics like conversation analysis (CA). However this suggestion raises several issues.

First, the value of a CA-style transcript is limited by the accuracy of the verbatim representation on which it is based (Section 5.3). If verbatim transcripts contain errors, adding technical detail will not help – and may actually mask deficiencies by making it even more difficult for listeners checking the transcript against the audio to notice errors (Section 4.3). The priority then, might be to ensure that interview transcribers produce reliable verbatim transcripts – not by insisting busy lawyers check the transcript against the audio, but by training, resourcing and managing interview transcribers in ways commensurate with courtroom transcribers (Section 5.1).

Second, learning even simplified CA transcription is difficult, especially for transcribers with no background in linguistics. While they may be taught some technicalities, they may retain misconceptions about language and speech that undermine their ability to use the teaching effectively (at least, this is a common outcome when training in phonetics is provided to assist English pronunciation teachers, see Burri et al., 2017).

Third, the detail in a CA transcript necessarily reflects the transcriber's understanding of its context and purpose (Section 5.3). This is not a problem for research transcripts, where end-users share the same context and purpose as transcribers. With interviews, however, end-users (especially lawyers on opposing sides) need to form their own independent interpretation of the interview in light of their own purposes, with minimal influence from the interpretations of others.

Finally, and most importantly, no transcript can represent all the information in the audio, as discussed at length above. Using any transcript, even one with detailed and accurate annotation, without reference to the audio, inevitably causes end-users to miss or misinterpret aspects of the content – as has now been powerfully demonstrated, specifically in relation to police interviews, by Deamer et al. (in press). In a worst-case scenario, an annotated transcript could even serve, intentionally or not, to manipulate end-users' understanding of what was said in the interview, especially when speech is nuanced, emotional or otherwise open to varying interpretation.

For all these reasons and more, it is really essential for end-users of interview transcripts to listen to the recording personally. Unfortunately, as we have seen, this rarely happens. While one reason is time-poverty, another is the transduction misconception. Lawyers on both sides simply accept that the transcript is essentially equivalent to the audio:

> [contamination of interview data] appears to stem from a lack of recognition that changes in the format of linguistic data involve transformation of the data themselves. A first step in improving current practice, then, is to increase awareness of that simple fact (Haworth, 2018: 445).

To persuade busy lawyers to listen to the audio, then, one approach might be to institute education, especially for those on the defence side, in which linguists can explain the falsity of the transduction misconception, and demonstrate how listening to the audio can reveal information that might help win a case – hopefully thus motivating solicitors to request video recordings at the start of each case (or, better still, to get them routinely without need for a request).

To make the listening more efficient, it may be worth noting that substantial proportions of police interviews are taken up with routine information-exchange, which can be understood relatively well from a standard verbatim transcript (Section 4.1.2). One suggestion worth exploring, then, might be to ask transcribers to draw the attention of lawyer end-users to parts that most need to be listened to, simply via marginal notes indicating sections of the transcript where the language diverges, in any way, from straightforward information-giving. This takes less skill, and less interpretation, than a detailed CA transcript, but could help busy solicitors to use their listening time for the most salient parts of the interview. Of course it would be necessary to test this suggestion via ecologically valid, end-to-end research, involving linguists, transcribers and lawyers, to discover whether it works well in practice. If it does, ongoing training and management would be needed to maintain appropriate standards (cf. Richardson et al., 2022).

Finally, as before, it is impossible to leave this section without mentioning the topic of interviews that involve languages other than English. Linguists are already well aware of poor practice in communication during interviews between police and less proficient speakers of English (e.g., Eades, 2018; Bowen, 2021), and are undertaking valuable research to bring improvement (e.g., Hale et al., 2019). It is certain there must also be major issues in relation to how transcripts of interpreted interviews are produced and used (cf. NAJIT, 2019). However, to my knowledge little has yet been done even to document these issues (though see Gibbons, 1995), let alone to solve them. Of course, interviews requiring use of Deaf sign language raise their own issues.

7. CONCLUSION

This systematic review started by discussing the nature of transcription, and setting out a framework for understanding the factors that affect a transcript's reliability and suitability for purpose. It then demonstrated how the framework can explain the successful use of two types of transcript that superficially appear to share few characteristics in common, namely court transcripts and transcripts used in linguistic research. This demonstration emphasised that a transcript is not the product of an individual transcriber working in isolation, but of a range of roles and factors that interact in complex ways. Ensuring the reliability and usability of a transcript requires managing all of these roles and factors effectively, with good understanding of how the transcript will ultimately be interpreted by the end-user. It is successful management at this level that ensures the success of court transcripts and linguistic transcripts for their disparate purposes.

The review then turned to two fields in which use of transcripts has been shown to be highly problematic, namely forensic audio and police interviews used as evidence in court. Emphasising that solving the problems with these transcripts requires careful identification of exactly what causes the problems, it then subjected each to analysis of the factors indicated by the framework. This showed that in neither case can the problems be addressed effectively simply by bringing the transcripts more into line with those used in linguistics research. Developing effective solutions requires considering high-level system-design factors, especially the transcript's overall purpose, and the conditions under which end-users interpret it.

This suggests a need for two strands of research, one directed towards improving provision of transcripts in a range of legal contexts, and another directed towards improving legal procedures, to ensure that good transcripts, once available, are used well. An excellent model for this kind of double-stranded research-based engagement between linguists and judges is provided by development of the Australian *Recommended National Standards for Working with Interpreters in Courts and Tribunals* (JCCD, 2022) – already used as inspiration in seeking improvement for transcripts of forensic audio (Fraser, 2020c).

It is hoped that the analysis offered in this systematic review will contribute to improving transcription in all legal contexts. A further hope, however, is that the "framework for deciding how to create and evaluate transcripts for forensic and other purposes" offered here, suitably amended via interdisciplinary discussion, might also be applied more broadly, helping to consolidate transcription as a dedicated field of study within linguistic science. After all, transcripts form the foundation of a large proportion of research in many branches of linguistics.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

ACKNOWLEDGMENTS

This paper, even more than most academic research, stands on the shoulders of previous scholars, namely those many linguists who for decades have sought to overcome societal misconceptions about the capabilities and limitations of transcripts as a representation of speech. I particularly acknowledge the work of Kate Haworth, whose publications on police interviews helped me see transcription in legal contexts as a general issue, beyond forensic audio, which could benefit from a framework like the one offered here. I also acknowledge valuable ongoing discussion with generous colleagues, especially Debbie Loakes; and careful, constructive comments from two reviewers. Remaining shortcomings of the framework, which I hope will be improved via collegial debate, are of course my own responsibility.

REFERENCES

Baldwin, J. (1985). The police and tape recorders. *Crim. Law Rev.* 695–704.

Bowen, A. (2021). Intercultural translation of vague legal language: the right to silence in the Northern Territory of Australia. *Target. Int. J. Transl. Stud.* 33, 308–340. doi: 10.1075/target.19181.bow

Bucholtz, M. (2000). The politics of transcription. *J. Pragmat.* 32, 1439–1456. doi: 10.1016/S0378-2166(99)00094-6

Bucholtz, M. (2009). Captured on tape: professional hearing and competing entextualizations in the criminal justice system. *Text Talk Interdiscip. J. Lang. Discourse Commun. Stud.* 29, 503–523. doi: 10.1515/TEXT.2009.027

Burri, M., Baker, A., and Chen, H. (2017). "I feel like having a nervous breakdown": pre-service and in-service teachers' developing beliefs and knowledge about pronunciation instruction. *J. Second Lang. Pronunc.* 3, 109–135. doi: 10.1075/jslp.3.1.05bur

Burridge, K. (2017). The dark side of mondegreens: how a simple mishearing can lead to wrongful conviction. *The Conversation.* Available online at: http://theconversation.com/the-dark-side-of-mondegreens-how-a-simple-mishearing-can-lead-to-wrongful-conviction-78466 (accessed June 26, 2022).

Capus, N., and Griebel, C. (2021). The (in-)visibility of interpreters in legal wiretapping. *Int. J. Lang. Law* 10, 73–98. doi: 10.14762/111.2021.73

Cooke, M. (2009). Anglo/Aboriginal communication in the criminal justice process: a collective responsibility. *J. Judic. Adm.* 19, 26–35

Coulthard, M., Johnson, A., and Wright, D. (2017). *An Introduction to Forensic Linguistics: Language in Evidence, 2nd Edn.* London/New York, NY: Routledge.

Coulthard, M., May, A., and Sousa-Silva, R. (eds.). (2020). *The Routledge Handbook of Forensic Linguistics*, 2nd Edn. London/New York, NY: Routledge. doi: 10.4324/9780429030581

Daniels, P., and Bright, W. (1996). *The World's Writing Systems*. Oxford: Oxford University Press.

Deamer, F., Richardson, E., Basu, N., and Haworth, K. (in press). Exploring variability in interview interpretations. *Language and Law/Linguagem e Direito.*

DeFrancis, J. (1989). *Visible Speech: The Diverse Oneness of Writing Systems*. Honolulu, HI: University of Hawaii Press.

D'Ignazio, C., and Klein, L. (2020). *Data Feminism*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/11805.001.0001

Dixon, D. (2008). *Videotaping Police Interrogation.* University of New South Wales Faculty of Law Research Series. p. 28.

Dror, I., Thompson, W., Meissner, C., Kornfield, I., Krane, D., Saks, M., et al. (2015). Context management toolbox: a linear sequential unmasking (LSU) approach for minimizing cognitive bias in forensic decision making. *J. Forensic Sci.* 60, 1111–1112. doi: 10.1111/1556-4029.12805

Eades, D. (1996). "Verbatim courtroom transcripts and discourse analysis," in *Recent Developments in Forensic Linguistics*, ed H. Kniffka. Bern: Peter Lang. p. 241–254.

Eades, D. (2010). *Sociolinguistics and the Legal Process*. Bristol: Multilingual Matters. doi: 10.21832/9781847692559

Eades, D. (2018). Communicating the right to silence to Aboriginal suspects: lessons from Western Australia v Gibson. *J. Judic. Adm.* 28, 4–21

Edwards, J. (2008). "The transcription of discourse," in *The Handbook of Discourse Analysis*, eds D. Schiffrin, D. Tannen, and H. Hamilton (Oxford: Blackwell Publishing Ltd), p. 321–348.

Eugeni, C. (2020). The reporter's invisibility. *Tiro J. Prof. Report. Trans.* 2.

Fraser, H. (2013). Covert recordings as evidence in court: the return of police 'verballing'? *The Conversation.* Available online at: https://theconversation.com/covert-recordings-as-evidence-in-court-the-return-of-police-verballing-14072 (accessed June 26, 2022).

Fraser, H. (2014). Transcription of indistinct forensic recordings: problems and solutions from the perspective of phonetic science. *Lang. Law Linguagem e Direito* 1, 5–21.

Fraser, H. (2018). Forensic transcription: How confident false beliefs about language and speech threaten the right to a fair trial in Australia. *Aust. J. Linguist.* 38, 586–606. doi: 10.1080/07268602.2018.1510760

Fraser, H. (2019). Don't believe your ears: "Enhancing" forensic audio can mislead juries in criminal trials. *The Conversation.* Available online at: https://theconversation.com/dont-believe-your-ears-enhancing-forensic-audio-can-mislead-juries-in-criminal-trials-113844 (accessed June 26, 2022).

Fraser, H. (2020a). Enhancing forensic audio: what works, what doesn't, and why. *Griffith J. Law Hum. Dign.* 8, 85–102.

Fraser, H. (2020b). "Forensic transcription: the case for transcription as a dedicated area of linguistic science," in *The Routledge Handbook of Forensic Linguistics*, eds M. Coulthard, A. May, and R. Sousa-Silva (London/New York, NY: Routledge), 416–431. doi: 10.4324/9780429030581-33

Fraser, H. (2020c). Introducing the research hub for language in forensic evidence. *Judic. Officers Bull.* 32, 117–118.

Fraser, H. (2021a). How misconceptions about transcription affect the criminal justice system. *Tiro J. Profess. Report. Transc.* 3.

Fraser, H. (2021b). The development of legal procedures for using a transcript to assist the jury in understanding indistinct covert recordings used as evidence in Australian criminal trials: a history in three key cases. *Lang. Law Linguagem e Direito* 8, 59–75. doi: 10.21747/21833745/lanlaw/8_1a4

Fraser, H. (2022). "Forensic transcription: legal and scientific perspectives," in *Speaker Individuality in Phonetics and Speech Sciences: Speech Technology and Forensic Applications*, eds C. Bernardasci, D. Dipino, D. Garassino, E. Pellegrino, S. Negrinelli, and S. Schmid (Milano: Officinaventuno), 19–32.

Fraser, H., and Kinoshita, Y. (2021). Injustice arising from the unnoticed power of priming: how lawyers and even judges can be misled by unreliable transcripts of indistinct forensic audio. *Crim. Law J.* 45, 142–152.

Fraser, H., and Loakes, D. (2020). Acoustic injustice: the experience of listening to indistinct covert recordings presented as evidence in court. *Law Text Cult.* 24, 405–429.

French, P., and Fraser, H. (2018). Why "ad hoc experts" should not provide transcripts of indistinct forensic audio, and a proposal for a better approach. *Crim. Law J.* 42, 298–302.

Fry, H. (2021). What data can't do. *The New Yorker*. Available online at: https://www.newyorker.com/magazine/2021/03/29/what-data-cant-do (accessed June 26, 2022).

Gibbons, J. (1995). "What got lost? The place of electronic recordings and interpreters in police interviews," in *Language in Evidence: Issues Confronting Aboriginal and Multicultural Australia*, ed D. Eades (Sydney: UNSW Press).

Gilbert, D., and Heydon, G. (2021). Translated transcripts from covert recordings used for evidence in court: issues of reliability. *Front. Commun.* 6, 779227. doi: 10.3389/fcomm.2021.779227

Gillon, G. (2007). *Phonological Awareness: From Research to Practice*. New York, NY: Guilford Press.

Gonzáles, R., Vásquez, V., and Mikkelson, H. (2012). "Forensic transcription and translation," in *Fundamentals of Court Interpretation: Theory, Policy and Practice,* (Durham, NC: Carolina Academic Press), 965–1042.

Grant, T. (2022). *The Idea of Progress in Forensic Authorship Analysis*. Cambridge: Cambridge University Press. doi: 10.1017/9781108974714

Green, J., Franquiz, M., and Dixon, C. (1997). The myth of the objective transcript: Transcribing as a situated act. *TESOL Quart.* 31, 172–176.

Gurevich, O., Johnson, M., and Goldberg, A. (2010). Incidental verbatim memory for language. *Lang. Cognit.* 2, 45–78. doi: 10.1515/langcog.2010.003

Hale, S., Goodman-Delahunty, J., and Martschuk, N. (2019). Interpreter performance in police interviews. *Differences between trained interpreters and untrained bilinguals. Interpret. Transl. Train.* 13, 107–131. doi: 10.1080/1750399X.2018.1541649

Haworth, K. (2013). Audience design in the police interview: the interactional and judicial consequences of audience orientation. *Lang. Soc.* 42, 45–69. doi: 10.1017/S0047404512000899

Haworth, K. (2018). Tapes, transcripts and trials. *Int. J. Evid. Proof.* 22, 428–450. doi: 10.1177/1365712718798656

Hepburn, A., and Bolden, G. B. (2012). "The conversation analytic approach to transcription," in *The Handbook of Conversation Analysis*, eds J. Sidnell and T. Stivers (Oxford: Blackwell), 57–76. doi: 10.1002/9781118325001.ch4

Heselwood, B. (2013). *Phonetic Transcription in Theory and Practice*. Edinburgh: Edinburgh University Press. doi: 10.1515/9780748691012

Himmelmann, N. (2018). "Meeting the transcription challenge," in *Reflections on Language Documentation 20 Years After Himmelmann 1998*, eds B. McDonnell, A. Berez-Kroeker, and G. Holton (Honolulu: University of Hawaii Press), 33–40.

Hoffman, D. (2019). *The Case Against Reality: Why Evolution Hid the Truth from Our Eyes*. (New York, NY/London: W. W. Norton and Company).

JCCD (Judicial Council on Cultural Diversity) (2022). *Recommended National Standards for Working with Interpreters in Courts and Tribunals*, 2nd Edn. Available online at: https://jccd.org.au/wp-content/uploads/2022/04/JCDD-Recommended-National-Standards-for-Working-with-Interpreters-in-Courts-and-Tribunals-second-edition.pdf (accessed June 26, 2022).

Jefferson, G. (2004). "Glossary of transcript symbols with an introduction," in *Conversation Analysis: Studies from the First Generation*, ed G. Lerner (Amsterdam: Benjamins), 13–31. doi: 10.1075/pbns.125.02jef

Jenks, C. (2013). Working with transcripts: an abridged review of issues in transcription. *Lang. Linguist. Compass* 7, 251–261. doi: 10.1111/lnc3.12023

Jones, T., Kalbfield, J., Hancock, R., and Clark, R. (2019). Testifying while black. *Language* 95, e1–37. doi: 10.1353/lan.2019.0042

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar Straus Giroux.

Kara, H. (2022). *Qualitative Research for Quantitative Researchers*. London: Sage.

Knoch, U., and Macqueen, S. (2020). *Assessing English for Professional Purposes*. London/New York, NY: Routledge. doi: 10.4324/9780429340383

Komter, M. (2019). *The Suspect's Statement: Talk and Text in the Criminal Process*. Cambridge: Cambridge University Press doi: 10.1017/9781107445062

Linell, P. (1988). "The impact of literacy on the conception of language: the case of linguistics," in *The Written World*, ed R. Saljo (New York, NY: Springer), p. 41–58.

Loakes, D. (2022). Does Automatic Speech Recognition (ASR) have a role in the transcription of indistinct covert recordings for forensic purposes? *Front. Commun.* 7, 803452. doi: 10.3389/fcomm.2022.803452

Love, R. (2020). *Overcoming Challenges in Corpus Construction*. London/New York, NY: Routledge. doi: 10.4324/9780429429811

Love, R., and Wright, D. (2021). Specifying challenges in transcribing covert recordings: implications for forensic transcription. *Front. Commun.* 6, 797448. doi: 10.3389/fcomm.2021.797448

Mariottini, L. (2017). "Forensic interactions: power and (il)literacy in Spanish courtroom trials," in *Forensic Communication in Theory and Practice: A Study of Discourse Analysis and Transcription*, eds F. Orletti and L. Mariottini (Newcastle upon Tyne: Cambridge Scholars Publishing), p. 151–168.

Munday, J. (2016). *Introducing Translation Studies: Theories and Applications, 4th Edn.* London/New York, NY: Routledge. doi: 10.4324/9781315691862

NAJIT (National Association of Judiciary Interpreters and Translators). (2019). General guidelines and minimum requirements for transcript translation in legal settings. *NAJIT Position Papers Position Papers on Issues Affecting Court Interpreters and Translators*. Available online at: https://najit.org/position-papers/ (accessed June 26, 2022).

NSW Young Lawyers Criminal Law Committee. (2004). *Practitioner's Guide to Criminal Law*, 3rd Edn. Available online at: https://crimlawcommittee.wordpress.com/practitioners-guide-table-of-contents/ (accessed June 26, 2022).

Ochs, E. (1979). "Transcription as theory," in *Developmental Pragmatics*, eds E. Ochs and B. Schieffelin (New York: Academic Press), p. 43–71.

Olson, D. (1994). *The World on Paper: The Conceptual and Cognitive Implications of Writing and Reading*. Cambridge: Cambridge University Press.

Ong, W. (1982). *Orality and Literacy*. London: Methuen and co.

Park, J., and Bucholtz, M. (2009). Introduction. *Public transcripts: entextualization and linguistic representation in institutional contexts. Text Talk Interdiscipl. J. Lang. Disc. Commun. Stud.* 29, 485–502. doi: 10.1515/TEXT.2009.026

Pieraccini, R. (2012). *The Voice in the Machine: Building Computers that Understand Speech. Cambridge, MA*: MIT Press. doi: 10.7551/mitpress/9072.001.0001

Richardson, E., Haworth, K., and Deamer, F. (2022). For the record: questioning transcription processes in legal contexts. *Appl. Linguist.* 1–22. doi: 10.1093/applin/amac005. [Epub ahead of print].

Ritchie, S. (2020). *Science Fictions: How Fraud, Bias, Negligence and Hype Undermine the Search for Truth*. New York, NY: Metropolitan Books.

Roberts, A., and Ormerod, D. (2021). The full picture or too much information? Evidential use of body-worn camera recordings. *Crim. Law Rev.* 8, 620–641.

Roberts, C. (1997). *Transcribing talk: issues of representation*. 31, 167–172.

Scharf, H. (1989). The court reporter. *J. Legal History* 10, 191–227.

Urban, G. (1996). "*Entextualisation, Power and Replication,*" in *Natural Histories of Discourse*, eds M. Silverstein and G. Urban (Chicago: University of Chicago Press).

Voutilainen, E. (2018). The regulation of linguistic quality in the official speech-to-text reports of the Finnish parliament. *CoMe Stud. Commun. Linguist. Cult. Med.* 2, 61–73.

Wald, B. (1995). The problem of scholarly predisposition: G. Bailey, N. Maynor, & P. Cukor-Avila, eds., The emergence of Black English: Text and commentary. *Lang. Soc.* 24, 245–257.

Walsh, M. (1995). "Tainted evidence": literacy and traditional knowledge in an Aboriginal land claim," in *Language in Evidence: Issues Confronting Aboriginal and Multicultural Australia*, ed D. Eades (Sydney: UNSW Press), p. 97–124.

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Communication

ORIGINAL RESEARCH
published: 12 August 2022
doi: 10.3389/fcomm.2022.797145

Institutional and Academic Transcripts of Police Interrogations

*Martha Komter**

Netherlands Institute for the Study of Crime and Law Enforcement, Amsterdam, Netherlands

OPEN ACCESS

Edited by:
Juan Carlos Valle-Lisboa,
Universidad de la República, Uruguay

Reviewed by:
Lorenza Mondada,
University of Basel, Switzerland
Chuanyou Yuan,
Guangdong University of Foreign Studies, China
Renata Galatolo,
University of Bologna, Italy

Correspondence:
Martha Komter
mkomter@nscr.nl

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *18 October 2021*
Accepted: *03 March 2022*
Published: *12 August 2022*

Citation:
Komter M (2022) Institutional and Academic Transcripts of Police Interrogations. Front. Commun. 7:797145. doi: 10.3389/fcomm.2022.797145

The effects of working circumstances and intended uses on the transcripts of police interrogations cannot be underestimated. In the Netherlands, police transcripts are usually drawn up in the course of the interrogation by the interrogator or, when two police officers conduct the interrogation, by the reporting officer. Contemporaneous transcription involves the interrogators in a complex configuration of interactional commitments. They have to find a way to coordinate the talk and the typing, they must transcribe the talk of an event they themselves participate in, they must do justice to the suspects' story while also taking into account the intended readership of the police report, and they must produce a document that can serve as an official piece of evidence in the criminal case. In studying recorded police interrogations and their transcripts I realised that my own transcripts are also related to their intended uses and to my working circumstances. My transcriptions are much more detailed than those of the police, which draws the attention to the differences between them. The most noticeable difference is that police transcripts focus on substance and mine on interaction. Police transcripts are meant to be evidence of the offence and mine of the talk. But there are also similarities. Both police transcripts and those of mine are selective. Police transcripts orient to their relevance for building a case, mine orient to their relevance for my research questions. Both police transcripts and those of mine treat the transcript as the talk it is meant to represent. For a criminal case this means that in court suspects are held accountable for what the police wrote down as their statement, which disregards the fact that the police transcript is a coproduction.

Keywords: conversation analysis, police interrogations, transcription, multiactivity, ethnomethodology

INTRODUCTION

A feature characteristic of institutional life is the production and use of documents, many of which contain transcripts of spoken interaction. As these transcripts are usually written by employees of the institution, and as they are meant to accommodate the needs of their institutional users, I shall call them "institutional transcripts". "Academic transcripts" are drawn up not just to document what has taken place, but also to observe, analyse and understand it. The aim of this paper is to foster awareness of the affordances and limitations of institutional and academic transcripts for those who draw them up and for their professional users. To this end, I shall analyse police transcripts of suspect interrogations, and investigate my own academic transcripts by comparison. The focus will be on how the practical circumstances of transcribing may affect the transcripts.

I take an ethnomethodological and conversation analytic perspective. Whereas ethnomethdologists have studied texts or documents in their own right, CA studies tend to approach texts or documents as integral parts of many types of talk-in-interaction, especially institutional interaction (cf. Clayman, 1990; Drew, 2006; Mondada and Svinhufvud, 2016). The ethnomethodological view of considering documents as oriented to their future uses and as affected by the practical circumstances of their construction is documented in Garfinkel's work on clinic records (Garfinkel, 1967). Garfinkel (in collaboration with Egon Bittner) drew attention to the fact that documents do not merely describe and represent an outside reality, but that they can be understood as objects in their own right and with their own dynamics. The purpose of these documents is not so much to give an objective representation of the events, but to anticipate future readership and to make available displays of justifiable work or "correct procedures" (see also: Zimmerman, 1969; Smith, 1974, 2001; Harper, 1998; Watson, 2009; Lynch, 2015).

Conversation Analysts focus the attention on the sequential organisation of talk (Sacks et al., 1974). Each turn at talk displays the speaker's understanding of the previous turn and projects the range of activities available to the next speaker (Heritage, 1984). It is not the analyst's interpretations or intuitions that count, but the interpretation of the participants themselves as shown in the sequential organisation of their talk, which can then be an important resource for the analyst. Jefferson's work on transcription for conversation analysis (e.g., Jefferson, 1983, 2004) has become the standard for conversation analytic transcription. The idea is to capture as many elements in these transcripts as is necessary for a detailed analysis. Although transcription is meant to represent the original talk in some way, it is always selective and never to be seen as the ultimate representation. It has been observed that the choices made in transcriptions are linked to the contexts of their production and reception, such as purpose, anticipated audiences, and identity of the transcriber. Transcripts thus testify to the circumstances of their creation and intended use (Bucholtz, 2000: 1440; Mondada, 2007).

Initially, Conversation Analytic studies were based on audio materials. The increasing use of video recordings opened up new areas of research, including the study of gaze, gesture, body posture, and manipulation of artifacts (e.g. Goodwin, 1996; Mondada, 2018). This led to studies of multiple simultaneous activities. The question to be answered is then how these different activities are managed and coordinated in time (Haddington et al., 2014; Mondada, 2014). Mondada (2014) has proposed a systematic ordering of multiple activities based on their temporal position in the interaction. One end of the continuum is occupied by activities that are engaged in simultaneously (the parallel order), the other by activities that remain separate and alternate mutually (the exclusive order). In between are those activities that are coordinated and intertwined with one another (the embedded order). Most often multiple activities are managed by switching from one type of organisation to another.

My research into the ways in which police officers interrogate suspects and report their talk is focused on the organisation of talking and typing, and on the effects of practical circumstances on the talk, the typing and the texts of the transcripts (Komter, 2019). Studying the transcripts of police officers made me think about those of my own, so I decided to investigate the possible effects of my own working conditions and purposes on my transcriptions.

My materials include 34 audio recordings of police interrogations of "ordinary" street crimes, the police reports[1] of these interrogations, and my transcripts of the interrogations.[2] Because police officers are aware of the risks of their job, risks that may involve putting unacceptable pressure on suspects to confess, most of the police interrogations that we were allowed to record concern common street crimes such as drug dealing, robbery, or theft.

It is not my intention to present my materials and my practices as characteristic of institutional and academic transcription, but rather as examples of specific instances of transcripts of Dutch police interrogations. The fragments presented here are chosen not only to reflect the various conditions under which police officers perform their dual tasks of interrogating and reporting, but also to demonstrate and account for the choices I made for my transcriptions. In the following sections I shall first discuss some of the interactional arrangements in police interrogations for combining talking and typing, after which I examine the practical circumstances of my own transcriptions and the bases of the choices I made in transcribing these interrogations.

PRACTICAL CIRCUMSTANCES OF POLICE REPORTING

A characteristic feature of Dutch interrogations is the practice of contemporaneous transcription, which means that police officers must find a way of coordinating talking and typing. The organisation of talking and typing varies with the number of interrogators. "solo" interrogations are conducted by a single interrogator, who has to combine and coordinate talking and typing. In solo interrogations the typing alternates with the talk as question-answer-typing sequences (Komter, 2002–2003, 2006; Van Charldorp, 2011).

In "duo" interrogations the interactional organisation of the event is different: it affords opportunities for a division of labour between the two police officers and it provides for different forms of speakership and recipiency. The interactional organisation of the talk is more complex than in the "solo" interrogations as there is also room for interaction between the two police officers and between the reporting officer and the suspect. The usual division of labour in "duo" interrogations is that one of the police officers does the typing and the other does (most of) the questioning. This results in a simultaneous production of the talk and the typing, and for an orientation to interrogating and statement taking as appropriate simultaneous activities. In other words, in solo interrogations the activities are organised serially, and in duo interrogations concurrently (see: Haddington et al., 2014).

[1] Police reports are documents that contain the necessary administrative items and the police transcript of the interrogation (see: Komter, 2019). They are used in court as official pieces of evidence.

[2] Of these interrogations and police reports, 20 were collected by me and 14 by Tessa van Charldorp.

Solo Interrogations, Monologue Style

When asked, police officers consider contemporaneous transcription in solo interrogations a necessary evil, as it detracts attention from what they consider to be the core business of the event: interrogating the suspect (Malsch et al., 2012). This is corroborated by my findings that show how investigative questioning may be incompatible with contemporaneous transcribing, especially during antagonistic episodes in interrogations conducted by a single interrogator (Komter, 2002–2003, 2003, 2019).

Police manuals and instructions urge interrogators to start the interrogations with open questions about what happened. The idea is that open questions stimulate suspects to feel at ease and to tell their own version of the events. This enables the interrogator to report the suspect's "own words", which makes it more difficult for the suspect to withdraw his statement afterwards. Moreover, the length of the answers to open questions will provide the interrogator with enough material to ask new questions (Van den Adel, 1997).

However, the advice to start the interrogation with an open question does not take into account that open questions generate undirected answers, which may not contain the information required for a legally adequate piece of evidence. Another constraint on the management of open questions is, that the answers may be too long to remember and to write down in one go. In a number of interrogations in my materials police officers start with an open question about what happened without writing anything down, after which they recycle the story and report it bit by bit.

The next fragments are from an interrogation for a case of theft. The suspect initially denies her involvement in the events but eventually she confesses (see Komter, 2003). After the exchanges about the suspect's personal details and her living circumstances (the "social interrogation"), the interrogator (P) begins the interrogation proper by asking the suspect (S) to tell him what happened. He then recapitulates what she told him. According to the suspect, the events took place at "the market" (the text written down in the police report is transcribed in bold, underneath the lines that indicate P's typing):[3]

(1)
1. P: So yesterday you went to the market with
 your children.
2. S: Yes.
3. P: ((types, 6 s:))
Yesterday,
4. P: To the market, then we're talking about
 Waterlooplein I assume.
5. S: What do you say, yes.
6. P: Yes,
7. ((types, 17 s:))
I went to Waterlooplein, **together with my children**.
8. P: Uh (4) have you uh been to the stalls?[4]

We see here that every now and again the interrogation comes to a halt while the interrogator is typing. At the same time, the question-answer-typing (Q-A-T) format makes the typing an integral component of the interaction. It is noticeable that P stops his typing (line 4) in order to specify the location of the events as "Waterlooplein" instead of "the market." This is important information for the prosecutor, who has to indicate the time and place of the offence in the indictment. The monologue style of the report transforms the interaction into a seemingly volunteered narrative by the suspect.

P's recapitulation (line 1) works to round off the suspect's "free story" and to embark on the reporting of it. It is a formulation used to demonstrate understanding of the suspect's prior talk (Heritage and Watson, 1979). As it projects confirmation, it serves as a "candidate recordable" that elicits not only the suspect's agreement with the formulation but also with the text to be written next. P's typing (lines 3 and 7) transforms the interactional organisation of the talk into a question-answer-typing (Q-A-T) format. The Q-A-T format is found especially in the uncomplicated, routine episodes of the interrogations. It consists minimally of one question-answer exchange, but more often there is a series of questions and answers preceding the typing (Komter, 2006). During the typing, the suspect usually waits for the interrogator to ask the next question.

P's typing activities have "turn-like" features, as they start at transition relevance places in the suspect's talk, and they occupy the floor. Moreover, they can be understood as third position actions, serving as a sign of acceptance and understanding of the suspect's prior answer. The difference with conversational turn-taking is that the setting is "partially opaque" (Goodwin, 2000: 1508), in the sense that the suspect does not know what the interrogator is writing, nor how long the typing will last. Thus, as long as the typing occupies the floor, there is no transition relevance place for suspects to take the next turn. As interrogators generally take the turn after the typing, the Q-A-T format reinforces the interrogator's position of initiative and control.

As the tension in the interrogation increases, the interrogator suspends the typing for a while and directs his attention exclusively to the suspect instead of to the screen of the PC. The next fragment is part of the police transcript (the numbering is added by me):

[3] **Transcription conventions**

P	police interrogator
P^1	interrogating police officer
P^2	reporting police officer
S	suspect
full stop.	falling intonation
comma,	slightly rising intonation
question mark?	rising intonation
underlining	emphasis
(3)	pause of three seconds etc.
...	a few words omitted
=	latched utterances
( )	unclear utterance
(possible hearing)	possible hearing
((double brackets))	transcriber's note
shading	typing simultaneous to the talk.

[4] For the original Dutch examples see the **Appendix**.

(2)
1. **Then I bit the lady of the market stall where I bought the brooch in her wrist.**
2. **I did not bite hard. I bit her because she was pulling at me.**
3. **I have not told the whole truth, but I shall tell you the truth now.**
4. **I said that Clive took away the display-case from the market stall.**
5. **That is not so, for I actually took away the display-case myself from the market stall.**

Denying suspects do not usually change their position without inducement from the interrogators. The text gives no information about what actions the interrogator actually took, nor how much effort it took to persuade the suspect to confess. Indeed, a comparison with the talk in the interrogation shows that between lines 2 ("I bit her because she was pulling at me") and 3 ("I have not told the whole truth") there is half an hour of interaction that is not written down, in which the interrogator gradually steers the suspect toward her admitting not having told the truth. At this point the interrogator takes a break, in which the suspect goes to the toilet, after which the interrogator gives her a glass of water. He continues:

(3)
1. P: Right.
2. ((types, 8 s))
I haven't told
3. I now put I haven't told the whole truth, but I shall tell you the truth now.
4. Okay?
5. S: ((whispers:)) Okay.
6. ((types, 21 s))
The whole truth, but I shall tell you the truth now.

P's resumption of typing indicates that a different type of activity is relevant now beside interrogating her: from now on, he will be taking down her statement again. The whole episode of steering S toward a confession is retrospectively treated as "off the record". The talk in the interrogation will be talk-for-the-record again, the story that the suspect will tell will be the truth, and the truth will be recordable as piece of evidence.

The shift between the two activities of interrogating and typing is achieved explicitly; P does not only tell S that he types, but also what he types (line 3). Moreover, he asks for S's permission and agreement with the text to be written. In doing this, he constructs this moment as point of no return. With her support he writes down that she will tell the truth now, which involves her changing her story in such a way that a confession becomes relevant. P's articulation of what he is about to report suggests that it is now too late for her to go back on her promise, as the text written down constrains S's options.

The text of the next fragment is from an interrogation in a case of drug dealing. In a street in Amsterdam notorious for drug dealing activities, the police had been watching the suspect's actions for a while. In the course of his third drug deal he was arrested. P recapitulates S's description of his arrest (the written text is presented in the right hand column):

(4)

1.	P:	So you were arrested with that last person	**During the sales transaction with the latter person I was arrested**
2.	S:	Yes that lasted only half a minute.	**by two plain clothes police officers together with the buyer.**
3.		They had just been watching right.	

P recapitulates the suspect's prior talk with a formulation (line 1; Heritage and Watson, 1979) that projects a confirmation, the answer to which allows him to report that there has been a "proper arrest", because the suspect has been caught in the act. Moreover, he adds a lot of information in the police report that is not talked about in the interrogation. This can be attributed to the two "directions" of the police report: it is meant to look backward as representation of the talk in the interrogation, and forward in anticipation of the needs of future readers of the report. His additions and the stilted style in which the suspect seems to express himself suggest that the interrogator is orientated more to the prospective readers than to the suspect's original talk.

Let us now consider the interactional organisation of talking and typing. As S goes on talking after P has started typing, I shall transcribe the simultaneity of the talk and the typing, and suggest what text is typed when. The concurrent talk is transcribed by gray shading, to exhibit the simultaneousness of the talk and the typing.[5]

(5)

1.	P:	So you were arrested with that last person	
2.	S:	Yes that lasted only half a minute	
3.		they had just been watching right.	
4.	P:	((types, 7 s))	**During the sales transaction**
5.	S:	Yes what do I know,	**with that latter**
6.		I mean if I'd do that every day,	**person**
7.		then you could say I'd be dealing but uh	**I was arrested**
8.		if I'd do that every day yes.	**by two**
9.		Then I'd also say dealing	**plain clothes**
10.		but uh that's not the case.	**police officers**
11.	P:	((types another 5 s))	**together with the buyer.**
12.	P:	Look, the Criminal Code …	
13.		does not make that distinction	

[5] This is an approximation, as it is impossible to ascertain the exact placement of the text.

As in fragment 1, the interrogator recapitulates prior talk and listens to the suspect's confirmation before starting to write (lines 1–3). Although P allows the suspect to finish his utterance, the text of his subsequent typing shows that he only pays attention to S's confirmation ("Yes", line 2). This then provides him with the opportunity to reformulate and elaborate his summary, as his entry into the police report shows.

The episode starts off as a Q-A-T sequence. However, in this instance S does not wait for P to ask a next question, but picks up his talk 7 seconds after the start of the typing. The suspect not only takes the story further than the question asked for, but his elaborations also portray his doings as "normal" activities in everyday life. His additions resemble the "narrative expansions" identified by Galatolo and Drew (2006), that are produced to defend a person against a possible allocation of blame implied in the question. The absence of a slot for S's defensive elaborations, and the apparent urgency of his defensiveness, prompt his early response. When he is done, P completes his typing after 5 seconds (line 11). His next turn exhibits that he has heard the suspect's contributions (lines 12–13), but he does not write them down.

My materials show that interrogators tend to continue with their typing when suspects talk simultaneously, and that what suspects say simultaneously tends not to be written down. At the end of the interrogation the suspect reads the transcript, is asked if he agrees with it and signs it. In my materials, suspects never complain of items that have not been written down.

One of the arguments police officers gave for their dislike of contemporaneous reporting was that it interferes with the flow of the conversation (Malsch et al., 2012). On the other hand, police officers have no problems with picking up the thread of prior talk, because they have only to look at the screen to see where they have left off. In the next fragment from a case of shoplifting the last sentence on the screen reads: **On the ground floor I took a T-shirt worth Fl. 15,- from a rack and put it under my coat**. P continues:

(6)
1. P: Well you put that shirt under your coat and you left the shop without paying.
2. S: Yes.
3. P: And were you stopped outside or or uh
4. S: Yes.
5. P: in in the doorway or after the gates where exactly was that?
6. S: Outside.
7. P: In the street.
8. S: Yes.
9. P: ((types, 20 s))
10. **Then I walked out of the store without paying. Outside I was stopped.**

P reads from the screen in front of him what he has typed last, transforms it into a sentence addressed to the suspect ("you put that shirt under your coat", line 1) and proposes a "reasonable" future recordable ("you left the shop without paying" line 1). The suspect's response (line 2) is both a confirmation and permission for it to be written down. Thus, the transcript-thus-far is used as a resource to carry on the interrogation where it was left off, and as a means to take the suspect's story further.

In sum, solo interrogations are organised as a series of Q-A-T sequences, especially in the unproblematic parts of the interrogations. The Q-A-T sequence is accomplished by a piecemeal elicitation of "chunks" of information and by writing them down step by step. The typing is accompanied by a temporary shift from a mutual focus on the interaction to divergence, where the attention of the interrogator is directed toward the screen of his PC.

A constraint on the typing is the problem of reporting answers to open questions. In the more problematic episodes there may be a suspension of the typing, signifying that the unreported talk is "off the record" for the time being. This testifies to a potential incompatibility of talking and typing, as the interrogators" attention to the screen would reduce the intensity of their questioning. In addition, police interrogators may be reluctant to put their more adversarial actions on display.

Duo Interrogations, Question-Answer Style

The usual division of tasks in duo interrogations is that one police officer asks the questions and the other writes down the talk. There are various ways in which the interrogating officers encourage and inform the reporting officers' writing tasks. For example, interrogators sometimes explicitly instruct reporting officers on what to write, and in some cases they slow down their talk and articulate it as if dictating a text to the reporting officer. At a more implicit level, the interrogating officers may show an awareness of the reporting officers' tasks at hand by leaving pauses for the typing or by producing utterances that could facilitate the reporting, for example repeats of the suspects' answers (Komter, 2019). This shows that the division of labour is not just an instance of a participation format that consists of separate activities, but that it provides for the collaborative constitution of a shared stance (Goodwin, 1996: 375).

The next fragment, in the question-answer style,[6] is an example:

[6] My police materials contain transcripts in three writing styles: monologue style, question-answer style, and recontextualised monologue or "you ask me" style (see Komter, 2019).

(7)

1.	P^1:	I just want to talk about those fake	
2.		drugs right? (2)	
3.		How did you come by them. (1)	
4.	S:	I made them myself	
5.	P^1:	You made them yourself.=	**Question:**
6.	S:	=Yes. With what?	**how did**
7.	P^1	Yes that is the next question. Okay.	**you come**
8.	S:	With wheat flour and salt,	**by those**
9.	P^1:	Wheat flour and salt.	**fake drugs?**
10.		(3)	**Answer:**
11.		And where did you make that.	**I made them**
12.	S:	At home.	**myself**
13.	P^1:	Where is home.	**with wheat**
14.	S:	In the kitchen.	**flour and salt,**
15.	P^1:	No, what do you mean with home. (2)	**at home.**
16.	S:	At my uncle's house.=	**With that**
17.	P^1:	=At your uncle's house.	**I mean**
18.		(8)	**in the kitchen of my**
19.		Was your uncle at home too,	**uncle's**
20.		when you did that.	**house.**

As the typing is almost continuous, it is difficult to ascertain exactly when what text is typed. By the time P^1 asks his question (line 3) the reporting officer (P^2) is still typing up the prior talk. It can be suggested that at the same time she orients to the talk, as she suspends her typing during the suspect‘s answer (line 4). The first potential moment for her to report the question-answer exchange is after that, but it is possible that she is still finishing typing up prior talk. In either case, it will be clear that the typing lags behind, and that the pauses left by the interrogator are not long enough for her to keep up with the talk.

One of the ways in which interrogators take the work of the reporting officers into account is to repeat the suspect's answer, followed by a pause. There are three repeats in this fragment (lines 5, 9 and 17). The first repeat is followed immediately by the suspect's confirmation and by his production of what would be the next logical question (line 6). The suspect's answer to this question is then followed by the second repeat (line 9). This time the interrogating officer (P^1) is in the position to leave a pause after the repeat (line 10), as the suspect waits for the next question. P^1's next question (line 11) does not at first receive what he considers to be a complete answer, as is evidenced by his further questioning about the meaning of the suspect's answer "at home" (line 12). The suspect's final answer "at my uncle's house" (line 16) is then accepted by P^1 with a repeat followed by an eight second pause (lines 17–18).

The combination of repeats and pauses attends both to the "recordability" and to the "typability" of the talk. P^1s choice of repeats suggests the recordability of the substance of the text to be written down and his leaving pauses promotes the "typability" of this text (cf. Moerman, 1988: 54). The inclusion of the pauses is not P^1s decision alone, as S takes the opportunity to respond to P^1s first repeat (lines 5-6). Although the pauses facilitate the typing, they are much shorter than the "typing turns" in the solo interrogations. P^1 apparently relies on P^2's capacity to listen and type at the same time.

It can be noted that four questions are asked (lines 3, 7, 11 and 15), whereas only one question is reported (lines 5–9, right hand column). Thus, the suspect is reported to answer "more than the question". This is common practice in question-answer police reports, as it is a way for the reporting officer to deal with the constraints of time. In the question-answer style transcripts in my materials about one in six of the questions asked are written down (De Boer, 2014), resulting in a "monologisation" of the Q-A style reports (Komter, 2019).

The next excerpt shows problems not only with the intelligibility of the suspect's talk, but also with the teamwork. The suspect is a man from Sudan, who speaks a kind of Dutch that is difficult to understand. The interrogators suspect that he is an illegal immigrant and that he has been staying in The Netherlands for some time. The suspect does not want to answer the repeated questions by the interrogator about how he travelled to the Netherlands. The case related interrogation begins with P^1 recapitulating the conditions of the suspect's arrest, followed by a question about the duration of his stay thus far. In the meantime the reporting officer is still writing down the suspect's prior answer:

(8)

1.	P^1:	Okay we have picked you up there at Park.	**I don't want**
2.		Park 345. How long have you been there.	**to answer**
3.	S:	Uh I uh I come there yesterday because	**that any**
4.		I have a w- that woman who lives there,	**more.**
5.	P^1:	yes,	**Question:**
6.	S:	and uh her son is a good friend of mine.	**You were**
7.		and so then it is also often ( )	**arrested**
8.		and I have with him ( ) telephone	**in a house**
9.		or something ( )	**at Park.**
10.		and I told him like I have him because	**How**
11.		I come from uh I come yesterday to the	**long have**
12.		Netherlands with my (family) then	**you lived**
13.	S:	then with my (partner), come ( )	**there?**

14.	P^1:	Yesterday you come where?	
15.	S:	hm?	
16.	P^1:	What do you say?	
17.	S:	I come yesterday here to Park.	
18.	P^1:	Yes,	
19.	S:	Yes. Because I must uh in Amsterdam,	**Answer:**
20.		come and get a few things of mine with my	**The woman**
21.		(family) and then I ask that boy that have I	**who lives there,**
22.		something place to sleep? And so he has me	
23.		uh (and my mother too) perhaps you can	
24.		sleep for me	
25.	P^1:	((to P^2)) Can you still follow it? (2)	
26.	P^2:	No. (4)	
27.	P^1:	So you sleep there since yesterday,	
28.	S:	Yes.	
29.	P^1:	And you asked a friend?	
30.	S:	Yes a friend of mine is uh a son	**her son was**
31.		of that woman.	**a good friend**
32.	P^1:	And what is his name.	**of mine.**

Let us first examine the talk. The suspect answers the interrogator's question about the duration of his stay immediately ("I come there yesterday", line 3), but then he continues by giving what seems to become an account ("because … that woman who lives there", lines 3-4). P^1 encourages him to proceed with a continuer "yes" (line 5), after which the suspect goes on with a long uninterrupted turn (lines 6-13). His account is rambling and difficult to understand, but P^1 gives him the scope to expand and does not ask for clarification until line 14. One phrase that can be understood is S's virtual repetition of "I come there yesterday" (lines 3 and 11). This is then taken up by P^1 for further detailing (line 14). After the suspect provides the answer ("I come yesterday here to Park", line 17) P^1 utters another "yes" continuer (line 18), which is followed by what appears to be the suspect's motivation for coming to the address where he was arrested.

At the end of this P^1 turns toward P^2 to ask if she can still follow it (lines 25–26). Here the participation status of the participants changes: the interrogator draws the reporting officer into the interaction, while the suspect is temporarily excluded. The shortness of P^{2}s answer displays an orientation to minimal intrusiveness and characterise the exchange as a form of "byplay", which does not terminate their prior alignment but holds it in abeyance to be reengaged at a next moment (Goffman, 1981: 155). The 4 second pause marks an interactional "no man's land" after which the original participation format is reinstated. P^1 "recycles" the suspect's narrative by repeating some items in combination with some further questioning (lines 27–32).

There are three periods of typing in this episode (lines 1–13, 19–21 and 30–32). P^2 stops the typing for a short while when P^1 asks questions for clarification and S answers (lines 14–18). When the typing is halted a second time this may have been a sign for P^1 to ask P^2: "can you still follow it" (line 25). On P^2's negative answer (line 26), P^1's subsequent recycling of the suspect's original answer (line 27) may be produced as a "candidate recordable" to enable P^2's reporting of it (see the formulations in fragments 1 and 4).

However, if we look at the text that is typed up contemporaneously by P^2, we see that she wrote down the question (lines 5–13, right hand column) in the course of the talk between P^1 and the suspect, but missed the answer ("I come there yesterday", line 3). Instead, in spite of P^1's question for clarification and the suspect's answer about the day of his arrival (lines 14 and 17), and in spite of P^1's reformulation (line 27) she wrote down the account about "the woman who lives there" and her son (right hand column, lines 19–21 and 30–32). This text corresponds with the suspect's talk directly following his answer (lines 4 and 6).

These troubles may be attributed to the fact that the suspect's talk is rather unintelligible and that P^1 allows him some scope for continuing his narrative. P^1 appears to listen to the suspect's account as a "free" story, and to give him the opportunity to present his version of the events without interference. As mentioned above, this occurs quite often in solo interrogations, after which the interrogator recycles the suspect's story as a Q-A-T format to accommodate the typing. In duo interrogations, as the example shows, the "free story" may be incompatible with a parallel organisation of talking and typing.

Because of the differential pace of talking and typing, P^2 writes down a selection or summary of the talk. P^2 has selected the item of "the woman who lives there... her son is a good friend of me" (lines 4 and 6) for inclusion in the report. It can be expected that her problems are a result of the circumstance that from the moment that she misses the suspect's answer (line 3: "I come there yesterday") and writes down the next item (lines 4 and 6: "the woman who lives there...") she listens for possible continuations of the text on the screen. These troubles are likely to result from diverging orientations: P^1 listens for the story or for elements in the story to be taken up later, while P^2 listens for the typing and for the text: she has to combine the writing down of previous talk with listening for what to write next, while taking into account the text already on the screen. It is the kind of "practical listening" that exhibits their different tasks at hand.

Summary

In "duo" interrogations the typing has a less prominent position than in solo interrogations: it does not occupy the floor and the moments of typing onset or typing completion have less sequential relevance for the talk. Duo interrogations show various degrees of "teamwork". Interrogators may facilitate the writing by repeating an answer and by leaving pauses for the typing, or they may follow their own plan and leave it up to the reporting officer to decide what to write. Reporting officers are dependent on the interrogating officers for allowing them the time to write, and interrogating officers depend on the reporting officers' skill in keeping up with the talk and selecting the relevant items for the report.

However, when interrogating officers leave pauses for the reporting officer after a "recordable" answer of the suspect, these pauses are usually not sufficient to complete the reporting of prior talk. This makes for a more complex writing task than in the 'solo" interrogations as, beside the problems resulting from the constraints of time, the reporting officers have to remember and write up past talk, see to it that the text of the current writing is in line with the text already on the screen, and at the same time listen to current talk for future "recordables". The simultaneity and the differential pace of talking and typing affect the typing more than in the "solo" interrogations; it may result in mishearings, in a more selective reporting, and in a "monologisation" of Q-A style reports.

CONCLUSION

I have presented the fragments of police transcripts as examples of the coordination of the talk and the typing in solo and duo interrogations, and of the ramifications of contemporaneous talking and typing. In a broader sense, these fragments can be seen as instances of the impact of practical circumstances and purposes on the actions of the interrogators and on the texts of their transcripts.

Contemporaneous transcription inevitably leads to selective transcripts. The fragments shown here show two writing styles: the monologue and the question-answer style. The monologue style reads as a statement volunteered by the suspect, the question-answer style includes the interrogators" activities. Although the question-answer style police transcripts are more transparent, this is deceptive as most of the questions asked are not reported. Whatever the writing style, the police transcript is always a summary of the talk that focuses on substance rather than on interaction.

The practice of contemporaneous transcription of police interrogations entails a coordination of talking and typing. In solo interrogations this is predominantly accomplished exclusively, where the two activities alternate as question-answer-typing sequences. The separation of talking and typing is achieved by the suspects' waiting for the interrogator to finish the typing, and by the interrogators' disregard of the suspect's contributions during the typing. However, this organisation can develop into a more parallel organisation when suspects choose to add elaborations to their answer during the typing.

In duo interrogations there is usually a division of tasks, which allows the talking and the typing to be produced simultaneously. The temporal organisation of the two activities is more precarious than in the solo interrogations. The interrogator takes into account the tasks of the reporting officer, by repeating the reportable items and by leaving pauses for the writing. But a repeat does not necessarily result in the reporting of the required answer, and the pauses are usually too short for the reporting officer to keep up with the talk. This may result in a suspension of the typing or in a misrepresentation of the talk.

Thus, the talk, the typing and the text are inextricably interwoven. The talk is not merely a search for the truth about what happened but it is also directed at eliciting recordable answers that may contribute to building a case. The typing is not merely an activity for reporting what has been said but it is also part of the interaction between the interrogator and the suspect or, tacitly or explicitly, between the interrogator and the reporting officer. And, especially in the solo interrogations, the police transcript is not merely a document in which what is said is laid down, but it actively informs and directs the interrogation.

PRACTICAL CIRCUMSTANCES OF ACADEMIC TRANSCRIPTION

One of the practical circumstances that researchers have to deal with is the nature and quality of the recordings. The first series of 20 interrogations was collected around the turn of the century, when interrogations for "ordinary" street crime were usually conducted by one interrogator, and when the usual format was the monologue style. After a series of miscarriages of justice in the first decade of the century, it became more common to conduct the interrogations with two police officers, in the question-answer style. So the second collection of 14 interrogations differs in reporting style and number of interrogators.

Police interrogations are difficult to come by. During the entry negotiations I had to appease the worries of the officials I approached who were afraid that the recording process might interfere with the management of the interrogations. As the recording equipment we used was small, and as I thought it would be adequate for our purposes, I opted for audio recordings. If I had known the importance of the typing for the organisation of police interrogations beforehand, and if I had known that I wanted to include the texts of the police reports in my transcriptions, I would have tried to install some kind of a text tracking device through a connection between the audio-recorder and the computer that would enable me to trace exactly what was typed when. And to be able to analyse the embodied manifestations of the interrogators' dual attention to the screen and the suspect, I would have preferred video instead of audio recordings.

It is a feature of academic transcription that the research questions develop in the course of getting familiarised with the materials through transcribing them. This entails a constant movement between research questions and transcription (Mondada, 2007: 810). In the course of this process, I had to make decisions about whether to insert the text of the police report in the transcripts, how to transcribe the talk and the typing, the coordination of talking and typing, the amount of detail, and the translation. And I had to reconsider these choices whenever I thought there were better ones.

The Talk and the Text

In the early stages of my work I decided to insert the text of the police reports into my transcripts, in order to show comparisons of the talk with the text. This was easy for the Q-A-T sequences, as I transcribed the typing as transcriber's note, for example:

((types, 20 seconds)) and underneath that the corresponding text in bold (see fragments 1, 3 and 6). I got into trouble when the typing and the talk co-occurred. I solved that by constructing two columns, with the interaction in the left hand column and the corresponding text of the police report in the right hand column (see fragment 4).

However, this did not give any insight into the moments in the interrogation in which the texts were typed by the police officer. So I reconstructed what was typed up when. This was more or less easy in the interrogations with one interrogator (see fragment 5), but more difficult in the interrogations with two interrogators where talking and typing co-occur (fragments 7 and 8). The reconstruction of the moments when the texts were typed is based on an inspection of the following:

1. the correspondence between the talk and the text;
2. the differential pace between the talk and the typing;
3. the text follows the talk;
4. the length of the talk and the approximate length of the typing;
5. break off of typing may signify a completion of the text thus far.

This is the most problematic feature of my transcription, but the nature of my recordings makes it impossible to be more exact. For the problematic episodes I have "try-typed" the Dutch text and compared its duration with the duration of the talk in the audio recording of the episode. These ways of reconstructing what was typed when shows that in these circumstances the text lags considerably behind the talk.

Talking and Typing

As I became familiarised with the materials, I soon realised that the typing was more than just a pause in the talk or a background noise, because what I first transcribed as pauses were much noisier and longer than conversational pauses, and they clearly embodied specific activities of the police officers. Moreover, I had to find a solution to the problem of transcribing the co-occurrence of talking and typing.

There is no standard way of transcribing keystrokes. Zimmerman (1992) uses dashes to indicate keyboard activity. Whalen's transcription (Whalen, 1995) uses different symbols to indicate keystrokes, space bar, tab, back-tab, return, cursor, and arrow keys. Van Charldorp distinguishes between louder (X) and softer (x) keystrokes (Van Charldorp, 2011). Greatbatch et al. (1995) use symbols that differentiate between keystrokes, keystrokes that are pressed with greater force than normal, and return keystrokes.

In those cases where there was co-occurrence of talking and typing I used ### symbols to indicate the typing, and the overlap symbol [to indicate at what moments the talk and the typing co-occurred (see Komter, 2006). The problem with this notation is that it creates the impression that an audio recording allows the transcriber to hear and transcribe every single keystroke separately. I therefore decided to transcribe the talk and the typing not as two different lines but as one, the typing marked by a shade of gray covering the simultaneous talk, as a more direct way to accentuate the simultaneity of the two different activities (fragments 5, 7 and 8; Komter, 2019). I realise that this involves a loss of detail regarding variations in keystroke activity.

Amount of Detail

The basic principle of Conversation Analytic transcription is "to get as much of the actual sound as possible into our transcripts, while still making them accessible to linguistically unsophisticated readers" (Sacks et al., 1974: 734). While my transcripts give a more complete and more detailed account of the talk than do the police transcripts, they are less detailed than the usual Jeffersonian transcript notations.

Decisions about the degree of detail in academic transcripts depend on their relevance for the research questions and analytic perspective (Hepburn and Bolden, 2012: 73–74). And, conversely, the research questions may be adapted on the basis of what emerges in the process of transcription. As research questions may change in the course of getting familiarised with the data, it would be sensible to start out with a detailed Jeffersonian transcription, to diminish the chance that you are missing something essential (Jefferson, 1983).

On the other hand, there are practical considerations. The bulk of my materials led me initially to make more global transcriptions, until I could decide what phenomena were worth studying in more depth. As my research questions became more definitive, I adapted the transcription accordingly. Moreover, as I aimed for a broader audience, I was faced with the choice between detail and readability. I made use of Jeffersonian transcription notations as much as I thought I needed and added some of my own when I thought I needed them.

Translation

As I published most of my analyses in English I had to translate the original Dutch transcripts. Translations can never represent the phonetic details of the original talk, so only those features of the standard transcript notation have been preserved that are compatible with the translation: intonation, stress, pauses and overlap. Thus, it is inevitable that translation increases the distance between the transcript and the original talk. The challenge is to capture in the translation the salient details of the original language.

The usual way to present translated transcriptions to an English speaking readership is a three-line transcription, where the first line is the original transcript, the second line a word-by-word translation into English, and the third line an idiomatic translation that is meant to capture the conversational style of the talk (cf. Hepburn and Bolden, 2012: 68–69). For my purpose this turned out to be impractical, as I decided to transcribe the talk, the typing and the text in one and the same excerpt. So I chose the solution of presenting the original Dutch examples in an appendix (see **Appendix**). Another argument against the three-line transcription is that, when the publication has restrictions on the size of the article, the inclusion of the transcript in the original language leaves less space for analysis and discussion (Slembrouck, 2007).

DISCUSSION

The transformation of talk into writing allows for the transportation of the resulting written texts to readers who may use these texts in the performance of their professional tasks. As this is a crucial element of professional practices, this holds for the professionals in the criminal law process but also for academics who study and transcribe the talk. In fact, the transformation of talk into writing is one of the basic tools of conversation analysis, as transcription is the instrument for making talk in interaction available for inspection, reproduction and publication. Although the transcriptions made in conversation analytic studies are obviously constructed to be more accurate and complete representations of the talk-in-interaction than police transcripts, the principle is the same: talk is transformed into written materials that are easier to manage than live talk because they are fixed and transportable, so that they can be made accessible to a particular readership to serve specific ends. Below I shall discuss differences and similarities between police transcripts and my academic transcripts related to participation, purpose, relevance and selectivity, and to the status and treatment of the transcripts.

Participation

Police officers are participants, who are transcribing the talk in the interrogation that they are conducting. They are active speakers and hearers, monitoring the suspects' ongoing talk for inserting their own contributions and responses. They listen for understanding and for responding, but also for the recordability of the suspects' answers. On top of this, they must make a transcript, while being involved in the moment-by-moment contingencies of the configurations of their interactional commitments. A feature of contemporaneous transcription is that the completion of the interrogation coincides with the completion of the police transcript. When the participants have signed the police report, both the interrogation and the report are brought to an end.

Whereas the police transcripts were completed and ready to be sent off to the desks of those who would deal with the case, mine had yet to begin. I collected the recordings and the police reports and, in the relative peace and quiet of my office I could begin to play and replay the recordings, not only to understand what the participants were saying, but also to inspect more closely the phenomena that I discovered in the materials as I progressed with the transcription. This is a solo-activity as it is not embedded in interaction with others. It is similar to the activities in solo interrogations as it involves a continual shift of attention between the recording equipment and the screen of the PC, between listening and writing. Because the transcription relies on recordings instead of on participation, academic transcripts are only completed when they appear in print. But even then, they remain open for discussion and revision (cf. Bucholtz, 2007).

The transcription of audio or video recordings of talk and action involves a change of perspective, as the unique and ephemeral moments of the event are reproduced as moments in the recording, which can be played and replayed by observers who were not necessarily present at the time and did not take part in the interrogations. Thus, the sources of transcripts, live or recorded interaction, affect participation and perspective, and are therefore sources of differences between the texts of the transcripts.

Purpose

Another source of differences is the orientation to the intended uses of the transcripts. Police transcripts are meant to serve the legal professionals who will deal with the case in later stages of the criminal process as basis for their decision-making. Police transcripts are oriented to what the suspects have told the interrogators about "the facts", rather than to the interactional contexts of the creation of the transcripts. They are summaries of the interrogations, not only because of the circumstances of contemporaneous transcribing but also because judges are satisfied with a police report that contains a "factual representation" of what the suspect told the police (Franken, 2010: 406), rather than being burdened with a verbatim transcript.

My aim is, among other things, to observe and analyse the work of police officers in the interrogation room for academic publication. The aim of my transcripts is to gain insight into the processes by which police transcripts are produced. A result of the differences in purpose and use of the transcripts is that my transcripts are much more detailed and cover the whole of the interaction in the interrogations. Moreover, my transcripts changed in the course of my research as my understanding of the role of the typing, of the coordination of talking and typing and of the impact of the written texts on the interaction increased.

Thus, the purpose of police transcripts is to create a document that can serve as evidence in a criminal case; academic transcripts can also be considered as evidence, but they are evidence of the talk, not of the offence. Although the purposes of the two types of transcript differ, they are similar in that they are both meant to be a representation of the talk, and they are both "recipient designed", as they take into account their future readership.

Relevance and Selectivity

When asked, police officers say that they do not aspire to transcribe the whole interrogation, but that they only write down what is relevant for the case (Malsch et al., 2012). One may wonder what they mean by "relevant". It has been observed for the UK that what is written down in the police transcript (the ROTI), is more relevant for the prosecution than for the defense (Haworth, 2018). In my Dutch police transcripts I found that there is an orientation to building a case, but not specifically for prosecutors.[7]

As I have shown, my transcriptions were modified and adapted to what I thought at that moment was relevant for my research questions and necessary for my analyses. Another type of academic selectivity is the choice of fragments to be analysed and discussed in the publications. Police transcripts are meant to cover all the relevant items of the entire interrogations, and so are mine in the first instance. But when I come across a phenomenon

[7] This may be related to differences between the accusatorial and inquisitorial criminal law systems.

worthy of further study, I do a "data run" in my materials to find similar or dissimilar instances. From these instances I select those fragments that I can build upon to further my analyses. And when preparing a publication I make another selection of instances that will fit the organisation of the publication and the publication standards.

As judges, prosecutors and defense lawyers select items from the police reports in the execution of their professional tasks, so do I. Thus, police transcripts and my academic transcripts are different in the amount of detail they contain, but similar in that they are constructed on the basis of their relevance for the uses to which they are put.

The Status and Treatment of Transcripts

What struck me when I studied the references to and quotations from the police reports by the judges in court was that, even when it was clear that the sentences they read aloud would never be uttered like that by a suspect (see for example fragment 2), judges treated the suspects as having said what was written down in the police report as their own production, and held them accountable for it (Komter, 2019). This can be attributed to the language ideologies of decontextualised fragments and of narrator authorship (Eades, 2012: 447–448), which encompass a disregard of the interactional context in which the suspect's statement was elicited, and ignore the co-authorship of the police transcripts.

I realised that I am doing something similar. I refer to my transcription as if it were the talk rather than a representation of the talk (see my treatment of the examples 1–8). I also realised that this is common practice in Conversation Analytic or Discourse Analytic research publications (but see: Haworth, 2018). My research aims are to analyse the talk, not the transcript, and to compare the police transcript with the talk in the interrogation, not with my transcript. Yet, when it comes to write down and publish my results, I can only demonstrate differences between two kinds of text: the institutional transcripts of the police officers and my academic transcripts[8].

The legal professionals who deal with police transcripts later on in the criminal process must know that what they read cannot be exactly the same as what the suspect actually said, and academic professionals know that transcripts can never capture all the details of talk in interaction. The habit of treating the text as the talk, both by institutional and academic professionals, can be explained by the focus on their primary tasks. For legal professionals these are to study and decide on a criminal case, for academics to study the talk. Institutional and academic professionals, for all practical purposes, take transcripts at their face value, as this is part of their professional routine and instrumental for getting their work done. At the same time, academic and institutional professionals should be aware of the specific limitations of their transcripts, as the stakes are high. For criminal law practice the quality of the transcripts may ultimately affect the justice of the criminal process, for academic research the validity of the findings.

[8] Occasionally, publications contain links to the sound clips of the articles. See: https://www.sscnet.ucla.edu/soc/faculty/schegloff/sound-clips.html.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because the data are subject to restrictions of the Dutch Prosecution Office. Requests to access the datasets should be directed to the Dutch Prosecution Office.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

This publication was funded by the VUvereniging publication fund.

ACKNOWLEDGMENTS

I am grateful for the pertinent and insightful comments of the anonymous reviewers.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2022.797145/full#supplementary-material

REFERENCES

Bucholtz, M. (2000). The politics of transcription. *J. Pragmatics* 32, 1439–1465. doi: 10.1016/S0378-2166(99)00094-6

Bucholtz, M. (2007). Variation in transcription. *Discourse Stud.* 9, 784–808. doi: 10.1177/1461445607082580

Clayman, S. (1990). From talk to text: newspaper accounts of reporter-source interactions. *Media Cult. Soc.* 12, 79–103. doi: 10.1177/016344390012001005

De Boer, M. (2014). *"Ik Zal Het in Die Bewoordingen op Papier Zetten". De Verschillen Tussen Politieverhoren en Processen-Verbaal. ["I Shall Put It on Paper in Those Words". The Differences Between Police Interrogations and Police Reports]*. NSCR Report.

Drew, P. (2006). "When documents 'speak': documents, language and interaction," in *Talking Research: Talk and Interaction in Social Research Methods*, eds. P. Drew, G. Raymond, and D. Weinberg (London: Sage), 63–80.

Eades, D. (2012). The social consequences of language ideologies in courtroom cross-examination. *Lang. Soc.* 41, 471–497. doi: 10.1017/S0047404512000474

Franken, A. A. (2010). Regels voor het strafdossier [Rules for the criminal case file]. *Delikt en Delinkwent* 40, 403–418.

Galatolo, R., and Drew, P. (2006). Narrative expansions as defensive practices in courtroom testimony. *Text Talk*. 26/6, 661–698. doi: 10.1515/TEXT.2006.028

Garfinkel, H. (1967). *Studies in Ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall.

Goffman, E. (1981). "Footing," in *Forms of talk*, E. Goffman (Oxford: Basil Blackwell), 124–159.

Goodwin, C. (1996). "Transparent vision," in *Interaction and Grammar,* E. Ochs, E. A. Schegloff, S.A. Thompson (eds.). (Cambridge ST: Cambridge University Press), 370–404.

Goodwin, C. (2000). Action and embodiment within situated human interaction. *J. Pragmatics* 32, 1489–1522. doi: 10.1016/S0378-2166(99)00096-X

Greatbatch, D., Heath, C., Luff, P., and Campion, P. (1995). "Conversation analysis: human-computer interaction and the general practice consultation," in *Perspectives on HCI. Diverse Approaches*, A. F. Monk, and G. N. Gilbert (London: Academic Press), 199–222.

Haddington, P., Keisanen, T., Mondada, L., and Nevile, M. (2014). "Towards multiactivity as a social and interactional phenomenon," in *Multiactivity in Social Interaction: Beyond Multitasking*, eds. P. Haddington, T. Keisanen, L. Mondada, and M. Nevile (Amsterdam; Philadelphia: John Benjamins), 3–32.

Harper, R. H. R. (1998). *Inside the IMF. An Ethnography of Documents, Technology and Organisational Action*. San Diego, CA: Academic Press.

Haworth, K. (2018). Tapes, transcripts and trials: the routine contamination of police interview evidence. *Int. J. Evid. Proof.* 22, 428–450. doi: 10.1177/1365712718798656

Hepburn, A., and Bolden, G. B. (2012). "The Conversation Analytic approach to transcription," in *The Handbook of Conversation Analysis*, eds. J. Sidnell, and T. Stivers. Chichester: Wiley Blackwell, 57–76.

Heritage, J. (1984). *Garfinkel and Ethnomethodology*. Cambridge: Polity Press.

Heritage, J., and Watson, R. (1979). "Formulations as conversational objects," in *Everyday Language. Studies in Ethnomethodology*, ed. G. Psathas (New York, NY: Irvington Press), 123–162.

Jefferson, G. (1983). Issues in the transcription of naturally-occurring talk: caricature versus capturing pronunciational particulars. *Tilburg Pap. Lang. Lit.* 34, 1–12.

Jefferson, G. (2004). "Glossary of transcript symbols with an introduction," in *Conversion Analysis: Studies From the First Generation*, ed. G. H. Lerner (Amsterdam; Philadelphia: John Benjamins), 13–31.

Komter, M. L. (2002–2003). The construction of records in Dutch police interrogations. *Inf. Des. J. Docum. Des.* 11, 201–213. doi: 10.1075/idj.11.2.12kom

Komter, M. L. (2003). The interactional dynamics of eliciting a confession in a Dutch police interrogation. *Res. Lang. Soc. Interact.* 36, 433–470. doi: 10.1207/S15327973RLSI3604_5

Komter, M. L. (2006). From talk to text: the interactional construction of a police record. *Res. Lang. Soc. Interact.* 39, 201–228. doi: 10.1207/s15327973rlsi3903_2

Komter, M. L. (2019). *The Suspect's Statement: Talk and Text in the Criminal Process.* Cambridge: Cambridge University Press.

Lynch, M. (2015). "Turning a witness: the textual and interactional production of a statement in adversarial testimony," in *Law at Work. Studies in Legal Ethnomethods*, eds. B. Duprez, M. Lynch, and T. Berard (Oxord: Oxford University Press), 163–189.

Malsch, M., de Keijser, J., de Gruyter, M., Komter, M. L., and Elffers, H. (2012). *Het Opmaken Van Proces-Verbaal Van Een Verdachtenverhoor: Ervaringen en Oordelen Van Verbalisanten [The Construction of a Police Report of Suspect Interrogations: Experiences and Opinions of Reporting Officers]*. Amsterdam: NSCR report.

Moerman, M. (1988). *Talking Culture. Ethnography and Conversation Analysis.* Philadelphia, PA: University of Pennsylvania Press.

Mondada, L. (2007). Commentary: transcripts variations and the indexicality of transcribing practices. *Discourse Stud.* 9, 809–820. doi: 10.1177/1461445607082581

Mondada, L. (2014). "The temporal orders of multiactivity: operating and demonstrating in the surgical theatre," in *Multiactivity in Social Interaction. Beyond Multitasking*, eds. P. Haddington, T. Keisanen, L. Mondada, and M. Nevile (Amsterdam; Philadelphia: John Benjamins), 33–75.

Mondada, L. (2018). Multiple temporalities of language and body in interaction: Challenges for transcribing multimodality. *Res. Lang. Soc. Interact.* 51, 85–106. doi: 10.1080/08351813.2018.1413878

Mondada, L., and Svinhufvud, K. (2016). Writing-in-interaction. Studying writing as a multimodal phenomenon in social interaction. *Lang. Dialog.* 6, 1–53. doi: 10.1075/ld.6.1.01mon

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language* 50, 696–735. doi: 10.1353/lan.1974.0010

Slembrouck, S. (2007). Transcription - the extended directions of data histories: a response to M. Bucholtz's 'Variation in transcription'. *Discourse Stud.* 9, 822–827. doi: 10.1177/1461445607082582

Smith, D. (1974). The social construction of documentary reality. *Sociol. Inquiry* 44, 257–268. doi: 10.1111/j.1475-682X.1974.tb01159.x

Smith, D. (2001). Texts and the ontologies of organization. *Stud. Cult. Organ. Soc.* 7, 159–198. doi: 10.1080/10245280108523557

Van Charldorp, T. (2011). *From Police Interrogation to Police Report*. Oisterwijk: BOXpress.

Van den Adel, H. M. (1997). *Handleiding Verdachtenverhoor. Handhaving, Controle en Opsporing in de Praktijk (Manual for Suspect Interrogation. Upholding the Law, Control and Detection in Practice)*. Den Haag: VUGA.

Watson, R. (2009). *Analysing Practical and Professional Texts: A Naturalistic Approach*. Farnham: Ashgate.

Whalen, J. (1995). "A technology of order production: computer-aided dispatch in public safety communication," in *Situated Order. Studies in the Social Organization of Talk and Embodied Activities*, eds. P. Ten Have, and G. Psathas (Washington, DC: University Press of America), 187–230.

Zimmerman, D. H. (1969). "Record-keeping and the intake-process in a public welfare agency," in *On Record: Files and Dossiers in American Life*, ed. S. Wheeler (New York, NY: Russell Sage Foundation), 319–354.

Zimmerman, D. H. (1992). "The interactional organization of calls for emergency assistance," in *Talk at Work*, eds. P. Drew, and J. Heritage (Cambridge: Cambridge University Press), 418–469.

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 11 April 2023
DOI 10.3389/fcomm.2023.1096639

OPEN ACCESS

EDITED BY
Felicity Deamer,
Aston University, United Kingdom

REVIEWED BY
Heidi Salaets,
KU Leuven, Belgium
Franca Orletti,
University Rome Three, Italy

*CORRESPONDENCE
Miranda Lai
✉ Miranda.lai@rmit.edu.au

SPECIALTY SECTION
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

RECEIVED 12 November 2022
ACCEPTED 20 March 2023
PUBLISHED 11 April 2023

CITATION
Lai M (2023) Transcribing and translating forensic speech evidence containing foreign languages—An Australian perspective. *Front. Commun.* 8:1096639. doi: 10.3389/fcomm.2023.1096639

Transcribing and translating forensic speech evidence containing foreign languages—An Australian perspective

Miranda Lai*

Translating and Interpreting, Royal Melbourne Institute of Technology (RMIT University), Melbourne, VIC, Australia

There is a growing body of literature on forensic transcription of covert recordings obtained by clandestine law enforcement operations. Due to the nature of these operations, the quality of the recordings, particularly those obtained by planting listening devices in a car or a house, is often extremely poor. When tendering such recordings as evidence in court for prosecuting an alleged crime, a transcript will often accompany the recording to assist the triers of fact (i.e., judges and jurors) to hear better. In the context of multilingual and multicultural Australia, often such forensic recordings may contain languages other than English, and therefore a translation into English is required to facilitate understanding of the verbal exchanges in the recording. Little is known, however, about the people engaged by law enforcement to undertake these forensic translation tasks, what qualification and training they possess, how they carry out the tasks, and if there is a system to safeguard the quality and reliability of their translation output. This paper reports on an online survey conducted in Australia on professional interpreters and translators who have been engaged to perform this type of work. Descriptive statistics and thematic analysis of text answers provide a qualitative account of the status quo which has not been documented before. Deficiencies of the current practice and its associated risks are identified. Recommendations are proposed as the first step to address the issues identified.

KEYWORDS

forensic transcription, forensic translation, forensic speech evidence, interception/covert recording, legal translation, legal translator, legal interpreter

1. Introduction

Law enforcement agencies at times need to engage in clandestine operations to obtain private communications to solve or prevent crimes. In an increasingly globalized world where crimes do not observe national or linguistic boundaries, covert recordings law enforcement obtain often contain foreign languages. Australia is a case in point. Professional translators and interpreters are, therefore, often engaged by law enforcement in these situations to overcome language barriers, thereby allowing investigators to carry out their investigative tasks and/or to prepare forensic linguistic evidence for court trials. For investigative purposes, professional interpreters may be employed to listen to live or covertly recorded telephonic communications and asked to provide investigators with either the gist or a full interpretation of the exchanges under surveillance; they may also be asked to identify matters of interest or scour for specific items of information instructed by the investigator. For evidentiary purposes, although the actual recording is regarded as the primary evidence and the transcript as secondary (Gilbert and Heydon, 2021), triers of fact (i.e., judges and jurors) must rely on the translation into English of the original utterances in the audio to access the meaning of the exchanges spoken in a foreign language that they do not understand.

This paper reports on an online survey conducted in Australia on the experiences of translators and interpreters involved in forensic transcription and translation (FTT) for law enforcement for both investigative and evidentiary purposes. It provides insights into this under-researched interdisciplinary area of criminal justice and translating and interpreting (T&I) studies to establish an understanding of current practice and issues which need urgent attention.

2. Background

Wiretapping operations conducted by law enforcement can be categorized into two macro-types: telephone intercepts, which are achieved by telephonic listening interventions, and environmental recordings, which are made by planting listening devices in the environment of the targeted speaker (Fraser, 2014; Romito, 2017). It should be noted, though, that with the advancement of communication technologies, the former has now become more relevant to interceptions of private messages via mobile phone, Voice Over Internet Protocol (Butterfield et al., 2016), social media post, or email. The audio quality of the latter (i.e., environmental recordings such as the bugging of a house or a car) is normally extremely poor (Fraser, 2017), "to the extent that, without prior knowledge of the contents, few if any words can be clearly identified" (Fraser and Stevenson, 2014; p. 206). These covert recordings may be used to serve two purposes: investigative or evidentiary (French and Harrison, 2006; Haworth, 2010; Fraser, 2014). For the former, information from the covert recording is used to help law enforcement "uncover the facts surrounding an alleged crime" (Fraser, 2014; p. 8), for example, when the persons of interest will be meeting, where, and to do what. If successful, the outcome of the operation becomes evidence in the court trial (Fraser, 2014). In these situations, when investigators are faced with poor quality covert recordings, they can combine their insights on the case at hand and form an educated guess about what is possibly being said in the unclear or indistinct audio, thereby deciding their next action or the direction they should take in their active investigation. They do not need to justify to anyone how they "interpret" the indistinct audio to reach what they think the utterances are. On the other hand, when a case enters legal proceedings and if the covertly obtained recordings are going to be used by law enforcement as evidence to prove guilt, the recordings become forensic speech evidence and serve evidentiary purposes. The recordings may "capture a criminal offense being committed or can contain incriminating (or exculpating) material, including admissions of guilt, involvement, or knowledge of criminal activity" (Love and Wright, 2021; pp. 1–2). Fishman (2006) aptly describes the evidentiary value of conversations captured in covert recordings:

> Few, if any, forms of evidence are likely to be as probative—or as devastating. We see this most often in criminal cases: rather than rely on the testimony of witnesses who may be vulnerable to various forms of impeachment, a prosecutor simply allows a defendant's words [caught on recording] to speak for themselves. (p. 475)

Fishman (2006) further asserts that a jury's ability to use such evidence depends on two qualities of the recording: audibility and intelligibility. The former relates to whether the listener can hear what is on the recording, while the latter is about whether the listener is able to understand what is being said. When covert recordings with poor audibility and/or intelligibility are presented in Australian courts, the law allows the jury to be provided with a transcript prepared by police to help jurors hear better relevant utterances and attribute each to a speaker (Fraser and Loakes, 2020). These indistinct recordings are often transcribed by police detectives or officers involved in the case, or what Fraser (2014) calls "involved transcribers" (p. 12), with no training in transcription at all.

Transcription is highly complex, meticulous, and onerous even for clear recordings (Jenks, 2013). For covert recordings, it is clearly not the intention of the speaker to be (over) heard by a third person, therefore the possible "messiness" of the talk unlike a monitored talk, e.g., a courtroom exchange or police interview, which will be much more orderly. Transcribing covert recordings is particularly challenging because the "ground truth," that is, the accurate, incontestable knowledge of what was really said, is not available (Fraser and Loakes, 2020; p. 416). It is, therefore, problematic that the police transcribers may be "hindered by having contextual information that is potentially unreliable (having not yet been tested by the trial process)" (Fraser and Loakes, 2020; p. 417). Using untrained police officers who have a vested interest in the influence of the transcript on a case gives rise to potential inaccuracy (Love and Wright, 2021). There has been growing concerns about unreliable transcripts and their priming effect on jurors. Empirical evidence has shown that once triers of fact are presented incorrect and misleading transcripts, they are unable to unseen them (e.g., Fraser et al., 2011; Fraser, 2014, 2021, Fraser and Stevenson, 2014), or in Fraser and Loakes (2020) term, to "reset their perception to give equal consideration to alternative interpretations" (p. 418), and their confidence does not seem to diminish considering their "inability" to hear (e.g., Fraser, 2018; Fraser and Kinoshita, 2021). Unreliable transcripts, therefore, give "extraordinary privilege for the police interpretation of indistinct covert recordings" (Fraser and Loakes, 2020; p. 418) and increase the risk of innocent people being convicted and the guilty set free (Gilbert and Heydon, 2021).

What is described so far is also true when the covert recordings contain languages other than English (LOTEs) in the Australia context. In such situations, regardless of the audio quality being acceptable, poor, or indistinct, law enforcement is unable to transcribe nor translate the audios themselves. Little is known about who perform FTT tasks, what translation approaches are adopted, and how the quality and reliability of the translation into English is attained and assessed. This paper intends to address this gap of knowledge.

3. Literature review

Scholarship on transcribing covert recordings containing LOTEs and its implications is scant. As a starting point, such FTT tasks obviously must be undertaken by people who are speakers of the same foreign language as in the audio, and in Australia and other Anglophone criminal jurisdictions, interpreters and translators are often engaged; similarly, in European countries such as Belgium "sworn translators-interpreters" are engaged to provide the service for legal wiretapping (Salaets et al., 2015), i.e. intercepted communication, while in Switzerland "intercept interpreters" are engaged (Capus and Griebel, 2021; Capus and Grisot, 2022; Capus and Havelka, 2022). American legal interpreting scholars González et al. (2012) regard FTT as "one of the most demanding and rapidly growing areas of legal interpretation" (p. 965), and therefore devote an entire chapter to this topic in their seminal volume, *Fundamentals of Court Interpretation: Theory, Policy, and Practice*. They assert that the primary purpose FTT serves is to "provide an impartial, accurate, complete, legally equivalent, and contextually sound transcription/translation from the SL [source language] to the TL [target language]" (González et al., 2012, p. 991), while advocating the need for specialized training for FTT in response to the hybrid nature of a task that calls for interpreting, translating, and task-specific skills (see also Mikkelson, 2016). Sections 3.4 and 3.5 will cover the scholarly views about the nature of the work and the required skills and knowledge.

It should be mentioned that González et al. (2012) chapter on FTT has a different focus from the current paper. Their chapter is mostly concerned with transcribing and translating police interviews in the US, both custodial and noncustodial, where "putative interpreters" [Calmeyer, 2010, as cited in González et al. (2012); p. 967] are used, that is, where police officers who have unspecified Spanish language competence double as interpreters, therefore creating miscommunication and harming the interviewee's defense. In these circumstances, T&I practitioners do not deal with covert recordings of suspected criminal activities. Rather, they deal with police interview recordings, which are generally of better audio quality, and all participants to the interview are aware of the recording taking place (i.e., overt recording). However, regardless of whether recordings are overt or covert, the principles that González et al. (2012) advocate—to produce quality and reliable FTT—are equally applicable. This will be explicated in Sections 3.1–3.3.

It is also worth pointing out that the emerging European literature referenced before approaches the activities undertaken by "sworn translators-interpreters" (in Belgium) or "intercept interpreters" (in Switzerland) from a slightly different perspective. It is rightly concerned about how T&I practitioners' agency and work practices in the law enforcement operation, investigation, and prosecution phases, therefore their "visibility" or, rather, "invisibility" which leads to ethical and ontological questions in their respective inquisitorial systems. While the current paper focuses more specifically on the probity, quality, and reliability of forensic speech evidence used in the adversarial criminal justice system in Australia accompanied by translations produced by T&I practitioners, the commonalities in relation to the challenges and issues faced by Australian practitioners will be remarked upon where appropriate.

3.1. Two-step process

According to González et al. (2012), FTT should be a two-step process: first, producing an orthographic transcript of the original language caught in the recording; and then translating the transcript into the target language for forensic purposes (English in the case of the US). This is because that "without the critical step of transcribing the speech event into textual form, an accurate and verifiable translation is not possible" (p. 1006). Whether such an approach is followed by T&I practitioners is a separate matter, and the survey reported in Section 5 will shed light on the reality in Australia.

The starting point of the judiciary is often that all transcripts provided by the prosecution (whether in English or translation into English from a foreign language) are accurate and fit for purpose for trials (Gilbert, 2014), and from there the defense can attempt to create uncertainty in trials about the meaning alleged by the prosecution (González et al., 2012). Although, as mentioned before, the primary evidence is the audio and the transcript is secondary (Gilbert and Heydon, 2021), in reality, audio recordings are not necessarily played in court trials for practical reasons: if the audio is in English, reading the transcript is easier for the triers of fact to visualize the words, as opposed to listening to ephemeral sounds in the recording; and if the audio is in a foreign language, there is even less incentive to play it, since triers of fact will have to rely on the translation anyway. Either way, jurors rely heavily on the transcript, unless there is a particular point the prosecution or the defense attempt to make about the recording, in which case the audio may be played. If the utterances in a foreign language in the translation provided to the jury are disputed by the defense, often the court interpreter may be asked to listen to the recording on the spot and provide their version of translation for counsels to further explore and confirm meaning. In theory, the prosecution will make the transcript available to the defense before trial for the defense to check and mount challenges to its accuracy; if it is a translation, the defense can employ their own T&I services to verify and rectify points of differences to arrive at an agreed version with the prosecution. However, in reality, the defense often does not have the resources nor sophistication to undertake such checking. In the current system, no one really knows if the translation produced by T&I practitioners is accurate (Gilbert, 2014). In the US context, González et al. (2012) assert that once the translation is entered into evidence without objection, "defense attorneys lose the opportunity to appeal, challenging the reliability of the evidence, and the LEP [Limited English Proficient] defendant faces a greater risk for wrongful conviction" (p. 977). According to Capus and Griebel (2021), intercepted communication is often not transcribed first in Switzerland either, and "different procedures seem to be utilized within the Swiss cantons and police stations regarding whether a transcript is produced in the original language before translation" (Capus and Havelka, 2022; p. 1830). The recommended two-step process engenders a better audit trail (Gilbert and Heydon, 2021) for the accused to "determine whether the transcript accurately corresponds to the recording [in the original foreign language], even though he/she may not be in a position to evaluate the accuracy of the translation" [National Association of Judiciary Interpreters and Translators (NAJIT, 2019; p. 5)]. González et al.

(2012) hold that the constitutional rights of the defendant are infringed when they are only provided with a translation into English without a transcript of what they are alleged to have said in the foreign language in the recording.

3.2. Verbatim orthographic transcript

In relation to producing orthographic transcript in LOTE, that is the first step of the two-step process for evidentiary purposes, scholarly views converge on Fishman's (2006) "mirror the tape" rule (pp. 494–495), which is to include what can actually be heard on tape. Further, González et al. (2012) assert that "all the linguistic, sociolinguistic, pragmatic, and discoursal elements of the speech event" (p. 992) in the audio should be transcribed, and that "clearly discernible paralinguistic features, such as pauses, changes in voice, tone, volume, silences or hesitations, hedges, false starts, or interjections, also need to be documented via the application of the legend system" [González et al., 2012; p. 992; see also Mikkelson (2016)]. The suggested legend system referred to is intended to enable the transcript reader to reconstruct meaning more holistically (Mikkelson, 2016) when there are "paralinguistic or sociolinguistic elements that may not be explicitly stated, but are present and do carry meaning" (González et al., 2012, pp. 1039–1041). Appendix 1 shows the LOTE transcription guidelines González et al. (2012) propose. The conventions and symbols they recommend using largely conform with the Jeffersonian transcription system (Jefferson, 2004) used to transcribe English discourse.

It should be noted, though, that transcripts can never be a full representation of spoken discourse, which comes with an almost infinite number of nuances and layers of social interaction due to limitations of space (Jenks, 2013), therefore the possibility and practicality of including all details as suggested by González et al. (2012). Considering the purpose of the transcript (and its subsequent translation) advocated by Capus and Griebel (2021) holds much truth. There should be communication between the transcriber and the user of the transcript to agree on the desired level of details required for the transcript or when/where detailed discoursal information is required, as this has implications for the time it takes to produce the transcript, therefore the cost.

3.3. Translation of transcript

Once the transcript captures all necessary linguistic, paralinguistic, and extralinguistic elements (if required), an impartial, accurate, complete, legally equivalent, and contextually sound translation can then be produced, without editing, summarizing, deleting, or adding any information, while conserving the non-English speaker's language level, style, tone, and intent (González et al., 2012). González et al. (2012) go so far as to suggest that T&I practitioners should, in producing the translation, clarify in a footnote when "gesture, feature, or utterance is culturally bound or contains significant linguistic or sociolinguistic information" (p. 992).

Gilbert (2014; Gilbert and Heydon, 2021) documents various FTT issues from Vietnamese into English in drug related cases heard in the Victorian County Court in Australia. Notably the Vietnamese term "ấy", which is an exophoric or anaphoric reference word similar to the term "it" in English, was translated numerous times in the telephone intercepts as "thingy". The Crown alleged that "thingy" was a coded word for drugs. Yet there is no evidence in the original utterances that such a coded word or any other word exists that can be translated as "thingy" within the context of the communication. According to González et al. (2012), a literal translation approach in these high-stake situations should be used, because "the potential for prejudice is too great" (p. 991), and they recommend that the meanings of coded words be left to be professed in testimony as expert opinion by police.

The National Association of Judicial Interpreters and Translators (NAJIT, 2019) in the US endorses the two-stage process of FTT, namely, transcribing in the original language first before translation. They acknowledge FTT to be very time-consuming and exacting, citing an industry standard of up to one hour of transcribing work for every minute of conversation in a forensic recording, which does not include the subsequent translation. NAJIT (2019) further asserts that given all that is at stake in a criminal matter, there is no justification for cutting corners (see also Mikkelson, 2016; p. 69). It should be noted that in reality this NAJIT proposition will be hard to attain since the FTT costs will be prohibitive. Maintaining a balance between readability and accuracy (Tilley, 2003) should be achievable, though, through communication between the transcriber and person commissioning the work as suggested in the previous section so the FTT outcome is adequate to serve the intended purpose.

3.4. Intermodal translation

FTT is fundamentally a "translational activity *sui generis*" (Capus and Havelka, 2022; p. 1817), in that it entails an auditory input in the SL and a written output in the TL, which distinguishes it from conventional translation (text input to written output) and conventional interpreting (auditory input to oral output). Influential Russian linguist Jakobson (1959) delineates three ways of deciphering verbal signs:

> (a) *intralingual translation* or *rewording*, an interpretation of verbal signs by means of other signs of the same language; (b) *interlingual translation* or *translation proper*, an interpretation of verbal signs by means of some other language; and (c) *intersemiotic translation* or *transmutation*, an interpretation of verbal signs by means of signs of nonverbal sign systems. (p. 233)

This Jakobson's framework is insufficient to describe the hybridity of FTT activities. Israeli translation theorist Toury (1994/[1986]) further delineates translation under Jakobson's typology into *intersemiotic* versus *intrasemiotic*, where the latter (which FTT applies) is further divided into *intrasystemic* (i.e., intralingual) translation versus *intersystemic* (i.e., interlingual) translation.

The two steps of FTT advocated by González et al. (2012) involve acts of translation: the first step corresponds to Jakobson (1959) intralingual translation as well as Toury's intrasemiotic and intrasystemic translation, while being a kind of intermodal transfer (Kaindl, 2012), i.e., from auditory to written. The second step is in Jakobson's term interlingual translation as well as a kind of intramodal transfer (Kaindl, 2012), i.e., both input and output are in written form, while it is intersystemic in Toury's term. Table 1 summarizes the two-step transcribe–translate process and how they correspond to the different translation typologies.

The first step of the two-step process—is no different from monolingual transcription from spoken English to written English which, as Fraser (2022) aptly points out, requires interpretation and decision-making by both its creator and by its end-user, and that no transcript is ever "the" transcript, rather "a" transcript. In this sense, Orletti and Moriottini (2017) also acknowledge that the transcriber "inevitably makes selections" (p. 3), and therefore transcription is never a neutral action. T&I practitioners engaged to undertake FTT, like other types of interlingual transfer tasks, must have not only linguistic competence, but also intertextual, psychological, and narrative competence (Eco, 2001; p. 13). Available T&I scholarship does not have applicable models as yet for the intermodal operation of transcribing covert recordings (i.e., Step 1 in Table 1), nor interlingual translation from the foreign language in the transcript into English (i.e., Step 2 in Table 1), which should be the direction of future scholarly endeavor.

3.5. Required skills and knowledge for FTT

Bucholtz (2007) asserts that transcription is "a sociocultural practice of representing discourse" (p. 785), while Orletti (2017) describes it as "extracting chunks of a social interaction and fixating its 'flowing' on a printed page ... [by doing so, turning] those chunks into movable items that can be repositioned into other contexts" (p. 13). Italian scholars Paoloni and Zavattaro (2007; p. 139) remark on a lack of academic curriculum for training experts in dealing with intercepted telephone calls and undercover recordings, while Bellucci (2022, as cited in Orletti, 2017) echoes the same deficiency of specific training for both police professionals and experts of forensic transcription. To successfully perform forensic transcription (intralingually), Orletti (2017) states that one must possess linguistic, phonetic, dialectological, sociolinguistic, and technological competencies.

Considering FTT as a hybrid translational activity *sui generis,* the required knowledge and skills for T&I practitioners to undertake FTT, therefore, comes into question, as is explicated in the NAJIT (2019) position paper on FTT:

> Not all interpreters are adept at transforming the spoken word into written text with the accuracy required in the legal setting. By the same token, professional translators may lack the training to accurately transform live recorded extemporaneous speech into written form. Translators may also not be familiar with non-standard usage and jargon, as well as not being accustomed to documenting the errors and misspeaks that often color the speech of individuals with limited or no formal education. Consequently, not all translators can successfully render an authentic and accurate forensic transcription translation. (pp. 1–2)

González et al. (2012) observe that the field of FTT remains a "largely ungoverned, unlicensed, and nonprofessional practice," arguing that "until there is acceptance of this field as a subspeciality of interpreting and the establishment of credentialing or certification, there will be great variability in product quality" (p. 980). The authors go so far as to suggest a *master-level* FTT specialist, who is certified for their higher level of skills and expertise with additional knowledge, experience, and academic credentials, and who not only provides routine FTT services, but also specializes in reviewing FTT work performed by others when FTT evidentiary materials are challenged by any of the parties, or when the judge orders ad hoc independent review or independent transcription/translation.

In addition to primary skills of language proficiency, cultural knowledge, and linguistic knowledge as well as an understanding of forensic linguistics (Kredens et al., 2021) which is not dissimilar to competencies required for monolingual transcription, González et al. (2012) also propose the following five personal traits for T&I practitioners to possess for FTT tasks:

1. A highly attuned, perceptive ear
2. Analytic and problem-solving skills
3. Research skills
4. Organizational skills
5. Attention to detail

It appears that apart from the first trait, which is more specific to the task of transcription, the rest tend to be soft skills that are generic to a lot of professions.

3.6. Recommended FTT formatting

González et al. (2012) recommend a four-column presentation of FTT (as shown in Table 2) in order to be clear and accountable, a recommendation endorsed by NAJIT. The first column denotes line numbers for easy reference. The second column attributes the speaker to the utterance transcribed in the line and is distinguished by male or female voices represented as MV1 (male voice 1) or FV1 (female voice 1); and as far as the transcriber can tell whether the voice belongs to the same speaker in the same recording, or a different voice, therefore MV2, MV3, and so forth, or FV2, FV3, and so forth. The third is the verbatim orthographic transcription of the SL utterance, and finally the last column is the translation from the text in the third column.

3.7. Translation of text messages

With the advancement of communication technologies and the popularization of computer-mediated communication (CMC)—defined as text, images and other data received via computer (Wainfan and Davis, 2004; p. 4) either synchronously (e.g., online chat or text message) or asynchronously (e.g., webpage

TABLE 1 FTT typologies.

Two-step process	Jakobson's typology	Kaindl's typology	Toury's typology
Step 1: transcription [From SL spoken utterances to SL written utterances]	Intralingual translation	Intermodal transfer	Intrasemiotic translation + Intrasystemic translation +
Step 2: translation [From SL written utterances to TL written utterances]	Interlingual translation	Intramodal transfer	Intrasemiotic translation + Intersystemic translation

TABLE 2 Recommended FTT format.

Line #	Speaker	Source language transcription: Spanish	Target language translation: English
61	MV1	*Betito, mira, yo no te voy a chingar . . .*	Betito, look, I'm not going to screw you over
62	MV2	*Yo sé. . . yo sé. . . pero yo no conozco a ese vato y. . .*	I know. . . I know. . . but I don't know that dude and. . .
63	MV2	*¿Y qué?... tú sabes que el lo-. . . el loco* [U] *allá con tu ruca.*	So what? . . . you know that foo-. . . that fool [U] over there with your old lady.
64	MV3	[U] [Loud motor in background]. . . *¿Cuánto traes?*	[U] [Loud motor in background] . . . How much you got on you?
65	MV2	*Aquí traigo 2 kilos* [U]. . . *yerba. . . La carga, este, la carga está en mi camioneta.*	I've got 2 kilos here [U]. . . yerba[a] . . . The load, uh, is in my *camioneta.* [Translator's note: The term *camioneta* may mean a station wagon, pickup truck, camper, or van; it is not clear from the context which type of vehicle is meant.]
66	MV3	*¿Cuándo le dijistes que* [U]*? ¿Hoy domingo o mañana martes?* [sic]	When did you tell him [U]? Today Sunday or tomorrow Tuesday? [*sic*] [Electronic noise from 4:23 to 4:48]

[a] Primary denotation of yerba is "weed".
MV1, Male Voice 1; MV2, Male Voice 2; MV3, Male Voice 3.
Adapted from the table in González et al. (2012, p. 1036).

or email) (O'Hagan and Ashworth, 2002), private messages have become increasingly important in crime investigations. This has necessitated the engagement of translators to assist in converting such communications in the text format from a foreign language into English in a forensic context. According to Capus and Havelka (2022), intercept interpreters in Switzerland translate text messages as part of their work, together with live and recorded conversations. Text messaging, as a form of CMC, is a unique way of communicating, which manifests in written and visual structures, but embodies the characteristics of spoken discourse with all the elements and complexities of oral communication; this hybrid nature lends itself to "finger speech," in that it is as if the fingers speak the minds of the communicators (Cal-Meyer, 2016, para. 2). Similarly, live chats as another form of CMC are observed by O'Hagan and Ashworth (2002) to be like spoken discourse which are fraught with "anomalies such as misspellings and grammatical errors. . . [and are] characterized by the use of online jargon and topic fluidity" (p. 55).

In the Western Australia case R v Yang [2016] WASC 410 (Auslii, 2017), translations of text messages from Korean into English between a drug trafficking suspect with alleged accomplices came under question. The defense challenged a number of aspects of the translation of the text messages and argued that the approach taken by the translator was inconsistent with the AUSIT Code of Ethics in that a translator should preserve the "content and intent of the source message or text without omission or distortion" (AUSIT, 2012). Justice Fiannaca ruled that there were several deficiencies in the translator's evidence, including:

1. lack of translator's notes when translating a laughing emoticon into "ha"
2. lack of translator's notes when disregarding certain parts of the messages and clean up typographical errors in the messages, which could have been ambiguous
3. not reproducing all the laughing emoticons as were in the original text messages
4. repeated Korean expressive characters were translated into a single "oh" or "ah" sound, which lost its expressive characteristic
5. expressing an opinion about a conclusion to be drawn from a message (KordaMentha, 2018).

It is, therefore, important to note that the inaccuracies in the translation in this case led to limitations for the court to draw inferences from the affected messages, and the opinions provided by the translator outside of his field of expertise were manifestly disregarded (KordaMentha, 2018). His honor states that "the challenge [by the defense] . . . was not to the witness's [i.e., the translator's] impartiality, but to the accuracy of some of the translations and his methodology. More generally, it could be said that the challenge was to Mr. Y Lee's [i.e., the translator's] reliability as an expert" (Auslii, 2017, para 55). This case serves as a reminder for T&I practitioners to approach this type of forensic translation tasks with great caution and well-informed methodology, as the language in text messages comes with many challenges, according to Cal-Meyer (2016):

- linguistic uncertainties: e.g., grammatical inconsistencies and gaps; spontaneous use of abbreviations, onomatopoeias, alliterations, acronyms; spellcheckers altering the meaning of the message; limited use of discourse markers such as adverbs, conjunctions, prepositions, deictic pointers, and other referents of space, time, and persons/pronouns.

- pragmatic uncertainties: e.g., use of etomitcons, iconic symbols, visual representations familiar only to the texters; intermittent and interrupted conversations.
- cognitive uncertainties: e.g., fragmented and short speaker turns; whimsical ways to economize space on the screen; encapsulation or minimization of ideas, statements, propositions; limited use of cohesive devices; unclear anaphoric references; poor adherence to principles of relevance. Too long with no comments.

This type of translation, unlike translating conventional written text, was referred to as "transterpreting" by Ashworth (1997). Although he coined the term to describe the real-time translation output in the target language for live chats in what can be regarded as the prototype of today's online conference, except for the simultaneity required for "transterpreting", the rest of the translation challenges in relation to the nature of online chats identified by O'Hagan and Ashworth (2002) are very similar to Cal-Meyer's (2016) observations above for translating text messages in a forensic context.

4. Methods

An online survey was designed to collect descriptive statistics and qualitative data to answer an overarching research question: what is the current state of service provision for FTT by T&I practitioners in Australia? The study is important to generate new knowledge to complement the growing body of literature on forensic transcription practice and its evidentiary value in criminal trials in Anglophone countries, which has so far focused on issues arising from monolingual audio materials. The three sub questions of the study are:

1. Who among interpreters and translators are engaged to undertake FTT and what is the required training and/or credential?
2. How do they perform their FTT tasks?
3. What have been their reflections about their FTT experience?

The anonymous and voluntary online survey received ethics approval from the university the author is based, and email invitations to T&I practitioners nationwide with an embedded survey link was distributed through four language service agencies having national presence as well as through the newsletters of the National Accreditation Authority for Translators and Interpreters (NAATI) and the Australian Institute of Interpreters and Translators (AUSIT). The precise size of the survey population is difficult to ascertain, due to the fact that not everyone practicing as an interpreter and/or translator holds a NAATI credential, particularly for low-demand and new-arrival languages in which there are very few or no NAATI-credentialed practitioners. However, the 13,178 practitioners currently holding some form of NAATI credential can be regarded as a reference point; they are reported to cover 147 languages, including the Australian Sign Language, and a further 38 indigenous languages (NAATI, 2021). The survey was open from April 2019 with a closing date extended from the original six months to the end of 2019. The survey questionnaire contains nineteen questions (see Appendix 2), eighteen of which are multiple-choice with free text spaces for the respondent to elaborate on the answer they chose and the last Question 19 is open-ended to elicit further voluntary contribution from the respondents on anything they wished to say about FTT.

Purposive sampling was achieved by the explanation in the email invitation, which stated the purpose of the survey and invited those who had done FTT assignments to self-identify and participate. A total of 356 questionnaires were returned via the university's Qualtrics platform from which the survey was administered. Although the response rate is only less than 3% of the population, it should be noted that not all languages are required for FTT, and that some languages are required more frequently than others. In other words, the actual population of the current study should be T&I practitioners who are involved in FTT. However, presently there is no way to ascertain this more precise population. What can be sure is that the response rate against this more precise population, had it been available, would be much higher than 3%.

Not all respondents answered all questions. As all questions yielded more than 315 responses, except for Question 15 which had 256 answers, it was decided not to exclude questionnaires which missed some questions in order to capture the respondents' contribution to the maximum. Number counts and their corresponding percentages for Questions 1 to 18 were generated by Qualtrics reporting facility, and each question has slightly different overall count depending on how many respondents skipped the question. Free text contributions for Questions 1 to 18 was analyzed using a deductive approach, considering they were specific to the questions asked, and the number of contributions for each question was not large, and thus manual count of relevant meaning units was more efficient for the purpose of further enlightening the quantitative data. On the other hand, the contributions entered for Question 19 were coded using an inductive approach (Braun and Clarke, 2006) using Nvivo 12. The author underwent the analysis in two phases by first reading through the contributions a few times to familiarize herself with the content, which enabled her to identify the central issues (Patton, 2002) and then document her initial thoughts and impressions. This phase was in keeping with the first three steps recommended by Braun and Clarke (2006) for thematic analysis: (1) familiarization with data; (2) generate initial codes; and (3) search for themes. The author then examined the initial themes emerged from the first phase to evaluate their connections, similarities, and difference. This phase reflected the next two steps by Braun and Clark's: (4) reviewing themes, (5) defining and naming themes. Some meaning units were found to relate back to the specific questions and therefore were grouped with the questions to streamline the reporting. The results of the survey are presented in Section 5 below, followed by discussions based on the insights achieved.

5. Results

Of the 19 questions in the survey, Questions 1 and 2 were intended to form the profile of the respondents, which is reported under Section 5.1. Questions 3 to 17 were designed to build a picture of the FTT work practice in Australia. The statistics and insights from these questions were synthesized into seven topics and reported as sub-sections under Section 5.2. The only exception is Question 16 about practitioners testifying their FTT work in courts, which is presented in Section 5.4 separately. As a relevant enquiry but not strictly within the realms of FTT, Question 18 probed the respondents' experience in providing forensic translation services for text messages. This is reported under Section 5.3 separately. Lastly, Section 5.5 reports the three themes arising from the last Question 19 as an open invitation for further thoughts the respondents were willing to share about their FTT experiences.

5.1. Demography of respondents

The profile of the respondents is, to a large extent, a mixture of translating and/or interpreting practitioners, who have some T&I related education and practiced at the professional level for a long time. As can be seen in Table 3 below, among the returned questionnaires, there were 328 who identified as interpreters and 192 as translators. The system is unable to identify the number of respondents who practiced both, although it is safe to say some of them are both interpreters and translators. Only 3% of the respondents reported to have had FTT specific training. A total of 49 LOTEs were reported by the respondents, with the top five languages being Mandarin (n = 55), Arabic (n = 46), Persian (n = 32), Vietnamese (n = 37), and Cantonese (n = 22). The top languages for those who self-identified as translators share exactly the same trend, with the exception of Cantonese being replaced by Spanish. This is because Cantonese is a dialect and Chinese is the language Cantonese speakers read when it comes to translation services.

The majority of the respondent interpreters were NAATI Certified Professional Interpreters (62%), followed by 31% being NAATI Certified Provisional Interpreters[1] More than half of the respondent interpreters (54%) said they had more than 10 years of practicing experience, with the remaining divided among those who had 7–10 years (14%), 4-6 years (18%), and 1–3 years (14%) of experience. More than one in every ten participant interpreters (12%) had no T&I education at all, while just over a quarter (26%) had postgraduate T&I education, followed by 22% and 16% respectively having vocational training at the advanced diploma and diploma levels.

In relation to participant translators (some of whom may also be interpreters), a higher percentage (77%) were certified by NAATI at the professional level,[2] with equally small proportions who reported themselves to be Certified Advanced Translators (4%) and Recognized Practicing Translators (4%). The remaining 15% of participant translators practiced without any credentials. Similar to the trend for the participant interpreters, the highest proportion of this cohort (64%) had more than 10 years of experience practicing, and those with no formal training were slightly more (16%) than the participant interpreters as well as those who had postgraduate education (33%), followed roughly equally those with a bachelor's degree (12%) and those who had vocational advanced diploma training (11%).

5.2. FTT work practice

5.2.1. Engagement pattern and work frequency

T&I practitioners were predominantly engaged in FTT assignments through interpreting agencies or directly from law enforcement (68%, n = 234). A further 25% (n = 87) offers their services directly as a sole trader, while the last 7% (n = 23) stated other ways of being engaged in FTT assignments but provided no further elaboration.

As is shown in Table 4 below, the frequency of FTT assignments appeared to be low, with 58% (n = 189) of the 325 respondents who answered Question 4 said they perform the task less than once a year. Only 6% (n = 19) said they do it more than once a month on average, while the remaining 36% (n = 117) reported doing between one to twelve assignments per annum. Those who undertook FTT assignments most frequently (i.e., more than once a month on average) cover seven languages, while those who did it not as frequently (i.e., between one to twelve times per annum) spread across 19 languages. The five languages which appear in both categories are bolded in Table 4, indicating possible higher demands of them for FTT assignments.

It is noteworthy that over 80% of the respondents (82%, n = 266) said that they usually work alone in FTT assignments, with only 32 respondents (10%) who said they normally work as part of a team. Only eight text answers were further provided by those in the latter category, of which five said they work as part of a "critical team," or part of an "operation," or with police officers, pointing to possibly work related to the investigative stage of cases. Only one mention in these eight text answers relates to "team translation," pointing to the rare practice of engaging multiple T&I practitioners to check each other's work to ensure highest possible quality for transcription and translation. This reality is further corroborated by 83% of the respondents (n = 270) of Question 8 who said they were either never or rarely asked to check other practitioners' translation of forensic recordings.

5.2.2. Audio quality

Respondents predominantly described the audio quality of recordings for their FTT assignments as inconsistent, that is, sometimes good and sometimes bad (56%, n = 184), with almost equal proportions saying it is "normally good" (24%, n = 78) as opposed to "normally bad" (20%, n = 66). This points to the possibility that T&I practitioners work with both telephone intercepts, which normally have better audio quality, as well as covert recordings obtained through clandestine operations, which often feature extremely poor audio quality. The impact of bad audio quality is made clear in this participant's response: "Many a time, I had to listen to a section of the recording more than ten times. I was always worried about losing productivity in my attempts to create

[1] For more on different levels of interpreter certification and their corresponding levels of knowledge and competencies, refer to NAATI webpage https://www.naati.com.au/information-guides/descriptors-for-interpreting/.

[2] For more on different levels of translator certification and their corresponding levels of knowledge and competencies, refer to NAATI webpage https://www.naati.com.au/information-guides/descriptors-for-translating/.

TABLE 3 Summary of respondent demography.

<table>
<tr><th colspan="12">Total respondents ($N = 356$) covering 49 LOTEs</th></tr>
<tr><td colspan="12">FTT training</td></tr>
<tr><td colspan="4">No training</td><td colspan="8">97%</td></tr>
<tr><td colspan="4">Had training</td><td colspan="8">3%</td></tr>
<tr><td colspan="6">Interpreter ($n = 328$)</td><td colspan="6">Translator ($n = 192$)</td></tr>
<tr><td>Mand</td><td>Arb</td><td>Viet</td><td>Pers</td><td>Cant</td><td>Other</td><td>Mand</td><td>Arb</td><td>Viet</td><td>Pers</td><td>Spa</td><td>Other</td></tr>
<tr><td>55</td><td>46</td><td>37</td><td>32</td><td>22</td><td>136</td><td>57</td><td>31</td><td>21</td><td>14</td><td>12</td><td>37</td></tr>
<tr><td colspan="12">T&I credential</td></tr>
<tr><td colspan="4">Conference interpreter</td><td colspan="2">1%</td><td colspan="5">Certified advanced translators</td><td>4%</td></tr>
<tr><td colspan="4">Certified interpreter</td><td colspan="2">62%</td><td colspan="5">Certified translator</td><td>77%</td></tr>
<tr><td colspan="4">Certified provisional interpreter</td><td colspan="2">31%</td><td colspan="5"></td><td></td></tr>
<tr><td colspan="4">Recognized interpreter</td><td colspan="2">2%</td><td colspan="5">Recognized translator</td><td>4%</td></tr>
<tr><td colspan="4">No credential</td><td colspan="2">4%</td><td colspan="5">No credential</td><td>15%</td></tr>
<tr><td colspan="12">Years of practice</td></tr>
<tr><td colspan="4">1–3 yrs experience</td><td colspan="2">14%</td><td colspan="6">13%</td></tr>
<tr><td colspan="4">4–6 yrs experience</td><td colspan="2">18%</td><td colspan="6">10%</td></tr>
<tr><td colspan="4">7–10 yrs experience</td><td colspan="2">14%</td><td colspan="6">13%</td></tr>
<tr><td colspan="4">10+ yrs experience</td><td colspan="2">54%</td><td colspan="6">64%</td></tr>
<tr><td colspan="12">T&I education</td></tr>
<tr><td colspan="4">Postgraduate</td><td colspan="2">26%</td><td colspan="6">33%</td></tr>
<tr><td colspan="4">Bachelor</td><td colspan="2"></td><td colspan="6">12%</td></tr>
<tr><td colspan="4">Advanced Diploma (vocational)</td><td colspan="2">22%</td><td colspan="6">11%</td></tr>
<tr><td colspan="4">Diploma (vocational)</td><td colspan="2">16%</td><td colspan="6"></td></tr>
<tr><td colspan="4">No T&I education</td><td colspan="2">12%</td><td colspan="6">16%</td></tr>
<tr><td colspan="4">Other, e.g., short courses</td><td colspan="2">24%</td><td colspan="6">28%</td></tr>
</table>

TABLE 4 FTT work frequency and corresponding languages.

Frequency	Number of respondents	Percentage	Language counts in descending order
More than once per month	19	5.85%	**Arabic (5), Mandarin (5)** Khmer (3) Assyrian (1), **Burmese (1), Greek (1), Vietnamese (1)** Other unspecified (2)
Between 1 – 12 times per year	117	36.00%	**Mandarin (20)** **Arabic (16)** **Vietnamese (14)** Persian (11) Cantonese (7) **Greek (6)** Spanish (5) Bosnian/Croatian/Serbian (3), Italian (3), Punjabi (3), Turkish (3) Tamil (2) Bengali, (1), **Burmese (1)**, French (1), Hindi (1), Korean (1), Polish (1), Tagalog (1) Other unspecified (17)
Less than once a year	189	58.15%	
Total	325	100%	

excellent quality output." Overlapping talk in these environmental recordings was also commented on by respondents as increasing the challenging nature of FTT.

5.2.3. Briefing on assignments and provision of FTT protocols

When it comes to FTT work practices, a little over one third of the respondents (36%, $n = 117$) said that they did not usually receive a briefing about the case by the police officer in charge before they started translating the relevant forensic recording, while another roughly third (36%, $n = 117$) said that they were usually given a briefing. Information provided to them in the briefing was various, such as the nature and/or background of the case (e.g., "drug trafficking"), location, how the recording was done, people involved in the case, or, more bluntly, "how to pick up criminal activities" and "look for threatening evidence". The remaining 28% ($n = 92$) had mixed experiences that is, a briefing was not consistently received. Further text responses from respondents in this group included: "sometimes [a briefing is provided], if it is not classified"; "sometimes … [I may have to] start straightaway, especially when I do phone call interpreting"; "I can access the warrant to obtain a full picture"; "vague info, e.g., 'this is our crook, and he's calling his business partner'"; "number of people having conversation and mixing the languages"; and "a drug trafficking matter so that I need to understand some code words." Respondents were further probed on whether they were usually briefed by the investigator about how to approach the translation of the LOTE utterances into English, to which two thirds of the respondents said no (66%, $n = 213$), for example, "they just say, there it is. Go ahead. Do your best. Talk to me about progress/problems". Only 17% ($n = 56$) said yes, where the instructions they received included: "how to replay the recordings"; given "keywords" to look for in the recording; being advised to "type all the phone conversation in English"; "they want F and M for gender, or names if known;" "format [to use] and type of notes [to be inserted];" to produce "full [transcript] vs. 'interesting part' only"; and being told "don't guess," or "if … not clear or fragmented, leave them as they are." The remaining 17% represented a mixture of experience by the respondents where instructions were not consistently received on how to produce the translation. The text answers revealed that "it depends on the nature of the assignment"; "they only require a summary in English"; "they advise, e.g., only focus on relevant parts, do a summary, [or] do a full translation etc."; "told to do it verbatim"; "not much info. I feel I'm making up the rules as I go sometimes." One participant offered invaluable insights about working for a particular law enforcement agency:

> The syntax of my languages is different from that of English. Therefore, it's essential to listen to the full sentence before I can start translation. Sometimes, it may become helpful to a reader if I add the intended pronoun. For example, in my languages, a person would simply say, "How is/are?" This may be translated as "How are [you]?" or "How is/are [he/she/they]?" Unfortunately, the [name of agency] requires us to obtain special permission from a supervisor to write anything in square brackets, and generally, such permission is not granted! In my languages, there's only one pronoun for he and she. Sometimes, this creates a problem of gender recognition! Moreover, there are three different types of "you"—informal, formal, and honorable. The only punctuation marks that the [name of agency] allows are the full stop and a question mark! The use of an exclamation mark is discouraged.

5.2.4. Formatting instructions and transcribe-translate two-step process

In terms of formatting the translation for the forensic recordings, roughly two thirds of the respondents (64%, $n = 205$) said there were no instructions or guidance, with 19% ($n = 62$) saying they were advised about the required or preferred formatting. When probing what formatting instructions the respondents were given, they included templates or proformas provided in electronic formats by investigators, being told to follow a format that "should be admissible in court"; "put the accused and the other party in separate columns"; to "bold the words spoken in English by the individuals recorded"; requests to "identify who is speaking, e.g., speaker 1, speaker 2 etc."; and to include time stamps. The majority of the respondents (72%, $n = 236$) translated the audio in a foreign language directly into English, with only 20% ($n = 65$) saying they first transcribed the foreign language in the forensic recordings before translating it into English. The remaining 9% ($n = 28$) of respondents reported a mix of the two practices. From the additional text the respondents entered, it is interesting to note that four respondents explicitly said that they wrote the words down first to enable a better translation into English, suggesting the utilitarian focus of this step for their translation process, rather than from the point of view of providing a traceable record for legal processes. A further three respondents said that they sometimes undertake this two-step transcribe–translate process, as exemplified by this response: "It depends on what the client wants. Sometimes I transcribe in the source language then translate and give them both copies or just translate directly."

5.2.5. Speaker profiling

The respondents were also asked whether they had been asked to "profile" the speaker, i.e., "to give an opinion about what dialect they speak, or what region they might be from" as was explained in the question to ensure understanding. Most respondents either never (60%, $n = 195$) or rarely (14%, $n = 45$) found themselves in the situation, with only 4% ($n = 13$) saying they were asked all the time. The remaining 22% ($n = 71$) answered "sometimes". Regardless of the answer they chose, respondents' written responses indicated that they were mostly asked to comment on the accent (e.g., north or south) and the variety of the dialect heard on the recording; which country (if the language is spoken broadly) or region the speaker was from; or what tribe/ethnicity, education level, or social status could be ascribed to the speaker. The following response illustrates the complexities encountered by practitioners when faced with such requests:

> My language is Albanian. Albanian is spoken in the country Albania and also in Kosovo, where 99% of the population is Albanian. Albanian is also spoken in part of Macedonia, Greece and parts of Italy, where there is a large Albanian population. There are many dialects, and the times when I am asked about "profiling" the speakers, is when the criminals claim that they come from a certain region, but their dialect is from another region. It's a very complicated issue with Albanians.

On two occasions the text answers suggested that the practitioner was asked to discern which language was being spoken in the recording, for example, whether it was Russian or Ukrainian. No suggestion was made in the question as to whether practitioners should or should not respond to such requests, however, one participant wrote "[I told them] that they should get an opinion from a proper linguist or anthropologist who do [*sic*] have knowledge about the Indonesian dialects and accents," and another participant remarked that such practice "can be fraught with danger/traps. Best not to jump to conclusions." Similarly, another participant said "I try not to respond since such a response can be very subjective and may prejudice the case. I can usually tell the general region of the speaker but prefer not to be dogmatic."

5.2.6. Voice identification

As a relevant question, respondents were also asked whether they had been asked to "identify" the speakers in a forensic recording, with further explanation in the question to ensure understanding: "that is, to say who they are by comparing their voices to other voices either within the same recording, or in a separate recording." Similarly, more than eight out of every ten respondents said they were either never (77%, $n = 251$) or rarely (6%, $n = 19$) asked to perform such a task, while the remaining respondents sometimes (13%, $n = 43$) or always did so (3%, $n = 11$). Three text answers entered by respondents explicitly expressed that "I deny [*sic*] to do that giving the reason that I am not a voice expert", or similar reasons. Only one text response explicitly embraced the task by saying "voice recognition is an important part of our work." Surprisingly, one of the respondents who answered that they "rarely" performed such a task said in the text answer: "But we compare handwritten documents," pointing to a risky and unprecedented request for T&I practitioners to act as forensic experts in comparing handwriting supposedly written in a foreign language.

5.2.7. Confidence level and time given to perform FTT tasks

When asked to rate their confidence in their FTT performance on a slider scale from 0 to 4 (0 = *not confident at all*, 1 = *somewhat confident*, 2 = *moderately confident*, 3 = *very confident*, and 4 = *highly confident*), the 256 respondents who answered this question returned a mean score of 2.69 ($SD = 0.97$), that is, between a moderate and very confident level of self-assessed performance.

To further understand the practitioners' FTT experience, they were asked if they were, on average, given the time, information, and resources they needed to do an excellent job in translating forensic recordings. More than seven out of every ten respondents said either "all the time" (32%, $n = 105$) or "sometimes" (41%, $n = 132$), leaving a minority who said "rarely" (16%, $n = 51$) and "never" (11%, $n = 35$). However, the text entered by one participant who answered "never" is concerning: "I have never been given a recording of adequate quality to transcribe or to translate, nor the background or context of the case which would enable me to understand the situation well enough to translate accurately". Another participant who answered "rarely" was more understanding: "I think the police is trying to do their job as good [*sic*] as they can so I don't blame them." Of the 22 text answers further provided by respondents who answered this question, the major themes are: time constraints for translation output impact on translation quality (5 mentions); poor quality of audio hampered the translation quality (4 mentions); and the lack of case related contextual information impedes the deciphering of the interaction on the recording (2 mentions). One text answer was particularly illuminating regarding the different capacities to rewind and re-play audios generated by different recording devices by law enforcement, pointing to possible limitations they have on FTT outcomes:

> I find that the system used by the Federal Police, for example, allows you to slow down or speed up the recording and go back a few seconds and this is good when you need to re-listen to a particular part. However, recordings from listening devices use a different system that does not allow you to easily repeat a particular sentence and is very time-consuming.

5.3. Translation of text messages

Although not strictly in the realms of FTT as it does not involve transcription of recordings, respondents were asked if they had been engaged to translate text messages into English in forensic contexts, given the rising popularity of this means of communication. Over half (52%, $n = 168$) of the 326 respondents answered "yes." Of the 112 text answers further entered by these respondents, there were 39 mentions of the task being "straightforward," while 51 text answers related to the difficulties of the task. These challenges can be categorized into three broad groups:

1. Non-Latin-based languages using English alphabets in the text messaging without tone marks or diacritics, making it extremely hard to decipher meaning. The languages mentioned include Arabic, Chinese, Persian, and Vietnamese. As the following text answer explains: "Because there are tone marks in my language which are often missing in the text messages, the translator has to guess the meaning of the text which is sometimes not correct. The same spelling without tone marks has [several possible] different meanings."
2. Use of slang/street lingo, sociolect/dialect/non-standard language, idiosyncratic language, abbreviations, coded words, emojis, swear words, ambiguous language, incomplete sentences, typographical errors, bad grammar, lack of punctuation. The following participant's response illustrates such challenges:

> The issue is social media posts are often confusing, so you have to spend time to analyze the poster's language by scrolling through their previous posts to understand their language use … the Indonesian people are the king of abbreviations, they can come up with many different variations of non-standard acronyms or abbreviations. Those things might lead you to a completely different understanding.

3. Lack of context about the communication and lack of knowledge about the relationship between texters. As one participant put it: "The one big issue is in spoken Arabic. One statement can mean something and the exact opposite, for example the Arabic meaning of 'you are kind' can mean both 'kind' and 'mean' depending on the context."

5.4. Testify in court

Respondents were also asked if they were ever required to appear in court to answer questions from the prosecution and/or the defense about their translation of forensic recordings. Only 13% of the respondents ($n = 42$) said they were "sometimes," with a further 1% ($n = 3$) saying they were asked "all the time." The majority of respondents were either never (78%, $n = 255$) or rarely (9%, $n = 29$) required to do so. Of the 329 respondents who answered this question, 54 provided further text answers, ten of whom mentioned they had been subpoenaed but never had to testify in court either because they were eventually not called, the cases were settled before hearing, or the defense pleaded guilty. One of these ten respondents stated, "but I feel very nervous about the prospect of it." For those who did appear in court, they were most often questioned about the accuracy of their translation and asked to explain or justify their choice of words, as was described by one participant: "Why this specific meaning of the word(s) has been used [but] not other meanings of the word, when the word has many meaning". Similarly, another respondent stated that they were "queried on alternative possible interpretations," while an observation was made by a further respondent: "Sometimes defense wants to use words of less impact." The following response comprehensively summed up the challenging nature of FTT work and the prospect of having to swear in court on the accuracy of work which is generated from indistinct covert recordings containing information that is inherently hard to decipher:

> It is hard to transcribe without context, and we often don't have enough context to make full sense of what is being said. For example, who are the speakers, their relationship, how many there are, etc. When you work on a case for a longer period, you start to learn more context from other recordings, but then that info can affect what you hear, or think you hear, on future recordings. It's a very difficult job and the idea of having to be cross examined on my work, particularly the decision about whether I am sure enough about what I heard to swear it in front of a court and therefore include it, or not include it in the transcription... well it's challenging!

In a similar vein, another participant described the dilemma of whether to commit to what they think they hear or to play it safe by stating the segment is *indistinct*, in case they must appear in court to defend the transcript/translation:

> Very often the voice of the person whose phone is being intercepted is clear, but the interlocutor's voice is distorted. As an interpreter/translator you strain your ear, listen to the same part multiple times to make sure you can understand and translate, but sometimes this is just not possible. Or the lines are simply crackling or there is background noise, etc. When you produce a transcript/translation to be used in court, you need to be sure that it is correct, and very often I cannot be 100% sure that I'm hearing what I think I'm hearing. This is one added responsibility on the interpreter/translator and the dilemma arises as to whether to type what you think is being said or cover your back by typing "[indistinct]" if you know you won't be able to fully and satisfactorily back your choice in court if necessary.

Further, there was an honest revelation that "I do not like going to present myself [in court] as it is scary to be sitting there and the accused person seeing me and thinking I am working against them." A similar statement was made by another participant: "Am I going to see the accused who made the calls [in the recording]? Will that put me in any kind of danger if they see who I am? I think more information should be provided to interpreters in such scenarios to put their mind at ease."

5.5. Practitioner concerns

In closing the survey, respondents were given an opportunity to enter any free text they wish to comment on any aspect of FTT. Of the 72 text answers entered, the themes about poor audio quality and the lack of contextual information for cases were again dominant. Practitioners' concerns about the impacts of these limitations on their performance were palpable. In addition, a number of rare insights emerged which are categorized into the following three themes:

5.5.1. Working conditions and remuneration

The reality of the work is such that practitioners are mostly required to attend law enforcement offices in person due to data security and operational concerns. However, as one participant explains:

> When you work onsite at police premises, you may not have the benefit of having a little chat here and there, stretching your legs, etc. so you end up doing many straight hours looking at the computer, straining your ear, without a break, sometimes surrounded by people you haven't met before, and even a toilet break is stressful when you have to ask someone to escort you, unlock doors, etc. and then someone has to come and log you into the system again, etc.

On rare occasions when practitioners were allowed to work offsite, it was not ideal either, as the following response illustrates:

> I have the comfort of working at my own place; however, I lose the opportunity of having the agent/officer in charge at hand to ask any questions or to discuss any aspects that may arise, and the end result may be affected. Additionally, the audio software I use at home does not allow me to go back to an exact position in the recording or to slow it down to get more clarity.

A number of respondents commented on how they had to work under time pressure and the highly "complex," "demanding," "exhausting," and "draining" nature of FTT tasks. One participant remarked that the task may look easy, but in fact is very hard and requires a lot of focus, highlighting the importance of incorporating linguistic as well as emotional elements in the transcripts. Other respondents raised the issue of remuneration: "Current pay offered does not compensate the effort and time put into providing a proper and best possible English version of forensic recordings"; language service agencies which deploy practitioners to such assignments "usually want the job done in a short time and pay minimum fees without considering quality and time required"; and "some agencies pay only the interpreting rate even for transcription." A specific law enforcement agency was singled out by a participant as having tendered out FTT work to many language service agencies, and therefore every time a quote was requested for an assignment, numerous agencies competed for business basically on price, which "results in interpreters doing a taxing, complex job full of responsibility that is not commensurate with the pay."

5.5.2. Need for translation guidance and standardized work protocols

Respondents strongly conveyed their views on the lack of translation guidance and work protocols for FTT assignments, which was also reflected in Section 5.2.2 above. One participant made the insightful observation that a translation accompanying the forensic audio evidence "by its nature [already] disrupts the evidence." This participant went on to say that "police (and the judiciary) rarely understand this. When doing forensic transcription sometimes police cannot appreciate the complexities and implications for the evidence that the transcription constitutes. This should be of concern, understood and managed." Another participant captured the dilemma well by asking the following question: when faced with ambiguity of meaning in forensic recordings or text extracts with little contextual information provided "should translators ask the professional for more context and discuss about word choices, or should translators offer all the likely possibilities in the translation for the judge/jury to decide which meaning it should be?"

In the absence of any explicit translation guidelines offered by law enforcement who require FTT services, it is also not known how practitioners deal with coded words and whether their neutrality is maintained. One participant stated that "once I transcribed a tape recording for drug trafficking. They mentioned red and white buttons hidden under the bed. I would not interpret what they were but just translated as it was." Similarly, another participant also clearly articulated that for slang or coded words such as "a hit" or other drug terms, "these terms should be translated as they are. It is up to the law enforcer to work out what they mean and not the interpreter's job to conjecture." There is also a comment which concurred with the two-step transcribe—translate process: "Transcripts are essential when doing this job. If the client is using several translators and comparing their translations, a transcript makes sure we all have the same primary source. Without a transcript this is a futile exercise."

5.5.3. Need for specialist training and to define required competencies

Another strong theme emerging from the last free-text question in the survey is about the lack of specialist training nor clear definition of the competencies required for FTT. One comment remarked on the infrequent nature of the FTT assignments, and thus the need for the practitioner to "refresh, re-familiarize with equipment, program, find best work methods each time ... [which] can be difficult and challenging to work efficiently and quickly to produce an excellent result. Training sessions would be extremely valuable." Another comment suggested that "formal training as part of an advanced diploma or master's or as a separate long PD [professional development] should be offered."

Practitioners rightly asked the question about who should perform FTT tasks and what credential should be required, for example, "I am not sure if interpreters are qualified to do transcription. Is transcription a translation? If yes, only certified translators should do it"; and "if I don't have the credential of LOTE into English, should I refuse the request of forensic translation when I serve as an interpreter?" These queries culminated in the following participant's comment: "I believe practitioners need to have both certifications in translation and interpreting in order to carry out this kind of forensic work." Relevant to this, another participant suggested that NAATI should "test and award credentials for this area specifically, since I'm not sure that our current qualifications are applicable to the role."

6. Discussion

This study has brought to light the current state of service provision for FTT by T&I practitioners in Australia by pursuing three enquiries: *who* does it, *how* they do it, and *what* they think about it. The landscape of this under-explored area has been mapped for the first time through the findings reported above.

6.1. Who does it

We have come to understand that a mixture of practitioners who are either interpreters, translators, or both were variously engaged for FTT assignments. Although large proportions of them had credentials awarded by NAATI, had some T&I education, and were relatively experienced practitioners, very few of them had any FTT specific training, which is currently not widely available, if at all. The two-step process recommended by best practice FTT (see discussion in 3.1) points to two areas of specialist training required: transcription (from spoken LOTE to written LOTE), and translation (from LOTE into English). The need for training for the former is no different from monolingual settings, which has been advocated by scholars (Fraser and Stevenson, 2014; Romito, 2017; Fraser and Loakes, 2020; Fraser, 2021, 2022) in order to achieve accuracy and reliability in forensic contexts. The current study reveals the fact that the majority of the respondents undertook very infrequent FTT assignments, and unlike other areas such as community interpreting for healthcare, education, or social services, FTT does not constitute their bread and butter. On the one hand, it hampers developing expertise in this line of work as was reported in Section 5.5.3. However, this also makes it possible to focus on the higher-demand languages and consider prioritizing them for targeted training to start cultivating expertise in this specialist branch under legal interpreting and translation which has so far been neglected. This should improve the status quo where only 3% of the respondents had ever received relevant FTT training. If we disregard the row showing the lowest work frequency in Table 4, (i.e., those who did FTT assignments less than once per annum), Arabic and Mandarin no doubt feature most prominently in the other two categories of higher frequency, pointing to the possibility of recruiting selected practitioners from these two languages as the candidates for targeted training. Languages such as Burmese, Greek, and Vietnamese which appeared in both categories, may be considered when training can be expanded for larger language coverage. However, further triangulation of data on high-demand languages from law enforcement and language service agencies will be desirable to confirm if these languages reflect their demand profiles, or whether adjustments to add or take out certain languages are necessary. This is because the work frequency probed in the survey was self-reported, and there was no definition given as to what constitutes an assignment. For example, whether respondents regarded a long case spanning many weeks of FTT work at a law enforcement office as one assignment or several days of single assignments, is unknown, and therefore some languages of high demand might be missed or appear to rank lower in this study, or vice versa.

Although the respondents' average confidence level of their self-assessed FTT performance was between moderate to very confident, one may posit that the lack of training could manifest as a false sense of confidence and an ignorance of risks. Those who expressed unease about performing FTT when they are not credentialed translators from LOTE into English were right to question the probity. Interpreters are language professionals who specialize in listening to spoken discourse and converting it into spoken discourse in the TL (i.e., column 1 in Table 5 below), while translators specialize in reading written discourse and converting it into written discourse in the TL (i.e., column 2 in Table 5). Interlingual transcribers, however, listen to spoken discourse in the SL, but produce written discourse in the TL. This is why NAJIT (2019) position paper points out the deficiencies of either interpreters or translators undertaking FTT tasks. Mapping the hybrid set of competencies required for FTT and mandate that such tasks be performed by those who possess both T&I credentials should be the future direction to ensure quality output.

TABLE 5 FTT as a hybrid T&I task.

1. Interpreter	2. Translator	3. Interlingual transcriber
Listens to spoken discourse in SL	Reads written discourse in SL	Listens to spoken discourse in SL
Re-expresses the spoken discourse into TL	Re-expresses the written discourse into TL	Re-expresses in written discourse into TL

As a relevant issue to the enquiry of *who* does FTT, in addition to the concerns discussed so far about the lack of specialized training nor clarity on the required competencies, another concern is that some practitioners were asked to "profile" or to "identify" speakers in forensic recordings. A practitioner may be knowledgeable in the varieties of their LOTE, relevant accents, and their associated geographical differences; however, it is a dangerous practice to rely on an unverified non-expert to supply such information without any checking mechanisms. In relation to identifying speakers in recordings, it is understandable that in the same recording, it is necessary for the practitioner to discern different speakers and assign labels such as MV1 (Male Voice 1) or FV1 (Female Voice 1). The task by itself is challenging, as voice distortions often found in intercepted phone calls are not conducive to accuracy in identifying same speakers in a talking sequence in the same recording. It is even more challenging to ask practitioners to identify whether a certain voice belongs to the same person in different recordings. Without specialist training and stringent quality procedures, practitioners' contributions will be conjectural and unreliable. Law enforcement should refrain from soliciting such input from T&I practitioners undertaking FTT tasks, as the latter may feel pressured to respond to the request, while lacking the skills and competence to do so. Further, the finding about practitioners being asked to compare handwriting in a foreign language is even more concerning, as it is positively beyond T&I practitioners' field of expertise. If law enforcement relies on the practitioner for speaker profiling, voice identification, or even handwriting comparison, when such evidence is tendered in court and doubts are raised by defense, it will not qualify as expert opinion, which is exempted from the general rule that opinion evidence is inadmissible. For example, the state of New South Wales, Australia, Section 79(1) Evidence Act (NSW) defines an expert as a person who has specialized knowledge, based on the person's training, study, or experience. In this case, the practitioner would have failed on all three accounts rendering the evidence inadmissible.

6.2. How they do it

The current study shows conclusively the need for a set of protocols to govern:

- the competencies required to undertake FTT (i.e., ideally practitioners with both T&I credentials, and specifically from LOTE into English for the former)
- the production of FTT (i.e., a two-step process to ensure audit trail, when team translation or peer checking is required)
- the provision of case briefs (i.e., whether the nature of the assignment is investigative or evidentiary, when to introduce case information and how much information)

- the format of FTT (i.e., ideally the four-column presentation as recommended in Table 2)
- transcription conventions (i.e., uniform set of transcription symbols such are in Appendix 1), level of linguistic/paralinguistic/extralinguistic details required, and threshold for confidence level (i.e. how "sure" is sure enough to commit an indistinct utterance to words assigned to them in the transcript)
- the translation approach (i.e., how to represent uncertain meanings in uncertain contexts, when and how to provide translator's notes).

The protocols are needed in light of the fact that more than 70% of the respondents in the survey translated forensic recordings from a LOTE directly into English, as no requirement exists to mandate the production of a LOTE transcript before a translation is produced, thus losing the *audit trail* (Gilbert and Heydon, 2021). The current popular practice may create issues in trials if anyone challenges the translation, as there is no way to ascertain what was heard in the recording and how the spoken utterances were converted into the English translation. Further, <40% (36%) of the respondents usually received a briefing for their FTT tasks, with roughly only one in every five (19%) reported to have been given the instructions on formatting and style of the English translation. As forensic recordings are often of poor quality, and practitioners are entrusted with the unenviable task of deciphering communications devoid of context, it is important that a certain level of briefing (bearing in mind the priming effect a briefing might have on the practitioner, therefore the consideration of the timing and extent of a briefing) and a set of work instructions be in place for practitioners to follow, for example, whether and when they should take a more literal approach for words that do not seem congruent with the utterances, therefore the possibility of coded words. In this regard, close to 70% (66%) reported that they were not given instructions as to how to approach the translation work. The text contributions also point to a mixture of investigative and evidentiary FTT tasks the respondents were involved in, where practitioners were sometimes given key words by investigators to look for in the forensic recording, or told to scour for "interesting" information, or to produce a "summary"; whereas on other occasions they were told "don't guess" and to leave unclear or fragmented parts as they were. It is important for the practitioners to understand the nature of their engagement and the different criteria for translation exactitude and the extent of "interpretation" of meanings for investigative vs evidentiary purposes. The survey also confirms that few respondents worked in teams with other fellow practitioners or were asked to check others' work. Although it may not be practical to employ a team approach or have every translation peer reviewed, it should be reasonable to consider such an approach for major cases to ensure rigor in the translation tendered to courts. Anecdotally the current author is aware that some seasoned practitioners find it difficult to work in teams or to check other's work, because, in the absence of uniformed guidelines, everyone approaches FTT in their own way and it is not easy to justify to colleagues why your approach in a particular instance is more appropriate. This reinforces the importance of specialist training and work protocols discussed above, so when a team approach is adopted, everyone is on the same page and be able to work collaboratively.

Another implication of the lack of protocols to guide the approaches in their FTT work relates to the prospect of being subpoenaed to appear in court. Apart from the anxiety in the respondents about "outing" themselves in front of the accused, they were apprehensive that their translations might be challenged—on either what they hear (in the recording), or how they translate, or both. There is a real dilemma for the practitioner to "play safe" by resorting to saying *indistinct* whenever they have the slightest doubt about what they can or cannot hear, which may render their work of little use, or to try their best to discern the unclear LOTE utterances and stand by them after listening many times, which is no different from investigators transcribing indistinct audios containing exchanges spoken in English as ad-hoc experts without training. The practitioners' anxiety will be better managed if they have specialist training on transcribing indistinct audios which is currently lacking, and they are provided with guidelines on evaluating their confidence level of what they hear and when to commit what they hear to the transcript.

As acknowledged before, translating text messages in forensic contexts is not strictly FTT. However, it does closely related to the (forensic) translation part of FT"T". More than half of the respondents had been involved in translating text messages from foreign languages into English for forensic purposes, and the challenges they encountered concurred with those asserted by Cal-Meyer (2016) and O'Hagan and Ashworth (2002) (see Section 3.7). In the absence of any specialist training, practitioners have no choice but to do it the way they "think" is right. Given the growing popularity of CMC, needs for forensic translation of text messages will no doubt grow and extend to online chats, emails, and social media posts. Sensitizing translators on the linguistic, pragmatic, and cognitive features of these genres is becoming critical. Taking the lesson from R v Yang [2016] WASC 410 discussed in Section 3.7, translators must be acutely aware of their role boundaries by faithfully representing the *tenor* of the discourse in the target text—that is *how* something is said in addition to *what* is said (e.g., reproduce all emoticons or expressives in totality, annotate typographical errors that are in the source language), while refraining from expressing opinions drawn from the source text. Organizing professional development on this topic area may be a good starting point to address the training needs.

6.3. What they think about it

Lastly, the three major concerns expressed by the respondents echoed the discussions above about the lack of specialist training and work guidelines and protocols. Practitioners tried their best to respond to requests by law enforcement for whom they perform the FTT tasks, but they were not confident about whether they were doing the right thing in producing the best quality and highest reliability possible. The survey does not provide evidence of clear awareness in practitioners about the difference between investigative and evidentiary FTT tasks, nor is it able to confirm practitioners' commitment to neutrality as an independent professional, which is what T&I practitioners should abide by in Australia. What should also be acknowledged is that no amount of training on the part of T&I practitioners will address the issues identified in the current study if the status quo continues on the law enforcement side which is the main consumer of FTT services, as one participant noted: "law enforcement officers are hardly ever trained on how to work with interpreters and translators." Inadequate remuneration for community interpreters and translators has long been argued by T&I practitioners and the profession as a whole, which has serious implications for the sustainability of the industry and retention of quality practitioners. As is argued in Section 6.1, there is potential to focus specialist FTT training on a selected range of prioritized languages as an achievable starting point. It is perhaps more feasible to start by adequately remunerating those who have been through specialist training and, ideally, possess a national specialist credential which should be introduced in the future. Mapping the hybrid set of competencies for FTT as was argued in 6.1 will help formulating the new national specialist credential and designing specialist training in education programs. It will then be possible for other languages less frequently required for FTT tasks to work under these specialists' supervision, an idea similar to the *master-level* FTT specialist proposed by González et al. (2012).

7. Concluding remarks

This paper would be incomplete without acknowledging the limitations of the current study. The answers to the questionnaire were self-reported by the respondents, and therefore gathering data in further studies from law enforcement and language service agencies for triangulation is desirable to form a more holistic understanding of the FTT landscape and service provision in Australia. Although it is helpful to have a relatively large number of respondents in the current study, it should be borne in mind that significantly more respondents only undertook FTT tasks very infrequently, and therefore what is learned from the survey may be snapshots of distant experiences. Future studies should attempt more focused purposive sampling to recruit practitioners in languages of major demand to collect further insights into their experiences. This will be beneficial in informing possible future training design and collaborative practice with law enforcement.

To sum up, this study offers insights into the FTT landscape in Australia in terms of (1) the profile of the T&I practitioners who undertook FTT assignments; (2) their work practices and experiences interacting with law enforcement; and (3) their reflections and thoughts about this line of work. It reveals the mismatch between the level of competence required by FTT to serve the ultimate purpose of justice and the work practices law enforcement facilitates FTT. Similar to the more abundant scholarship in the space of forensic transcription in monolingual settings, this study echoes the position that practitioners engaged in FTT should have training in transcription to ensure quality and reliability. Additionally, the nature of FTT is such that it is not only intermodal (i.e., from audio input to written output) for the part of transcription, but also intersystemic (i.e., interlingual) for the part of translation. It is not possible for a trained monolingual transcriber to undertake FTT, since the person lacks proficiency in the source language (i.e., LOTE). The only feasible way is for T&I professionals to receive transcription training. On top of that, they must be attuned to the discourse features of covert recordings and intercepted private messaging, and understand the criteria for translation exactitude and the extent of "interpretation" of meanings appropriate to the forensic context under which their FTT service is required. The current T&I training is lagging behind these needs and available scholarship lacks applicable models for this branch of forensic translation. In line with the call for forensic transcription in monolingual settings to be treated as a branch of linguistic science (Fraser, 2021; Love and Wright, 2021), this study demonstrates similar support for specialization in the T&I studies, as well as the urgency to develop scholarship to guide and inform best practice for FTT.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the author, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by College Human Ethics Advisory Network, RMIT Human Research Ethics Committee. The participants provided their informed consent by opting in this study.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Acknowledgments

The author sincerely thanks the interpreters and translators who took part in this survey and shared their valuable insights into this under-researched topic. Gratitude is owed to Professor Helen Fraser who designed the questionnaire and provided guidance on the study.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2023.1096639/full#supplementary-material

References

Ashworth, D. (1997). *Transterpreting: A New Modality for Interpreting on the Internet Pan-Pacific Distance Learning Association Conference*. Honolulu, Hawaii.

AUSIT (2012). *AUSIT Code of ethics and code of conduct.* Australian Institute of Translators and Interpreters. Available online at: http://ausit.org/ausit/documents/code_of_ethics_full.pdf (accessed September 2, 2022).

Auslii (2017). *R v Yang [2016] WASC 410.* Available online at: http://www.austlii.edu.au/cgi-bin/viewdoc/au/cases/wa/WASC/2016/410.html (accessed October 1, 2022).

Bellucci, P. (2022). *A onor del vero: fondamenti di linguistica giudiziaria.* Torino: Utet.

Braun, V., and Clarke, V. (2006). Using thematic analysis in psychology. *Qual. Res. Psychol.* 3, 77–101. doi: 10.1191/1478088706qp063oa

Bucholtz, M. (2007). Variation in transcription. *Disc. Stud.* 9, 784–808. doi: 10.1177/1461445607082580

Butterfield, A., Ngondi, G. E., and Kerr, A. (2016). *A Dictionary of Computer Science.* Oxford: Oxford University Press. doi: 10.1093/acref/9780199688975.001.0001

Cal-Meyer, P. (2016). Decoding and constructing meaning from text messages: Pragmatic considerations for court interpreters. *Summer Volume*(2). Available online at: https://najit.org/proteus/decoding-constructing-meaning-text-messages-pragmatic-considerations-court-interpreters-summer-2016/ (accessed October 1, 2012).

Capus, N., and Griebel, C. (2021). The (In-)Visibility of interpreters in legal wiretapping—A cases study: how the swiss federal court clears or thickens the fog. *Int. J. Lang Law* 10, 73–98. doi: 10.14762/jll.2021.73

Capus, N., and Grisot, C. (2022). Ghostwriters of crime narratives: Constructing the story by referring to intercept interpreters' contributions in criminal case files. *Crime, Media, Culture.* 5, 304. doi: 10.1177/17416590221133304

Capus, N., and Havelka, I. (2022). Interpreting intercepted communication: a Sui Generis translational activity. *Int. J. Semiot. Law Revue internationale de Sémiotique juridique*, 35, 1817–1836. doi: 10.1007/s11196-021-09876-0

Eco, U. (2001). *Experiences in Translation.* Toronto, ON: University of Toronto Press.

Fishman, C. S. (2006). Recordings, transcripts, and translations as evidence. *Washington Law Rev.* 81, 473–523.

Fraser, H. (2014). Transcription of indistinct forensic recordings: Problems and solutions from the perspective of phonetic science. *Language and Law* 1, 5–21.

Fraser, H. (2017). *Transcription and interpretation of indistinct covert recordings used as evidence in court.* Available online at: https://forensictranscription.com.au/video-of-njca-talk/ (accessed November 25, 2017).

Fraser, H. (2018). 'Assisting' listeners to hear words that aren't there: dangers in using police transcripts of indistinct covert recordings. *Australian Journal of Forensic Sciences*, 50, 129–139. doi: 10.1080/00450618.2017.1340522

Fraser, H. (2021). "Forensic transcription: The case for transcription as a dedicated area of linguistic science," in *The Routledge Handbook of Forensic Linguistics (2nd ed.)*, eds M. Coulthard, A. Johnson, and R. Sousa-Silva (London: Routledge). doi: 10.4324/9780429030581-33

Fraser, H. (2022). A framework for deciding how to create and evaluate transcripts for forensic and other purposes [Systematic Review]. *Front. Commun.* 7, 8410. doi: 10.3389/fcomm.2022.898410

Fraser, H., and Kinoshita, Y. (2021). Injustice arising from the unnoticed power of priming: How lawyers and even judges can be misled by unreliable transcripts of indistinct forensic audio. *Crim. Law J.* 45(3). Available online at: https://search.informit.org/doi/full/10.3316/agispt.20210923053902

Fraser, H., and Loakes, D. (2020). Acoustic injustice: The experience of listening to indistinct covert recordings presented as evidence in court. *Law, Text, Culture* 24, 405–429. Available online at: https://ro.uow.edu.au/ltc/vol24/iss1/16

Fraser, H., and Stevenson, B. (2014). The power and persistence of contextual priming: More risks in using police transcripts to aid jurors' perception of poor quality covert recordings. *Int. J. Evid. Proof* 18, 205–229. doi: 10.1350/ijep.2014.18.3.453

Fraser, H., Stevenson, B., and Marks, T. (2011). Interpretation of a crisis call: persistence of a primed perception of a disputed utterance. *Int. J. Speech Lang. Law* 18, 261–292. doi: 10.1558/ijsll.v18i2.261

French, P., and Harrison, P. (2006). "Investigative and evidential applications of forensic speech science," in *Witness Testimony: Psychological, Investigative and Evidential Perspectives*, eds A. Heaton-Armstrong and E. Shepherd. Oxford University Press.

Gilbert, D. (2014). *Electronic Surveillance and Systemic Deficiencies In Language Capability: Implications for Australia's National Security.* [Doctoral thesis] Melbourne: RMIT University.

Gilbert, D., and Heydon, G. (2021). Translated transcripts from covert recordings used for evidence in court: issues of reliability. *Front. Commun.* 6, 9227. doi: 10.3389/fcomm.2021.779227

González, R. D., Vásquez, V. F., and Mikkelson, H. (2012). *Fundamentals of Court Interpretation: Theory, Policy, and Practice (2nd ed.).* Durham, NC: Carolina Academic Press.

Haworth, K. (2010). "Police interviews in the judicial process: Police interviews as evidence," in *The Routledge Handbook of Forensic Linguistics* eds M. Coulthard and A. Johnson (pp. 169-184). Routledge. doi: 10.4324/9780203855607.ch12

Jakobson, R. (1959). *On linguistic aspects of translation.* Available online at: http://www.stanford.edu/$\sim$eckert/PDF/jakobson.pdf (accessed October 01, 2022).

Jefferson, G. (2004). "Glossary of transcript symbols with an introduction," in *Conversation Analysis: Studies from the First Generation*, ed G. H. Lerner (John Benjamins) (pp. 13-31). doi: 10.1075/pbns.125.02jef

Jenks, C. J. (2013). Working with transcripts: an abridged review of issues in transcription. *Lang. Linguis. Compass* 7, 251–261. doi: 10.1111/lnc3.12023

Kaindl, K. (2012). "Multimodality and traslation," in *The Routledge handbook of translation studies* eds C. Millan and F. Bartrina (pp. 257-269). Routledge.

KordaMentha (2018). *Evidence lost in translation.* Available online at: https://kordamentha.com/news-and-insights/forensic-expert-evidence-lost-in-translation (accessed October 1, 2022).

Kredens, K., Monteoliva- García, E., and Morris, R. (2021). "Interpreting outside the courtroom - 'A shattered mirror?' Interpreting in law enforcement contexts outside the courtroom," in *The Routledge Handbook of Forensic Linguistics*, eds A. May, R. Sousa-Silva, and M. Coulthard (2nd ed., pp. 502-520). Routledge. doi: 10.4324/9780429030581-39

Love, R., and Wright, D. (2021). Specifying challenges in transcribing covert recordings: Implications for forensic transcription. *Front. Commun.* 6, 7448. doi: 10.3389/fcomm.2021.797448

Mikkelson, H. (2016). "Interpreting for law enforcement," in R *Fundamentals of Court Interpretation: Theory, Policy, and Practice*, eds González, V. F. Vasques, and H. Mikkelson (Eds.) (pp. 58-74). Carolina Academic Press.

NAATI (2021). *2020–21 Annual Report.* Available online at: https://www.naati.com.au/wp-content/uploads/2021/11/Annual-Report-2020-2021-1.pdf (accessed July 22, 2022).

NAJIT (2019). Position paper: General guidelines and requirements for transcription translation in a legal setting for users and practitioners. National Association of Judiciary Interpreters and Translators. Available online at: https://najit.org/wp-content/uploads/2016/09/Guidelines-and-Requirements-for-Transcription-Translation.pdf (accessed February 2, 2022).

O'Hagan, M., and Ashworth, D. (2002). Translation-mediated communication in a digital world: facing the challenges of globalization and localization. *Multilin Matt.* 3, 5820. doi: 10.21832/9781853595820

Orletti, F. (2017). "Transcribing intercepted telephone calls and uncovered recordings: An exercise of applied coversation analysis," in *Forensic communication in theory and practice: A study of discourse analysis and transcription*, eds Orletti and L. Mariottini (pp. 21-36). Cambridge Scholars Publishing.

Orletti, F., and Moriottini, L. (2017). Introduction: Forensic communication – From Theory to Practice In F. Orletti and L. Moriottini (Eds.), *Forensic Communication in Theory and Practice: A Study of Discourse Analysis and Transcription* (pp. 1-7). Cambridge Scholars Publishing.

Paoloni, A., and Zavattaro, D. (2007). *Intercettazioni telefoniche e ambientali*. Metodi, limiti e sviluppi nella trascrizione e verbalizzazione. Torino, Italy: Centro scientifico editore.

Patton, M. Q. (2002). *Qualitative Research and Evaluation Methods (3rd ed.)*. Thousand Oaks, CA: SAGE.

Romito, L. (2017). "A training program for expert forensic transcribers," in F. Orletti and L. Moriottini (Eds.), *Forensic communication in theory and practice: A study of discourse analysis and transcription* (pp. 47-62). Cambridge Scholars publishing.

Salaets, H., Alsulaiman, A., and Biesbrouck, S. (2015). Tap interpreting: From practice to norm. A Belgian case study. *Turjuman* 24, 11–49.

Tilley, S. A. (2003). "Challenging" research practices: turning a critical lens on the work oftTranscription. *Qual. Inq.* 9, 750–773. doi: 10.1177/1077800403255296

Toury, G. (1994/[1986]). Translation: A cultural-semiotic perspective. In T. Sebeok (Ed.), Encyclopedic dictionary of semiotics (Vol. 2, pp. 1111-1124). Mouton de Gruyter.

Wainfan, L., and Davis, P. K. (2004). *Challenges in Virtual Collaboration: Videoconferencing, Audioconferencing, and Computer-Mediated Communications*. Santa Monica, CA: Rand Corporation.

TYPE Correction
PUBLISHED 28 February 2024
DOI 10.3389/fcomm.2024.1383240

Corrigendum: Transcribing and translating forensic speech evidence containing foreign languages—An Australian perspective

Miranda Lai*

Translating and Interpreting, Royal Melbourne Institute of Technology (RMIT University), Melbourne, VIC, Australia

OPEN ACCESS

EDITED AND REVIEWED BY
Mila Vulchanova,
Norwegian University of Science and Technology, Norway

*CORRESPONDENCE
Miranda Lai
✉ miranda.lai@rmit.edu.au

RECEIVED 07 February 2024
ACCEPTED 19 February 2024
PUBLISHED 28 February 2024

CITATION
Lai M (2024) Corrigendum: Transcribing and translating forensic speech evidence containing foreign languages—An Australian perspective. *Front. Commun.* 9:1383240. doi: 10.3389/fcomm.2024.1383240

KEYWORDS
forensic transcription, forensic translation, forensic speech evidence, interception/covert recording, legal translation, legal translator, legal interpreter

A corrigendum on
Transcribing and translating forensic speech evidence containing foreign languages—An Australian perspective

by Lai, M. (2023). *Front. Commun.* 8:1096639. doi: 10.3389/fcomm.2023.1096639

In the published article, there was an error. The study that was analyzed in the article was an Australian first survey. Words to this effect appeared in four places. Due to the way one sentence was worded, it may be incorrectly construed that this was the first study on the topic. To avoid any unnecessary confusion or disputes in the future, all four mentions of the study have been amended.

A correction has been made to the Abstract. This sentence previously stated:

"This paper reports on the first ever survey conducted in Australia on professional interpreters and translators who have been engaged to perform this type of work."

The corrected sentence appears below:

"This paper reports on an online survey conducted in Australia on professional interpreters and translators who have been engaged to perform this type of work."

A correction has been made to Introduction, Paragraph 2. This sentence previously stated:

"This paper reports on a first ever study in Australia on the experiences of translators and interpreters involved in forensic transcription and translation (FTT) for law enforcement for both investigative and evidentiary purposes."

The corrected sentence appears below:

"This paper reports on an online survey conducted in Australia on the experiences of translators and interpreters involved in forensic transcription and translation (FTT) for law enforcement for both investigative and evidentiary purposes."

A correction has been made to the Methods, Paragraph 1. This sentence previously stated:

"An Australia-first survey was designed to collect descriptive statistics and qualitative data to answer an overarching research question: what is the current state of service provision for FTT by T&I practitioners in Australia?"

The corrected sentence appears below:

"An online survey was designed to collect descriptive statistics and qualitative data to answer an overarching research question: what is the current state of service provision for FTT by T&I practitioners in Australia?"

A correction has been made to the Concluding remarks, paragraph 2. This sentence previously stated:

"To sum up, this study offers first insights into the FTT landscape in Australia in terms of (1) the profile of the T&I practitioners who undertook FTT assignments; (2) their work practices and experiences interacting with law enforcement; and (3) their reflections and thoughts about this line of work."

The corrected sentence appears below:

"To sum up, this study offers insights into the FTT landscape in Australia in terms of (1) the profile of the T&I practitioners who undertook FTT assignments; (2) their work practices and experiences interacting with law enforcement; and (3) their reflections and thoughts about this line of work."

The authors apologize for these errors and state that they do not change the scientific conclusions of the article in any way. The original article has been updated.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

TYPE Original Research
PUBLISHED 13 July 2023
DOI 10.3389/fcomm.2023.1165233

Check for updates

OPEN ACCESS

EDITED BY
Kate Haworth,
Aston University, United Kingdom

REVIEWED BY
James Tompkinson,
Aston University, United Kingdom
Georgina Brown,
Lancaster University, United Kingdom

*CORRESPONDENCE
Lauren Harrington
✉ lauren.harrington@york.ac.uk

RECEIVED 13 February 2023
ACCEPTED 28 June 2023
PUBLISHED 13 July 2023

CITATION
Harrington L (2023) Incorporating automatic speech recognition methods into the transcription of police-suspect interviews: factors affecting automatic performance. *Front. Commun.* 8:1165233.
doi: 10.3389/fcomm.2023.1165233

Incorporating automatic speech recognition methods into the transcription of police-suspect interviews: factors affecting automatic performance

Lauren Harrington*

Department of Language and Linguistic Science, University of York, York, United Kingdom

Introduction: In England and Wales, transcripts of police-suspect interviews are often admitted as evidence in courts of law. Orthographic transcription is a time-consuming process and is usually carried out by untrained transcribers, resulting in records that contain summaries of large sections of the interview and paraphrased speech. The omission or inaccurate representation of important speech content could have serious consequences in a court of law. It is therefore clear that investigation into better solutions for police-interview transcription is required. This paper explores the possibility of incorporating automatic speech recognition (ASR) methods into the transcription process, with the goal of producing verbatim transcripts without sacrificing police time and money. We consider the potential viability of automatic transcripts as a "first" draft that would be manually corrected by police transcribers. The study additionally investigates the effects of audio quality, regional accent, and the ASR system used, as well as the types and magnitude of errors produced and their implications in the context of police-suspect interview transcripts.

Methods: Speech data was extracted from two forensically-relevant corpora, with speakers of two accents of British English: Standard Southern British English and West Yorkshire English (a non-standard regional variety). Both a high quality and degraded version of each file was transcribed using three commercially available ASR systems: Amazon, Google, and Rev.

Results: System performance varied depending on the ASR system and the audio quality, and while regional accent was not found to significantly predict word error rate, the distribution of errors varied substantially across the accents, with more potentially damaging errors produced for speakers of West Yorkshire English.

Discussion: The low word error rates and easily identifiable errors produced by Amazon suggest that the incorporation of ASR into the transcription of police-suspect interviews could be viable, though more work is required to investigate the effects of other contextual factors, such as multiple speakers and different types of background noise.

KEYWORDS
transcription, automatic speech recognition, forensic linguistics, automatic methods, phonetics

1. Introduction

Orthographic transcripts of spoken language can be admitted as evidence in courts of law in England and Wales in a number of scenarios. When the speech content of an audio or video recording is used as evidence, e.g., a threatening voicemail message, the recording is often accompanied by a transcript to assist the court in "making out what was said and who said it" (Fraser, 2020). These recordings tend to be of very poor quality such that the speech is often close to unintelligible without the aid of a transcript. However, this means that the transcript can be highly influential on what members of the court believe they hear in the recording, as highlighted by Fraser and Kinoshita (2021; see also Fraser et al., 2011). It is therefore crucial that transcripts presented alongside speech evidence are as accurate as possible since they can play an important role in listeners' perception of speech and speakers, potentially leading to miscarriages of justice in cases where an utterance is inaccurately interpreted as incriminating (Harrison and Wormald, in press).

Another use of orthographic transcripts in the legal system is transcripts of police-suspect interviews, which play an important role in the investigative process and are often admitted as evidence in court (Haworth, 2018). While the audio recording of the police-suspect interview is technically the "real" evidence in this context, the transcript is admissible as a "copy" and is often the only version of the police-suspect interview that is referred to in the courtroom (Haworth, 2018). Given that the court often does not hear the original audio recording, it is important that the transcripts are an accurate representation of the interview's contents. However, Haworth (2018, 2020) has identified issues with the transcripts created by police transcribers, including summarizing large sections of the interview, paraphrasing the speech content and inconsistent representation across transcribers. A verbatim record of the speech would be ideal, but this is a time-consuming and laborious task.

Automatic speech recognition (ASR) technology is rapidly improving and can produce transcripts in a fraction of the time it would take a human to complete the same task. Transcripts produced by an ASR system would require manual checking and correction, but the output would be a verbatim record of the full interview, eliminating the issue of potentially important information being inaccurately paraphrased or omitted. A computer-assisted transcription method could lead to more reliable evidence being presented to courts without a significant increase in the time spent producing the records.

When considering the incorporation of ASR into the transcription process, it is important to take into account factors that have a significant impact on ASR performance, such as audio quality and regional accents. Background noise has been shown to decrease the accuracy of ASR systems in a number of contexts (Lippmann, 1997; Littlefield and Hashemi-Sakhtsari, 2002) including for forensic-like audio recordings (Harrington et al., 2022; Loakes, 2022). In recent years, a growing body of research has focused on systematic bias within automatic systems, i.e., underperformance for certain demographic groups, and significant disparities in performance have been demonstrated across accents. Transcripts tend to be significantly less accurate for non-native speakers (DiChristofano et al., 2022) or speakers of non-standard regional varieties (Markl, 2022). However, a limitation of work in this area is the use of word error rate (WER) for evaluating performance. WER is the ratio of errors in a transcript to the total number of words spoken and can be useful to highlight differences in performance across groups. However, this metric does not provide insights into where and why systems produce errors, or how evidentially significant those errors could be.

This paper presents work on the topic of automatic speech recognition in the context of police-suspect interview transcription, employing a novel method of analysis that combines industry-standard measures alongside detailed phonetic and phonological analysis. While WER is useful for an overview of performance, incorporating fine-grained linguistic analysis into the method permits a deeper understanding of the aspects of speech that prove to be problematic for automatic systems. The performance of three commercial ASR systems is assessed with two regional accents, across different audio qualities; the purpose of this assessment is to evaluate how practical it would be for ASR systems to play a role in the transcription of police-suspect interviews.

2. Background

This section covers a range of topics relevant to the present study. Firstly, Section 2.1 outlines the current situation regarding police-suspect interview transcription in England and Wales, and highlights the issues. Automatic speech recognition (ASR) is offered as part of a potential solution, and Section 2.2 covers a brief history of ASR and its rapid improvement in recent years. Section 2.3 describes research on the use of ASR for transcribing forensic audio recordings, which leads into the potential incorporation of ASR in the transcription of comparatively better quality audio recordings, i.e., police-suspect interviews, in Section 2.4. Section 2.5 addresses potential speaker-related factors that may affect ASR performance, such as regional accent. Finally, Section 2.6 outlines the research aims of the present study.

2.1. Transcription of police-suspect interviews in England and Wales

In England and Wales, police-suspect interviews are recorded according to requirements of the Police and Criminal Evidence Act 1984. The audio recording is subsequently used to produce a Record of Taped Interview (ROTI), and if the case ends up going to trial, the ROTI is often admitted as evidence alongside the original audio recording. However, the transcript itself often becomes effectively "interchangeable [with] and (in essence) identical" (Haworth, 2018, p. 434) to the audio evidence in the eyes of the court, and is often used as a substitute for the original audio recording. Relying on the transcript as the primary source of the interview's contents could be problematic in cases where speech has been omitted or inaccurately represented.

The police interview transcribers, also known as ROTI clerks, tend to be employed as administrative staff, and the job-specific skills required often include proficiency in audio and copy typing and a specific typing speed (Tompkinson et al., 2022). ROTI clerks receive little to no training or guidance on the transcription process (Haworth, 2018), which has the potential to create a systematic lack of consistency in transcription production, even within police forces. This is highlighted by an example provided in Haworth (2018) in which three ROTI clerks transcribe an unanswered question in three unique ways: "no response," "no audible reply" and "defendant remained silent." Each representation could potentially generate varying interpretations of the interviewee's character. It is also worth noting that the 43 territorial police forces in England and Wales operate individually, which contributes to the issue of inconsistency in transcription and transcript production across forces.

Another issue with ROTIs is that much of the interview is summarized and the transcriber, untrained in legal issues and protocol, will ultimately decide what is deemed as important and worthy of full transcription. This decision-making process could lead to serious consequences given Section 34 of the Criminal Justice and Public Order Act 1994, which states that the court may draw inferences if something later relied upon as evidence is not mentioned during the initial interview stage.

In accordance with Haworth (2018), this assessment of problematic issues surrounding ROTIs does not serve as a critique of the clerks hired to produce the transcripts, but of the wider process. Transcription, particularly of long stretches of speech, is a time-consuming and labor-intensive task that can take four to five times the length of the audio recording to transcribe for research purposes (Walford, 2001; Punch and Oancea, 2014), and a time factor of 40 to 100 for difficult forensic recordings (Richard Rhodes, personal communication). It is also prone to human error, for example spelling and punctuation mistakes (Johnson et al., 2014) and omission or misrepresentation of short function words, discourse markers and filled pauses (Stolcke and Droppo, 2017; Zayats et al., 2019). Transcribing spoken language, even when producing a verbatim transcript, is a complex and inherently selective process which carries the inevitable risk of systematic and methodological bias (Jenks, 2013; Kowal and O'Connell, 2014). Transcripts carry social and linguistic information, therefore transcribers have an enormous amount of power over the way in which people are portrayed (Jenks, 2013).

Discrepancies concerning the portrayal of speakers have been reported within legal transcripts (e.g., US court reports, UK police interviews), with standardized language and "polished" grammar for professionals such as lawyers, expert witnesses and police interviewers but verbatim transcription or inconsistently-maintained dialect choices for lay witnesses or suspects (Walker, 1990; Coulthard, 2013). Similar inconsistencies were observed in ROTIs (Haworth, 2018), as well as an assumption revealed in focus group discussions with ROTI clerks that the interviewee will be charged with or convicted of an offense, as demonstrated through the use of terms such as "defendant" or "offender" to refer to interviewees (89% of references; Haworth, 2018, p. 440).

The use of ASR could address a number of the concerns regarding the production of police interview transcripts. Automatic systems can process a large amount of data in a fraction of the time it would take a human to do the same task. This could allow for interviews to be transcribed in full, rather than mostly summarized, while saving time, effort and money on behalf of the police. An automatic system would not apply social judgements to the role of interviewer and interviewee, and would therefore likely remain consistent in its treatment of speakers in this regard, given that only the speech content would be transcribed. Furthermore, an ASR system would likely be consistent in its representation of phenomena such as silences; for example, unanswered questions simply would not be transcribed, and therefore the system would not inject potentially subjective statements such as "defendant remained silent."

2.2. Automatic speech recognition

The field of automatic speech recognition (ASR) has received growing interest over the last decade given its expanding applications and rapid improvements in performance, though this technology has existed in different forms for over 70 years. The first speech recogniser was developed in 1952 at Bell Telephone Laboratories (now Bell Labs) in the United States and was capable of recognizing 10 unique numerical digits. By the 1960's systems were able to recognize individual phonemes and words, and the introduction of linear predictive coding (LPC) in the 1970's led to rapid development of speaker-specific speech recognition for isolated words and small vocabulary tasks (Wang et al., 2019). The 1980's saw the creation of large databases (O'Shaughnessy, 2008) and the implementation of a statistical method called the "Hidden Markov Model" (HMM) which allowed systems to recognize several thousand words and led to substantial progress in the recognition of continuous speech (Wang et al., 2019). Combining HMM with a probabilistic Gaussian Mixture Model (HMM-GMM) created a framework that was thoroughly and extensively researched throughout the 1990's and 2000's, and remained the dominant framework until the last decade when deep learning techniques have become prevalent (Wang et al., 2019). In recent years deep neural networks (DNN) have been implemented to create the HMM-DNN model, achieving performance well beyond its predecessor.

Modern state-of-the-art ASR systems are typically made up of two main components, an acoustic model and a language model, both of which are concerned with calculating probabilities. As a basic summary according to Siniscalchi and Lee (2021), the acoustic model recognizes speech as a set of sub-word units (i.e., phonemes or syllables) or whole word units. It is then tasked with calculating the probability that the observed speech signal corresponds to a possible string of words. The language model then calculates the probability that this string of words would occur in natural speech. This is often evaluated using n-grams, which calculate the probability of the next word in a sequence given the n previous words, based on extensive training on large text corpora. Both models contribute to the estimated orthographic transcription produced by the ASR system.

Adaptations to the architecture of ASR systems have led to huge improvements in accuracy, which can be illustrated by observing the reported word error rates (WER) on a commonly-used dataset for measuring ASR performance, such as the Switchboard corpus

(Godfrey and Holliman, 1993). This is a dataset of American English conversational telephone speech that is commonly used to benchmark ASR performance. The first reported assessment of speech recognition performance had a WER of around 78% (Gillick et al., 1993) and by 2005 state-of-the-art systems were yielding WER measures between 20 and 30% (Hain et al., 2005). Thanks to large amounts of training data and the application of machine learning algorithms, huge improvements in speech technology have been demonstrated in recent years. In 2016, Microsoft reported that their automatic system had achieved human parity, with a WER of 5.8% compared with a human error rate of 5.9% on a subset of the Switchboard data (Xiong et al., 2016). In 2021, IBM reported an even lower WER of 5.0% on a subset of the Switchboard data, reaching a new milestone for automatic speech recognition performance (Tüske et al., 2021).

It is crucial to acknowledge, though, that performance is relative to the materials being transcribed. Though trying to mimic spontaneous conversations, the Switchboard corpus contains "inherently artificial" (Szymański et al., 2020) spoken data due to factors such as the predefined list of topics, the localized vocabulary and the relatively non-spontaneous form of the conversations. These factors, paired with the relatively good audio quality, create conditions which are favorable to ASR systems, and while ASR may outperform human transcribers in some cases, there will be circumstances in which the reverse is true, especially in more challenging conditions such as forensic audio.

2.3. Automatic transcription of forensic audio recordings

Some work within the field of forensic transcription has considered whether automatic methods could be incorporated into the transcription of forensic audio samples, such as covert recordings. The audio quality of such recordings is generally poor given the real-world environments in which the recordings are made, and as a result of the equipment being deployed in a covert manner, rather than one designed to capture good-quality audio. They can also be very long, containing only a few sections of interest; it is often necessary to transcribe the recording in full to identify such sections, which is a time-consuming and arduous task for forensic practitioners.

Two studies in particular have explored automatic transcription in forensic-like contexts, the first of which uses an audio recording of a band rehearsal (Loakes, 2022), comparable to a covert recording. Two automatic transcription services (BAS Web Services and Descript) were employed to transcribe the 44 s recording containing the sounds of musical instruments and multiple speakers from a distance. BAS Web Services returned a system error when an orthographic transcription was requested, and when the in-built WebMINNI service was employed to segment the speech into phonemes, many sections of speech were identified as "non-human noise" and instrument noises were labeled as speech. Descript was also unsuccessful in its attempt to transcribe the speech, with the output containing only three distinct words ("yes," "yeah," and "okay"), a fraction of the total number of words uttered.

A second study on the topic of forensic transcription compared the performance of 12 commercial automatic transcription services using a 4-min telephone recording of a conversation between five people in a busy restaurant (Harrington et al., 2022). Talkers were positioned around a table upon which a mobile device was placed to record the audio, and all were aware of its presence. The transcripts produced by the automatic systems were of poor quality, making little sense and omitting large portions of speech, although this is not surprising given the high levels of background noise and numerous sections of overlapping speech.

A number of relatively clear single-speaker utterances were selected for further analysis, and results showed that even in cases of slightly better audio quality and more favorable speaking conditions, transcripts were far from accurate. The best performing system (Microsoft) produced transcripts in which 70% of words on average matched the ground truth transcript, though there was a high level of variability across utterances. Microsoft transcribed seven of the 19 utterances with over 85% accuracy, but many of the other transcriptions contained errors that could cause confusion over the meaning, or even mislead readers. For example, "*that would have to be huge*" was transcribed as "*that was absolutely huge*," changing the tense from conditional (something that could happen) to past (something that has happened). In many cases, the automatic transcript would need substantial editing to achieve an accurate portrayal of the speech content.

The findings of such research, though valuable, are unsurprising given that commercial ASR systems are not designed to deal with poor quality audio; they are often trained on relatively good quality materials more representative of general commercial applications. Following recent advances in learning techniques to improve ASR performance under multimedia noise, Mošner et al. (2019) demonstrated that a system trained on clean and noisy data achieved better performance (i.e., higher reductions in WER) than a system trained only on clean data. It seems that training data has a direct effect on the capabilities of ASR systems. There could potentially be a place for automatic systems within the field of forensic transcription if the training data used is comparable to the audio recordings that would ultimately be transcribed. However, it is impractical to expect commercial ASR systems to perform at an appropriate level for the type of data that forensic practitioners handle.

Given the current state of the technology, ASR should therefore not be employed for the transcription of poor quality audio such as covert recordings, though the question remains as to whether it could be incorporated for comparatively better quality audio samples, such as police interviews. This type of audio recording is much better suited to automatic transcription for many reasons. The quality of police-suspect interview recordings tends to be much better since the equipment utilized is built specifically for the purposes of recording audio, and all members present are aware of the recording process. The number of speakers is limited and known, and the question-and-answer format of the interview will most likely result in speech that is easier to transcribe, i.e., less overlapping speech. The level of background noise will also likely be much lower than a busy restaurant or a music practice room, although it must be noted that the audio quality of these interviews is not always ideal or comparable to studio quality audio. Reverberation, broadband noise or interference, the rustling of papers and the whirring of laptop fans (Richard Rhodes, personal communication) are examples of frequently occurring issues encountered within police interview recordings which can make some sections difficult to transcribe.

2.4. Incorporating automatic methods into police transcription

One approach to the use of automatic methods would be the use of an automatically-produced transcript as a starting point to which human judgements could be added i.e., "post-editing" an ASR output. Bokhove and Downey (2018) suggest that using automatic transcription services to create a "first draft" could be worthwhile in an effort to reduce the time and costs involved in human transcription. In their study, many of the errors made by the ASR system for interview data were relatively small and easily rectifiable, while recordings of a classroom study and a public hearing (with many speakers and microphones positioned far away from speakers) resulted in automatic transcriptions that deviated more substantially from the audio content. Nonetheless, Bokhove and Downey (2018) argue that, with little effort, reasonable "first versions" can be obtained through the use of freely available web services, and that these may serve as a useful first draft in a process which would involve multiple "cycles" or "rounds" of transcription (Paulus et al., 2013) regardless of the inclusion of automatic methods.

However, the baseline performance of the ASR system is a key issue in whether combining ASR and human transcription is viable. By artificially manipulating the accuracy of transcripts, Gaur et al. (2016) demonstrated that the time spent correcting an ASR output can exceed the time spent creating a transcript from scratch if the automatically-produced transcript is insufficiently accurate. By manipulating the WER of transcripts at rough intervals of 5% ranging between 15 and 55%, it was found that by the time the WER had reached 30% participants were able to complete the post-editing phase more quickly by typing out the content from scratch. However, participants only realized that the quality of the original transcript was a challenge when the WER reached around 45%. These findings suggest that post-editing an ASR output could reduce the time taken to produce a verbatim transcript provided that the WER does not exceed a certain level; however, if the WER consistently approaches 30% then the incorporation of automatic methods into the transcription process fails to be a worthwhile avenue of research.

There are, however, some issues with using WER as the defining metric of system performance, as highlighted by Papadopoulou et al. (2021). Firstly, WER can be expensive and time-consuming to calculate due to the requirement of manual transcriptions to use as a reference. Secondly, quantified error metrics do not take into account the cognitive effort necessary to revise the ASR transcripts into a "publishable" quality. A more useful metric for analyzing ASR outputs is the post-editing effort required. In their study, a single post-editor with intermediate experience in the field was tasked with post-editing transcripts produced by four commercial ASR systems (Amazon, Microsoft, Trint, and Otter). Both the time taken to edit the ASR output and the character-based Levenshtein distance between the automatic and post-edited transcripts were measured.

An interesting finding by Papadopoulou et al. (2021) is that the number of errors within a transcript does not always correlate with the amount of effort required for post-editing. Systems with the lowest error rates do not necessarily achieve the best scores in terms of the post-editing time and distance. Certain types of errors were shown to take longer to edit, such as those related to fluency, i.e., filler words, punctuation and segmentation. The authors also suggest that deletion and insertion errors are easily detectable and require less cognitive effort to edit than substitution errors. Although little justification for this claim is put forth in the paper, it does seem likely that deletions and insertions could be easier to identify given that the number of syllables will not match up between the speech content and the transcript. The post-editor may find substitutions more challenging to detect, especially if the phonetic content of the target word and transcribed word is similar. It is therefore crucial to consider the types of errors made, not just overall error rates, when assessing the viability of an automatic transcript as a first draft.

The study carried out by Papadopoulou et al. (2021) claims to be one of the first papers to evaluate the post-editing effort required to transform ASR outputs into useable transcripts and to conduct qualitative analysis on ASR transcription errors. Given that WER does not reveal sufficient information regarding the types of errors made and the difficulty of correcting those errors, there is a clear need for additional research on the topic of post-editing and alternative methods of analysis. This is particularly true when evaluating the practicality of incorporating ASR into the transcription process, as the effort required to transform an ASR output into a fit-for-purpose verbatim transcript is the main consideration in whether this approach is advantageous, rather than the number of errors in the initial transcript.

2.5. Automatic systems and speaker factors

Given that the speakers taking part in police-suspect interviews will come from a range of demographics, it is important to consider how this may affect the performance of automatic speech recognition systems. Factors relating to a speaker's linguistic background, such as accent, can prove challenging for an automatic transcription system. Previous work has demonstrated that the performance of ASR systems declines significantly when confronted with speech that diverges from the "standard" variety; this has been found for non-native-accented speech in English (Meyer et al., 2020; DiChristofano et al., 2022; Markl, 2022) and Dutch (Feng et al., 2021), as well as for non-standard regionally-accented speech in Brazilian Portuguese (Lima et al., 2019) and British English (Markl, 2022).

Markl (2022) compared the performance of Google and Amazon transcription services across multiple accents of British English. One hundred and two teenagers from London or Cambridge (South of England), Liverpool, Bradford, Leeds, or

Newcastle (North of England), Cardiff (Wales), Belfast (Northern Ireland), or Dublin (Republic of Ireland) were recorded reading a passage from a short story. Both systems demonstrated significantly worse performance, based on WER, for some of the non-standard regional accents compared with the more "standard" Southern English accents. Amazon performed best for speakers from Cambridge and showed a significant decline in performance for those from parts of Northern England (Newcastle, Bradford, and Liverpool) and Northern Ireland (Belfast). Much higher error rates were reported for Google for every variety of British English, likely as a result of much higher rates of deletion errors. Google performed best for speakers of London English and saw a significant drop in performance only for speakers from Belfast.

Many researchers have suggested that the composition of training datasets can cause bias within automatic systems (Tatman, 2017; Koenecke et al., 2020; Meyer et al., 2020; Feng et al., 2021) and that the underrepresentation of certain accents leads to a decline in performance for those varieties. Markl (2022) reports that certain substitution errors identified for speakers of non-standard regional accents of British English suggest that there is an overrepresentation of Southern accents in the training data or that acoustic models are being trained only on more prestigious Southern varieties, such as Received Pronunciation. Similarly, Wassink et al. (2022) claim that 20% of the errors within their data would be addressed by incorporating dialectal forms of ethnic varieties of American English (African American, ChicanX, and Native American) into the training of the automatic systems. The implementation of accent-dependent (or dialect-specific) acoustic models has been found to improve performance, particularly for varieties deviating more substantially from the standard variety, such as Indian English and African American Vernacular English (Vergyri et al., 2010; Dorn, 2019).

2.6. Research aims

The main aim of the present research is to assess ASR transcription errors across accents and audio qualities. The implications of such errors being retained in a transcript presented to the court will be considered, and methods of analysis that are appropriate for this particular context will be employed. This work is centered on the transcription of recordings resembling police interview data, and a further aim of this work is to consider the practicality of incorporating ASR into the transcription of police-suspect interviews.

The present study will explore in much greater detail the types of errors produced across two different accents of British English, and will focus not only on the distribution of three main error categories (deletions, substitutions, and insertions), but also on the distribution of word types that feature in the errors. For example, some substitutions may be more damaging than others, or more difficult to identify in the post-editing of a transcript. Errors will also be assessed from a phonological perspective in order to identify errors resulting from phonological differences across the accents and highlight particularly challenging phonetic variables for the automatic systems. Although both the acoustic and language model will affect ASR performance, the analysis and interpretation of errors in this study will focus on those which are most likely a reflection of the acoustic model.

In this study, recordings that are representative of police interviews in the UK (in terms of speech style and audio quality) are used, which are expected to degrade ASR performance compared with previous studies that have typically used high quality materials. The present study considers, from a practical perspective, whether this technology could be incorporated into the transcription process for police-suspect interviews.

The specific research questions are:

1. How do regional accent and audio quality affect the performance of different ASR systems?
2. What types of errors are produced by the ASR systems, and what are the implications of these errors?
3. To what extent could ASR systems produce a viable first draft for transcripts of police-suspect interviews?

3. Materials and methods

3.1. Stimuli

In order to explore differences in ASR performance across different regional accents, two varieties of British English were chosen for analysis: Standard Southern British English (SSBE) and West Yorkshire English (WYE). SSBE is a non-localized variety of British English spoken mostly in Southern parts of England, and although linguistic diversity is celebrated in contemporary Britain, SSBE is heard more frequently than other accents in public life (e.g., TV programmes and films), especially in media that is seen on an international scale, and acts as a teaching standard for British English (Lindsey, 2019). SSBE is referred to in this study as a "standard" variety. WYE is a non-standard regional variety of British English which shares characteristics with many other Northern English accents[1] and whose phonology diverges substantially from SSBE (Hickey, 2015).

Stimuli were extracted from two forensically-relevant corpora of British English: the Dynamic Variability in Speech database (DyViS; Nolan et al., 2009) and the West Yorkshire Regional English Database (WYRED; Gold et al., 2018). DyViS contains the speech of 100 young adult males from the South of England (the majority of whom had studied at the University of Cambridge) taking part in a number of simulated forensic tasks, such as a telephone call with an "accomplice" and a mock police interview. WYRED contains the speech of 180 young adult males from three parts of West Yorkshire (Kirklees, Bradford, and Wakefield) and was created to address the lack of forensically-relevant population data for varieties of British English other than SSBE. The collection

[1] West Yorkshire English shares some features (e.g., lack of TRAP-BATH and FOOT-STRUT splits) with General Northern English (GNE), an emerging variety of Northern British English which is the result of dialect leveling (Strycharczuk et al., 2020). However, there are some features that make WYE distinct from GNE, such as the monophthongization of vowels in words like "face" and "goat."

TABLE 1 Examples of linguistic content of stimuli from each speaker.

Speaker	Utterance
SSBE-1	And um there's also a boat house but that's obviously that's quite hard to see from there
SSBE-2	Not exactly I can't really remember their surnames but I might have known them I don't know
WYE-1	Uh can get a bit inebriated sometimes so not all the time no can't say
WYE-2	Yeah quarter of an hour half an hour something like that depending on traffic

procedures employed in the production of the DyViS database were closely followed for WYRED, resulting in very closely matched simulated forensic conditions.

The mock police interview contained a map task in which specific speech sounds were elicited through the use of visual stimuli. Participants assumed the role of a suspected drug trafficker and had to answer a series of questions regarding their whereabouts at the time of the crime, their daily routine and their work colleagues, among other things. Visual prompts were provided during the task, containing information about the events in question and incriminating facts shown in red text. Participants were advised to be cooperative but to deny or avoid mentioning any incriminating information. The speech was conversational and semi-spontaneous, and the question-and-answer format of the task was designed to replicate a police-suspect interview. On account of the focus on police-suspect interview transcription in this paper, the mock police interview task was selected for this study.

Two speakers of each accent were selected and eight short utterances were extracted per speaker, resulting in a total of 32 utterances. Much of the speech content in this task contained proper nouns such as the surnames of colleagues and place names. With the exception of two well-known brands, "*Skype*" and "*Doritos*," proper nouns were not included in the extracted utterances in order to avoid inflated error rates as a result of misspellings or due to the name not featuring in the ASR system's vocabulary. Other than filled pauses, which were extremely common in the spoken data, effort was also made to exclude disfluent sections. Disfluencies have been shown to be problematic for ASR systems (Zayats et al., 2019), therefore sections containing false starts or multiple repetitions were excluded in order to isolate differences in performance due to regional accent. Utterances ranged between 14 and 20 words in length and 3–6 s in duration, each containing a single speaker and unique linguistic content. Some examples of the utterances are included in Table 1 (and reference transcripts for all utterances can be found in Supplementary material).

To investigate the effects of low levels of background noise, such as that commonly found in real police interviews, the studio quality recordings were mixed with speech-shaped noise, derived from the HARVARD speech corpus. This was carried out using Praat (Boersma and Weenink, 2022), and the resulting files had an average signal-to-noise ratio (SNR) of 10 dB, such that intelligibility was not hugely impacted but the background noise was noticeable. The studio quality files had a much higher average SNR of 22 dB, reflecting the lack of background noise in these recordings. To summarize, a studio quality version and a poorer quality version (with added background noise) of each file was created, resulting in a total of 64 stimuli for automatic transcription.

3.2. Automatic transcription

Three commercially-available automatic transcription services were used to transcribe the audio files: Rev AI[2], Amazon Transcribe[3], and Google Cloud Speech-to-Text[4]. Many automatic transcription systems acknowledge that background noise and strongly accented speech can decrease transcription accuracy. Rev AI was chosen due to its claims of resilience against noisy audio and its Global English language model which is trained on "a multitude of... accents/dialects from all over the world" (Mishra, 2021). Services from Amazon and Google were chosen due to their frequent use in other studies involving ASR and the prevalent use of products from these technology companies in daily life. When uploading the files for automatic transcription, "British English" was selected as the language for Amazon and Google, and, since this option was not available for the third service, "Global English" was selected for Rev AI.

Reference (i.e., ground truth) transcripts were manually produced by the author for each utterance, using the studio quality recordings. The automatic transcripts produced by Amazon, Google, and Rev were compiled in a CSV file. Amazon and Google offer confidence levels for each word within the transcription but for the purpose of this research, only the final output (i.e., the highest probability word) was extracted.

3.3. Error analysis

Custom-built software was written to align the reference and automatic transcripts on a word-level basis, and each word pairing was assessed as a match or an error. Errors fall into three categories as outlined below:

- Deletion: the reference transcript contains a word but the automatic transcript does not.
- Insertion: the reference transcript does not contain a word but the automatic transcript does.
- Substitution: the words in the reference transcript and automatic transcript do not match.

From a forensic perspective, insertions, and substitutions are potentially more harmful than deletions, on the assumption that reduced information causes less damage than false information in case work (Tschäpe and Wagner, 2012). Table 2 shows an example of two potential transcriptions of the utterance "*packet of gum in*

[2] Rev AI accessed 12th November 2021.

[3] Amazon Transcribe accessed 17th October 2022.

[4] Google Cloud Speech-to-Text accessed 13th October 2022.

TABLE 2 Two potential transcriptions of the utterance "*packet of gum in the car*."

Reference	Packet	Of	Gum	In	The	Car
Transcript 1			Gum	In		Car
Transcript 2	**Pack**	**The**	**Gun**	In	The	Car

Deletions are represented by a shaded red cell and substitutions are represented by bolded red text.

the car," and demonstrates the different effect that substitutions can have in comparison with deletions. Both transcripts contain three errors, but the substitutions in transcript 2 could be much more damaging given the change in content and the new potentially incriminating interpretation of the utterance.

Some minor representational errors were observed, such as "*steak house*" transcribed as a compound noun "*steakhouse*" and numbers transcribed as digits. Since these substitutions do not constitute inaccuracies, rather slight changes in representation, the word pairing was marked as a match and these were not included as errors in the subsequent analysis. With regards to substitutions spanning multiple words, it was decided that the collective error would be marked as one substitution. For example, "*cut and*" transcribed as "*cutting*" was marked as a substitution rather than a combination of a substitution and a deletion, in an attempt to avoid inflated insertion and deletion rates.

Despite the limitations of WER, particularly in a forensic context, this metric can provide a brief overview of system performance across groups that can be used as a starting point for analysis. WER was therefore calculated for each utterance, by dividing the total number of errors (deletions, insertions, and substitutions) by the number of words in the reference transcript. The total number of each type of error in each condition was also calculated and compared to explore the differences across the ASR systems as well as the effects of regional accent and level of background noise. In order to explore in greater detail the types of words involved in errors, each error pairing was manually evaluated as involving content words, function words, filled pauses or a combination of these. Substitutions involving morphological alterations were also highlighted, and transcripts were assessed in terms of the effort required to transform the ASR output into a more accurate, verbatim transcript.

Errors were also assessed on a phonological level in order to explore whether varying phonetic realizations of features across accents could be responsible for transcription errors, with a particular focus on marked vocalic differences across SSBE and WYE. Substitutions involving content words in the Yorkshire English transcripts were analyzed by identifying which of Wells' lexical sets (i.e., group of words all sharing the same vowel phoneme; Wells, 1982) the words in the reference and automatic transcripts belong to as well as transcribing the speaker's production of the word, with the goal of better understanding why the error may have been made.

Four vocalic variables in particular were analyzed due to differences between the SSBE and WYE phonetic realizations (Wells, 1982; Hughes et al., 2005). These are outlined in Table 3, using Wells' (1982) lexical sets as a way of grouping words that share the same phoneme. Words in the BATH lexical set contain a long back vowel in SSBE, but typically contain a short front vowel in WYE, which overlaps with the production of the TRAP vowel [a] in both varieties. Words in the STRUT lexical set contain an unrounded low vowel in SSBE, but a rounded high vowel in WYE; the rounded high vowel [ʊ] is also produced in words belonging to the FOOT lexical set in both varieties. Words belonging to the FACE and GOAT lexical sets contain diphthongs in SSBE, but typically contain monophthongs in WYE.

TABLE 3 Phonetic realizations of four vocalic variables across the two varieties of British English analyzed in this study, Standard Southern British English (SSBE) and West Yorkshire English (WYE).

Lexical set	SSBE	WYE
BATH	[ɑː]	[a]
STRUT	[ʌ]	[ʊ]
FACE	[eɪ]	[eː ∼ ɛː]
GOAT	[əʊ]	[oː]

Variables are defined using Wells' (1982) lexical sets.

3.4. Statistical analysis

In order to evaluate which factors had a significant effect on word error rate, three linear mixed effects models were fitted using the lme4 package (Bates et al., 2015) in R. In each model, regional accent, audio quality or ASR system was included as a fixed effect, and all models included Speaker and Sentence as random effects to account for variation across speakers within accent groups and the unique linguistic content of each utterance. A separate "null" model was fitted including only the random effects, and the ANOVA function in R was used to compare each full model with the null model. Results of the model comparisons indicate whether the full model is better at accounting for the variability in the data, and therefore whether the fixed effect has a significant impact on word error rate. Results of the model outputs, containing an Estimate, Standard Error rate and a p-value, were then examined to evaluate the relationship between variables. A threshold of $\alpha = 0.05$ was used to determine statistical significance.

A three-way comparison was carried out for ASR system and in the first three models Amazon was used as a baseline, meaning that a comparison between Rev and Google had not been carried out. The "ASR system" variable was relevelled such that Rev became the baseline, and a fourth model was then fitted with ASR system as a fixed effect and Speaker and Sentence as random effects.

4. Results

4.1. ASR systems

The three automatic systems tested in this study performed with varying levels of success and were all clearly affected to some degree by the regional accent of the speaker and the level of background noise. Figure 1 shows WER in each condition for the three ASR systems. The four conditions are SSBE speech in

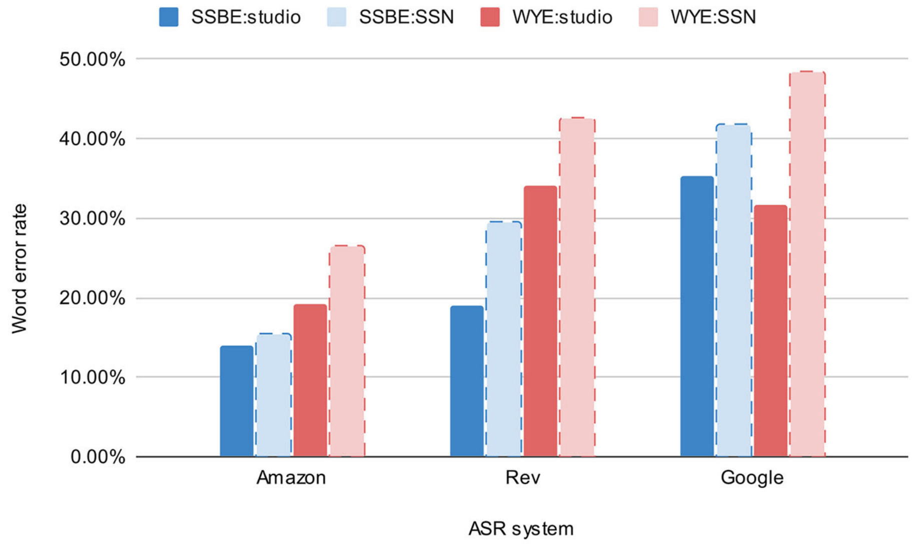


FIGURE 1
Average word error rate in each of the four conditions (SSBE studio, SSBE SSN, WYE studio, and WYE SSN) for all three ASR systems (Amazon, Rev, and Google). ASR systems are ordered from left to right according to lowest to highest average WER.

studio quality audio, SSBE speech in audio with added speech-shaped noise, WYE speech in studio quality audio and WYE speech in audio with added speech-shaped noise; these will henceforth be referred to as SSBE studio, SSBE SSN, WYE studio and WYE SSN, respectively. Amazon was the best performing system with the lowest word error rate (WER) in each of the four conditions compared with Rev and Google, and achieved its lowest WER (13.9%) in the SSBE studio condition and highest WER (26.4%) in the WYE SSN condition. Google was the worst performing system, achieving the highest WER in every condition except for WYE speech in studio quality, for which Rev performed worst with a WER of 34.1%.

Results of a model comparison between the null model and the model with ASR system as a fixed effect revealed that ASR system has a significant impact on WER [$\chi^2_{(2)} = 50.35$, $p < 0.0001$]. The summary output of the linear mixed effects model revealed that there was a significant difference in error rates between Amazon and both Rev ($\beta = 0.13$, SE $= 0.26$, $p < 0.001$) and Google ($\beta = 0.20$, SE $= 0.26$, $p < 0.001$). Rev achieved WERs that were on average 13% higher than those produced by Amazon, while Google produced WERs on average 20% higher than Amazon. When comparing the two worst performing systems, Google was found to produce significantly higher WERs than Rev ($\beta = 0.08$, SE $= 0.03$, $p < 0.005$).

A notable trend in the data was Google's high tendency toward deletion errors, with over double (and in some cases quadruple) the number of deletions that Amazon produced in the same condition. An example of this is the utterance "*not exactly I can't really remember their surnames but I might have known them I don't know*" which was transcribed in studio quality by Amazon as "*not exactly I can't remember their names but I might have known him I don't you*" (with one deletion and three substitutions) and by Google as "*not exactly I can't remember this sentence I don't know*" (with seven deletions and two substitutions).

4.2. Regional accent

There are some clear differences in performance between the two accents in this study. Word error rate is lower for SSBE than for WYE in all conditions except for Google in the WYE studio condition; however, the results of a model comparison between the null model and the model with regional accent as a fixed effect showed that the difference in performance across accents was not statistically significant [$\chi^2_{(1)} = 1.28$, $p = 0.26$]. This is likely due to the extremely small sample size and variation in system performance across the speakers of each accent. All ASR systems produced higher WERs for one of the SSBE speakers, which were on average 13 and 20% higher than for the other SSBE speaker in studio quality audio and speech-shaped noise audio, respectively. One of the WYE speakers also proved more challenging for the ASR systems, though the difference was most pronounced in studio quality where WERs were on average 10% higher than for the other WYE speaker. An average difference of 4% was observed between the WYE speakers in speech-shaped noise audio, which is likely a result of the highest WERs in the study being observed in this condition.

The most common type of error also varied across accents, with deletions featuring most frequently for SSBE speech (see Table 4) and substitutions featuring most frequently for WYE speech (see Table 5). As discussed earlier in this paper, substitution errors can be viewed as more harmful than deletion errors in forensic contexts given that incorrect information has the potential to be much more damaging than reduced information. Substitutions may also have a stronger priming effect than other types of errors on the post-editors who are correcting an ASR transcript.

55.6% of SSBE errors in studio quality audio and 62.7% of SSBE errors in speech-shaped noise audio were deletions. The number of deletions in SSBE was consistently higher than in WYE, though occasionally only by a relatively small margin. The majority of

TABLE 4 Counts of each error type (insertions, deletions, and substitutions) produced by each system for Standard Southern British English speech.

ASR system	Audio quality	INS	DEL	SUB	Total errors
Amazon	Studio	0	20	16	36
Amazon	SSN	0	26	14	40
Rev	Studio	0	25	25	50
Rev	SSN	2	43	30	75
Google	Studio	0	57	36	93
Google	SSN	1	73	37	111

SSN refers to the audio quality with added speech-shaped noise.

TABLE 5 Counts of each error type (insertions, deletions, and substitutions) produced by each system for West Yorkshire English speech.

ASR system	Audio quality	INS	DEL	SUB	Total errors
Amazon	Studio	1	13	33	47
Amazon	SSN	3	16	46	65
Rev	Studio	3	22	53	78
Rev	SSN	5	30	64	99
Google	Studio	4	39	42	85
Google	SSN	4	71	50	125

SSN refers to the audio quality with added speech-shaped noise.

deletion errors involved short function words, such as "*a*" and "*to*," which made up between 61.5 and 80% of all deletion errors for Rev and Google. Amazon made the fewest deletion errors out of all the ASR systems, and the majority of the deletions for SSBE speech involved the omission of filled pauses. The deletion of content words was much less frequent, accounting for 17.9% of all deletion errors for Rev and 16.3% of all deletion errors for Google. Amazon was the only system for which content words were never deleted.

Substitutions accounted for the most frequently occurring type of error for West Yorkshire English speech, with an average of 62.5% of all errors in studio condition and 58.5% of all errors in the speech-shaped noise condition involving the substitution of words or phrases. The only condition in which substitutions were not the most frequently occurring type of error for WYE speakers was Google in the speech-shaped noise condition where deletions constituted 71 of the 125 errors. The distribution of word types involved in substitution errors also differed across accents. The majority of substitutions for WYE speech involved content words while most substitutions for SSBE speech involved function words (Figure 2).

Despite substitutions relating to function words accounting for a minority of substitution errors in WYE, there were more of this type of error in WYE than in SSBE. For Amazon and Rev, the number of content word-related substitutions was between 2 and 5 times higher for Yorkshire English than for SSBE, and the smaller increase for Google was likely a result of higher numbers of substitutions for SSBE speakers.

4.3. Audio quality

Higher error rates (by an average of 8%) were observed in speech-shaped noise audio compared with studio quality audio for all systems and for both accents. The results of a model comparison between the null model and the model with audio quality as a fixed effect showed that this difference was statistically significant [$\chi^2_{(1)} = 11.42$, $p < 0.001$], and examination of the model output confirmed that WER was significantly higher in the degraded audio condition ($\beta = 0.08$, SE $= 0.02$, $p < 0.001$). An increase was observed in the number of insertions and deletions in all conditions when comparing the transcripts of the studio quality recordings to the recordings with added speech-shaped noise. Rev and Google in particular show large increases in the number of deletions from studio condition to the speech-shaped noise condition. A very similar number of substitutions was observed across the audio qualities in SSBE, but the number of substitutions in WYE was 19–40% higher in the speech-shaped noise condition. The change in audio quality also affected the distribution of word types involved in substitutions. While the majority of substitution errors in SSBE were related to function words in studio quality audio, a majority involved content words in the speech-shaped noise condition for both Rev and Google. Not only was Amazon the highest performing system overall, it was also the least affected by the addition of background noise.

4.4. Phonological variables

Many errors within the West Yorkshire English data could be explained by a phonetic realization deviating from what might be expected based on the assumed underlying acoustic models. This was especially true in the case of vowels where the phonology deviates markedly from SSBE. Given that previous studies suggest an overrepresentation of more "standard" (in this context, Southern British) varieties in training data, we may expect to see the ASR

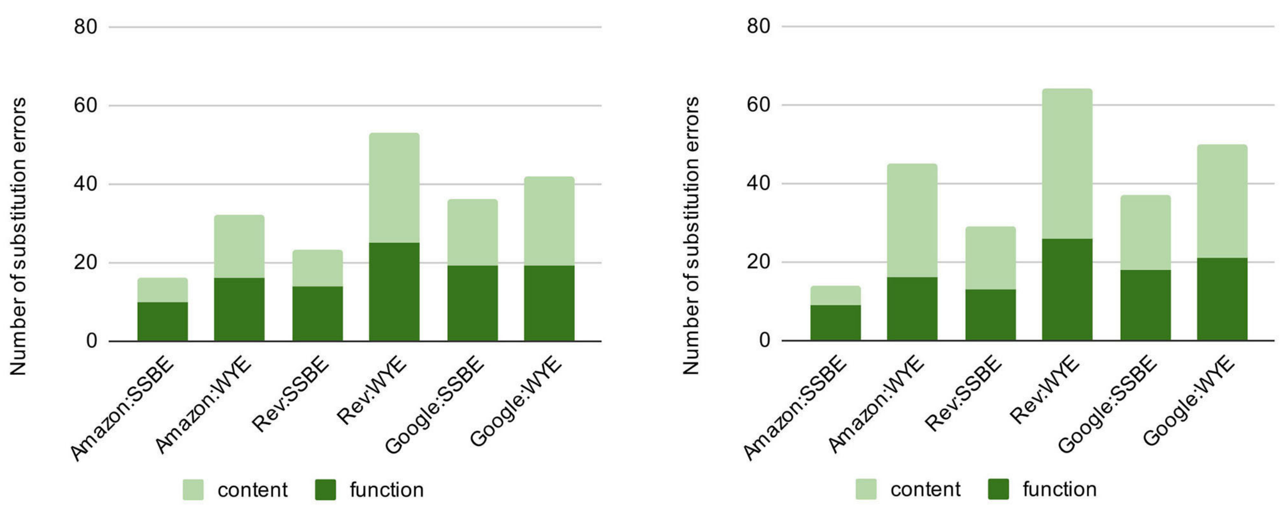


FIGURE 2
Number of substitution errors produced by each ASR system in each accent, in studio condition **(left)** and speech-shaped noise condition **(right)**. ASR systems are ordered from left to right according to lowest to highest average WER.

systems struggling with some of the non-standard pronunciations of words by Yorkshire speakers. To explore this, four vowels which are well-known to differ in quality, length, or number of articulatory targets across SSBE and WYE were chosen for more in-depth analysis.

4.4.1. BATH

Words belonging to the BATH lexical set contain different vowels within the two accents: the long back vowel [ɑː] in SSBE and, like many other varieties from the North of England, the short front vowel [a] in WYE. There were few occurrences of words belonging to the BATH lexical set in the Yorkshire data, though there were two utterances of the word "*staff*," one by each of the Yorkshire speakers, which were produced with a short front vowel, i.e., [staf], rather than a long back vowel, i.e., [stɑːf]. All three systems correctly transcribed this word for one speaker but not for the other. The pronunciations themselves were very similar but the surrounding context of the word was likely the cause of this issue. In the successful case, "*staff*" was uttered at the beginning of an intonational phrase but in the other occurrence it was preceded by a non-standard pronunciation of "*with*" [wɪʔ]. Omission of word-final fricatives, most commonly in function words, is a common process in some varieties of Yorkshire English (Stoddart et al., 1999). In this case, the voiced dental fricative /ð/ has been replaced with a glottal stop, resulting in the utterance [wɪʔstaf] which Rev and Google both analyzed as one word, transcribing "*waste*" and "*Wigston*," respectively. Amazon mistranscribed the word "*staff*" as "*stuff*," a substitution which could be the result of the Yorkshire vowel being replaced with the closest alternative that creates a word in Standard Southern British English. Since [staf] in this case is not recognized as the word "*staff*," the closest SSBE alternative is the word "*stuff*" which contains a low central vowel [ʌ] that is closer within the vowel space to the uttered vowel than [ɑː].

4.4.2. STRUT

There is a systemic difference between SSBE and WYE with regards to the number of phonemes in each accent's phonological inventory, whereby the SSBE STRUT vowel/ʌ/does not feature in WYE. Instead, [ʊ] is produced in words belonging to both the STRUT and FOOT lexical sets. Many words containing this vowel were correctly transcribed within the Yorkshire data, though some occurrences resulted in phonologically-motivated substitutions. The word "*bus*," pronounced [bʊs] by the Yorkshire speaker, was correctly transcribed by Amazon and Google but proved challenging for Rev which replaced it with "*books*," a word containing [ʊ] in SSBE and belonging to the FOOT lexical set. A similar pattern was observed for the word "*cut*," pronounced [kʊʔ], which Amazon and Google transcribed (almost correctly) as the present participle "*cutting*," while Rev substituted it with a word from the FOOT lexical set, "*couldn't*."

The word "*muddy*," pronounced [mʊdɪ] by the Yorkshire speaker, proved challenging for all three systems. In both audio qualities, Amazon mistranscribed this word as "*moody*"/mu:di/, retaining the consonants but replacing the vowel with the closest alternative that creates a plausible word. Interestingly, Rev and Google both transcribed "*much*" in place of "*muddy*," correctly recognizing the word uttered as belonging to the STRUT lexical set despite the high rounded quality of the vowel [ʊ].

Another example of a Yorkshire word belonging to the STRUT lexical set that proved to be challenging for the ASR systems was "*haircut*," pronounced [ɛːkʊʔ], though this was likely due to the h-dropping that takes places in word-initial position. Google semi-successfully transcribed "*cut*," ignoring the first vowel in the word, while Amazon and Rev transcribed "*airport*" and "*accurate*," respectively. The lack of /h/ at the beginning of "*haircut*" had a clear impact on the words consequently transcribed, since both begin with a vowel. This seems to have then had an effect on the vowel transcribed in the second syllable, as these systems transcribed final syllables containing the vowels [ɔ:] or [ʊ] in SSBE.

4.4.3. FACE

Words belonging to the FACE lexical set are subject to realizational differences across the accents; the FACE vowel is realized as the diphthong [eɪ] in SSBE but as the long monophthong [e:] in WYE. Most words containing this vowel were transcribed correctly, e.g., "*rains*" and "*place*," despite the monophthongal quality of the vowel produced by the Yorkshire speaker. However, some occurrences of [e:] proved challenging. For example, the word "potatoes," pronounced [p(ə)te:ʔəz] with a glottal stop in place of the second alveolar plosive, was incorrectly transcribed as "*tears*," "*debt is*," and "*date is*" by Amazon, Rev and Google, respectively. While Google transcribes a word containing the correct vowel [eɪ] ("*date*"), the other systems transcribe words containing the vowels [ɛə] and [ɛ], which share similar vocalic qualities with the front mid vowel uttered by the speaker in terms of vowel height, frontness and steady state (or very little articulatory movement). Given that Rev and Google both transcribe words containing /t/ after the FACE vowel, it seems unlikely that the mistranscriptions are a result of the glottal stop, and are rather a direct result of the monophthongal realization of the FACE vowel.

4.4.4. GOAT

Words belonging to the GOAT lexical set vary in their phonetic realization across the two accents, such that the diphthong [əʊ] features in SSBE but a long monophthong features in WYE, which can be realized in a number of ways. Traditionally this was produced as a back vowel [o:] but it has undergone a process of fronting (Watt and Tillotson, 2001; Finnegan and Hickey, 2015) to [ɵː] for many younger speakers, including the two Yorkshire speakers in this study. Some words containing this vowel were transcribed without issue, such as "*own*" and "*go*," though it should be noted that the latter was relatively diphthongal in quality given the phonological environment: the following word "in" begins with a vowel therefore a [w]-like sound is inserted, leading to movement during the vowel and creating a sound much closer to the SSBE diphthong [əʊ].

Other words containing the fronted monophthong proved more challenging for the systems, such as "*drove*" which was mistranscribed as "*drew if*," "*do if*," and "*if*" by Amazon, Rev, and Google, respectively. Amazon and Rev replace [ɵː] with words containing the vowel [u:], an alternative long monophthong produced in a relatively similar part of the vowel space, followed by [ɪ] and the voiceless version of the labiodental fricative. Google omitted the GOAT vowel, transcribing only the word "if" in studio quality audio and deleting the word completely in the speech-shaped noise condition. The word "*road*," pronounced [ɹɵːḍ], was also mistranscribed by two of the systems as "word" (/wɜ:d/), whereby the central monophthongal quality of the vowel was retained but the height was slightly adjusted to give [ɜ:].

4.5. Post-editing

In order to assess the possibility of incorporating an ASR output into the transcription process, it is necessary to assess the effort required to transform the ASR output into a more accurate (verbatim) transcript. The best performing system, Amazon, was evaluated in terms of the frequency and types of errors produced, as well as the difficulty of error identification within the data. Deletion and insertion errors may be more easily detectable than substitution errors, as suggested by Papadopoulou et al. (2021), in many contexts; in principle, these errors should stand out as missing or extraneous when the transcriber listens to the audio, while substitution errors may be more challenging to identify, especially if closely resembling the speech sounds in the audio recording.

In studio quality, 20 deletions were produced for SSBE speech and 13 for WYE speech, and in both cases, more than half of the deletion errors involved the omission of filled pauses. The rest of the deletions involved function words, and in almost all cases the transcription remained relatively unchanged in terms of semantic meaning, e.g., "*I can't **really** remember*" → "*I can't remember*," or "*half an hour **something like that** depending on traffic*" → "*half an hour depending on traffic*." In the speech-shaped noise condition, 26 deletions were produced for SSBE and 16 for WYE. Fifty percent of the errors for SSBE involved filled pauses while the majority of WYE deletions (11/16) involved function words, and most deletions did not affect the semantic meaning of the utterance, e.g., "*except **for** when it rains*" → "*except when it rains*" or "*he's a tour guide **and** I knew **him** from secondary school*" → "*he's a tour guide I knew from [a] secondary school*." Furthermore, some of the deletions occurred in instances where a pronoun or determiner, e.g., "*I*" or "*a*," had been repeated, such that the transcript contained only one instance of each word.

Insertions were extremely rare within the data, particularly for Amazon which did not produce any insertions for SSBE and only inserted 1–3 words in the WYE transcripts. In studio quality, the only insertion to be made was "*I knew him from secondary school*" → "*I knew [him] from **a** secondary school*," which is easily detectable given that the insertion of the determiner sounds unnatural in this context. The same insertion was made in the SSN condition, along with the insertion of first-person pronoun "*I*" and determiner "*a*."

Substitutions may require more cognitive effort to identify, particularly in cases where the word in the transcript closely resembles the word that is uttered. First, the substitution of content words was assessed given that this type of mistranscription could lead to serious errors in forensic contexts, e.g., if a non-incriminating word such as "*gum*" is substituted with an incriminating alternative like "*gun*." In studio quality, six content words in SSBE and 16 in WYE were subject to substitution errors. The majority of SSBE substitutions in this case involved morphological alterations, such as a change in tense (e.g., "*finish*" → "*finished*") or omission of an affix (e.g., "*surnames*" → "*names*"). Due to the phonetic similarity of the target and transcribed word, these substitutions could be difficult to notice in a post-editing phase, and an uncorrected change in tense could, in some circumstances, have a significant impact on the meaning of the utterance. However, the morphological alterations in the data were all relatively clear; either the change in tense was held in stark contrast to the tense used in the rest of the utterance, or it was coupled with another error which would indicate that the section needs closer review.

TABLE 6 Examples from the data of substitution errors involving pronouns.

Accent	Reference transcript	Automatic transcript
SSBE	I couldn't put a name to **a** face	I couldn't put a name to **her** face
SSBE	I might have known **them**	I might have known **him**
WYE	**Uh** can get a bit inebriated	**You** can get a bit inebriated

Words involved in substitutions are highlighted in bold text.

The remaining two errors were relatively easy to identify from the context of the utterance; the utterance-final phrase "*I don't* ***know***" was mistranscribed as "*I don't* ***you***" and "*a really big* ***yew tree*** *right next to it*" was mistranscribed as "*a really big* ***utility*** *right next to it.*" A much bigger proportion (11/16) of the WYE content-based substitutions involved non-morphological alterations, but the majority of these were easy to identify from context alone, such as the phrase "*it's bit uh cut and chop with staff*" which was transcribed by Amazon as "*it's bitter cutting chocolate stuff.*" The words directly preceding this part of the utterance referenced the frequent hiring of new staff, therefore the reference to "*cutting chocolate*" seems misplaced in this context. Other WYE substitutions included "*airport*" in place of "*giving him an* ***haircut***" and "*moody*" in place of "*when it rains it gets very* ***muddy***."

In the speech-shaped noise condition, there were a very similar number of content-based substitutions in SSBE (5) while the number increased substantially for WYE from 16 to 29, only six of which involved morphological alterations. The rest of the errors were relatively clear from context, e.g., "*I had a bit of* ***dessert***" → "*I had a bit of* ***Giza***" when talking about lunch or "*did have a* ***sack of potatoes***" → "*did have a* ***sacrum tears***," making them easy to identify when comparing the audio recording and the ASR transcript, and potentially even from simply reading the transcript through without audio.

The substitution of function words could be more difficult to detect in some cases as short grammatical words are generally paid little conscious attention and glossed over in reading tasks (Van Petten and Kutas, 1991; Chung and Pennebaker, 2007), and the meaning of the utterance often remains unchanged. For example, there is little difference between "*go* ***in*** *get my drinks*" and "*go* ***and*** *get my drinks*" in the context of visiting a pub. Substitutions involving function words featured around 10 times in SSBE and 16 times in WYE in both audio qualities, and the majority of these were relatively inconsequential, e.g., "***the*** *steak house*" → "***a*** *steak house*" and "***that's*** *quite hard to see*" → "***it's*** *quite hard to see.*" However, a number of the errors involved the substitution of pronouns (see Table 6), which could be extremely difficult to notice due to similar pronunciations, but could be problematic within a forensic context if left uncorrected.

5. Discussion

5.1. ASR performance

The present study set out to investigate the reliability of ASR transcripts with simulated police interview recordings by exploring the impact of regional accent and audio quality on the transcription performance of three popular commercially-available ASR systems. Results revealed that the ASR system used and the audio quality of the recording had a significant effect on word error rate, and though regional accent was not found to significantly predict WER, clear differences were observed across the two accents in terms of the frequency and types of errors made.

5.1.1. ASR system and audio quality

With regards to the commercial ASR systems chosen for this study, Amazon Transcribe was clearly the best-performing system, consistently achieving the lowest WER in each condition: 13.9 and 15.4% for SSBE in studio quality and the speech-shaped noise condition, respectively, and 19.2 and 26.4% for WYE in studio quality and the speech-shaped noise condition, respectively. Google Cloud Speech-to-Text achieved the highest WER in almost every condition, and error rates for this ASR system were significantly higher than those for both Amazon and Rev, as well as consistently above 30%. Rev AI had the most variable performance, ranging from 19.0 to 42.5%. The patterns observed across accents and audio qualities were relatively consistent within each system, but the specific reason behind the difference in performance across systems is not clear, especially given the "black box" nature of proprietary automatic systems. The addition of speech-shaped noise to the audio recordings was found to have a significant effect on word error rate, leading to a higher frequency of errors in almost every condition. However, it must be noted that Amazon Transcribe, the best performing system, was the least affected by the addition of speech-shaped noise, with WERs increasing by only 1.5% in SSBE and 7.2% in WYE between the two audio qualities.

5.1.2. Regional accent

Word error rate was not found to be significantly impacted by regional accent in this study, although this was likely due to variation between speakers and the small sample size. A clear pattern emerged whereby one speaker of each accent was favored by the ASR systems, and performance for the best WYE speaker was roughly level with performance for the worst SSBE speaker.

Variation in system performance within an accent group has recently been investigated by Harrison and Wormald (in press), a study in which test data from a sociolinguistically-homogenous group was transcribed using Amazon Transcribe. Despite demographic factors such as age, accent and educational background as well as the content of the recordings being relatively controlled, a high level of variability was observed across speakers, with word error rates ranging from 11 to 33%. The variation across speakers observed in this study is therefore unsurprising, although the systematic effects of variety may emerge on a larger data set, as reported by Markl (2022).

Despite the lack of a statistically significant difference in WER across the accents, a higher number of errors were produced for the West Yorkshire English speech compared with the Standard Southern British English speech, and the majority of errors for the non-standard regional accent involved the substitution of words or phrases. Substitution errors can be extremely damaging in forensic contexts, particularly when the quality of the audio is poor. It is possible that deletion and insertion errors will be easier to identify alongside the audio within a transcript, but if the listeners have been "primed" by an alternative interpretation of a word or phrase (i.e., a substitution) then the identification of that error will in all likelihood be more challenging.

There are a number of factors likely contributing to the disparity in performance between accents. Modern ASR systems tend to involve two components, an acoustic model and a language model. Research on performance gaps between accent groups suggests that many ASR performance issues concerning "accented" speech stem from an insufficiently-trained acoustic model, which is caused by a lack of representation of non-standard accents in training data (Vergyri et al., 2010; Dorn, 2019; Markl, 2022). There were many errors within the Yorkshire data that can be attributed to a phonetic realization deviating from SSBE, a large number of which involved vowels for which phonemic and realizational differences are observed across the accents. Numerous errors were likely the result of a combination of vocalic and consonantal differences between SSBE and WYE; for example, the combination of h-dropping and a Northern realization of the STRUT vowel in "*haircut*" led to substantial substitutions by two of the systems.

Although the main focus of the fine-grained phonetic analysis was on errors seemingly caused by issues with the acoustic model, there were some errors that could not be attributed to acoustics and instead were likely a reflection of the language model. The language model calculates the conditional probability of words in a sequence, i.e., how likely is it that word *D* will follow on from words *A*, *B*, and *C*. Utterances containing non-standard grammar are therefore likely to cause problems for ASR systems, a few examples of which were observed in the Yorkshire data. The lack of a subject pronoun in the utterance "*did have a sack of potatoes in front*" led to the insertion of the pronouns "*I*" and "*you*" by Rev and Google respectively, both positioned after the verb "*did*." The omission of the determiner in the phrase "*in the front*" led to the insertion of the verb "*is*" before this phrase, i.e., "*is in front*," by both Rev and Google. Another example of an error likely resulting from the language model is the insertion of the indefinite article into the phrase "*from secondary school*," transcribed by Amazon as "*from **a** secondary school*." Having reviewed the audio, there is no phonetic explanation for this insertion given that the nasal [m] is immediately followed by the fricative [s], therefore this insertion is likely due to the language model calculating that the sequence of words including "*a*" is more probable.

5.1.3. Error analysis

A WER of 5% is generally accepted as a good quality transcript (Microsoft Azure Cognitive Services, 2022) but if the errors within that transcript lead to significant changes to the content, then that transcript could be harmful in a court of law. WER alone cannot indicate whether a system is good enough to use in a legal setting, such as the transcription of police-suspect interviews. Fine-grained phonetic analysis of the errors produced is a much more informative approach that can highlight any major issues with a system such as high rates of substitution errors. This type of analysis could also help to identify common issues in ASR transcripts that could subsequently be built into training for police transcribers, if a computer-assisted approach to police-suspect interview transcription was adopted. However, this method of analysis is extremely labor-intensive in nature and is therefore not feasible for large data sets. A combination of the two approaches, in which WER is calculated for a large data set and a subset of the data is subject to more detailed analysis of the frequency, type and magnitude of the errors, may be more suitable.

5.2. Post-editing

One of the aims of this paper is to investigate the possibility of incorporating automatic transcription into the production of police interview transcripts. The transcripts produced by the three commercial ASR systems in this experiment would not be suitable for use without manual correction, which is to be expected given that this is a commonly acknowledged issue in the field of automatic speech recognition (Errattahi et al., 2018). The question to be addressed is therefore whether the automatic transcripts could act as a first draft which is then reviewed and corrected by a human transcriber.

Gaur et al. (2016) found that editing an ASR output actually takes longer than producing a transcript from scratch once the WER surpasses 30%. Given that the average WER for Google exceeded 30% in every condition, and in all but one condition the WER for Rev was more than 29%, neither of these systems would be adequate for the purpose of producing a first draft of a transcript to be corrected by a human transcriber. In contrast, WERs produced by Amazon ranged from 13.9 to 26.4%, falling into the range of "acceptable but additional training should be considered" according to Microsoft Azure documentation (Microsoft Azure Cognitive Services, 2022). Gaur et al. (2016) found that participants benefitted from the ASR transcript provided the word error rate was low, i.e., below 30%. It is therefore possible that utilizing the Amazon transcripts as a first draft to be edited could reduce the time necessary to produce verbatim transcripts.

Closer inspection of the transcripts produced by Amazon revealed that many of the errors should, in principle, be easy to identify or would be relatively inconsequential if left uncorrected. For example, over 50% of the deletion errors in studio quality audio involved the omission of filled pauses like "*uh*" and "*um*," which is unlikely to have a substantial effect on the reader's perception of the speech and the speaker. Most deletions in speech-shaped noise audio involved short function words, and in almost all cases the meaning of the utterance was unaffected by their omission. Insertions were very rare within the data but were quite easily identifiable from context or were paired with a substitution error. The substitution of content words, particularly for the Yorkshire English speech, was generally evident from context since the resulting transcript was often ungrammatical or non-sensical, and substitution errors involving function words generally made no difference to the meaning of the utterance. The exception to this was the substitution of pronouns and content words with morphologically-related terms (though cases of the latter in this data were relatively easy to identify); these errors would likely be much harder to spot due to the phonetic similarity between the word uttered and the substituted term.

5.2.1. Potential challenges

A potential challenge with the task of correcting a transcript is that post-editors could be "primed" (i.e., heavily influenced) by the content of the ASR output to such an extent that errors go unnoticed. Research in the field of forensic transcription has found that seeing an inaccurate version of a transcript can cause people to "hear" the error in the audio (Fraser et al., 2011; Fraser and Kinoshita, 2021). However, the quality of audio recordings in forensic cases is often extremely poor and the speech is "indistinct," resulting in a reliance on top-down information such as expectations about the speech content (Fraser, 2003). In the case of police-suspect interviews, where the audio quality is often relatively good in comparison to forensic recordings, transcribers may be less susceptible to the effects of priming. It is also worth noting that many of the errors produced by the ASR systems were easy to identify from contextual knowledge or due to the non-sensical nature of the substitution. For example, one ASR transcript contained "*giving him an airport*" in place of "*given him an (h)aircut*" which, despite the similar phonetic content, is unlikely to influence a post-editor due to the implausibility of the utterance. Minor deletion errors, such as the omission of filled pauses, could be more challenging to identify in a transcript, though in many cases this would likely be inconsequential with regards to the readability of the transcript and the reader's perception of the speech and speaker.

Another potential issue is that errors in transcripts with a low WER may be more difficult to identify. As suggested by Sperber et al. (2016), post-editors may miss errors due to a lack of attention, and this effect would likely be increased in cases where the transcript is almost completely accurate and an excessive amount of confidence is placed in the performance of the automatic system. It is possible that the user interface employed could help to address this problem. Sperber et al. (2016) suggest two methods for focusing transcriber attention and therefore decreasing the chance of missing transcription errors: highlighting low-confidence words in red, and typing from-scratch with the ASR hypothesis visible. Both methods were shown to improve the quality of the transcript (i.e., decrease WER) and reduce the time taken, and it was also found that different strategies work best for different levels of WER. Retyping (with the ASR output visible) gave the best results for segments with a high WER, while editing the ASR transcript text gave the best results for low WER segments.

5.3. Future work

This study used a small sample of commercially-available automatic speech recognition systems and has shown that not all ASR systems are suitable for the task of producing a "first draft" transcript, as evidenced by the frequency of errors produced by Rev AI and Google Speech-to-Text. However, promising performance was demonstrated by one of the systems tested and further analysis of the errors suggests that post-editing an ASR transcript, provided it is of adequate quality, is a worthwhile topic to explore in the context of police-suspect interviews. This approach could facilitate the production of verbatim transcripts of interviews without a substantially higher time requirement than the current practice of summarizing the majority of the recording.

Future work on this topic should focus on two areas: ASR performance in a range of audio, speaker and speech conditions, and post-editing. In the present study, the addition of speech-shaped noise to the recordings may not have created an audio quality representative of real police-suspect interview data. It would therefore be interesting to use real recordings to investigate the capabilities of this technology. Other factors that may impact the system's performance and would be present in police-suspect interviews include different levels and types of background noise, multiple speakers, other regional accents, and longer stretches of speech.

More research is also required on the topic of post-editing. Papadopoulou et al. (2021) claims to be one of the first studies to employ qualitative analysis on automatic transcription errors and to evaluate the post-editing effort required in correcting ASR transcripts. Incorporating ASR outputs into the transcription process has been investigated by others, though these studies tend to focus on optimizing efficiency (Sperber et al., 2016, 2017) or simply report on the use of a computer-assisted transcription approach, e.g., for meetings of the National Congress of Japan (Akita et al., 2009) or for speeches in the Icelandic parliament (Fong et al., 2018). Transcripts have been found to be highly influential on the perception of speech content when the audio quality of the recording is extremely poor, but more research is required on the priming effects of ASR transcripts in the context of post-editing police-suspect interviews, i.e., on comparatively better quality audio. Furthermore, it is crucial to investigate the practicalities of correcting an ASR transcript of a police-suspect interview. For example, how many errors are missed by post-editors, and what are the consequences of leaving those errors in the transcript? How long does it take to correct an ASR transcript of a full police-suspect interview, and how does this compare to the current time taken to create ROTIs? Future research should explore these questions as the incorporation of automatic speech recognition into the transcription process could be extremely beneficial.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical approval was not required for the study involving human data in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required in accordance with the national legislation and the institutional requirements.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Funding

This work was supported by the Arts and Humanities Research Council (grant number: AH/R012733/1) through the White Rose College of the Arts and Humanities. Open access publication fees were covered by the York Open Access Fund at the University of York.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2023.1165233/full#supplementary-material

References

Akita, Y., Mimura, M. and Kawahara, T. (2009). "Automatic transcription system for meetings of the japanese national congress," in *Interspeech 2009* (ISCA), 84–87.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using 'lme4.' *J. Stat. Softw*. 67, 1–48. doi: 10.18637/jss.v067.i01

Boersma, P., and Weenink, D. (2022). *Praat: Doing Phonetics by Computer*. Available online at: http://www.praat.org/ (accessed November, 2022).

Bokhove, C., and Downey, C. (2018). Automated generation of 'good enough' transcripts as a first step to transcription of audio-recorded data. *Methodol. Innov*. 11, 205979911879074. doi: 10.1177/2059799118790743

Chung, C., and Pennebaker, J. W. (2007). The psychological functions of function words. *Soc. Commun*. 1, 343–359. doi: 10.4324/9780203837702

Coulthard, M. (2013). The official version: audience manipulation in police records of interviews with suspects. *Texts Practices* 16, 174–186. doi: 10.4324/9780203431382-16

DiChristofano, A., Shuster, H., Chandra, S., and Patwari, N. (2022). *Performance Disparities Between Accents in Automatic Speech Recognition*. arXiv [cs.CL]. Available online at: http://arxiv.org/abs/2208.01157

Dorn, R. (2019). "Dialect-specific models for automatic speech recognition of African American vernacular English," in *Proceedings of the Student Research Workshop Associated with RANLP 2019. Student Research Workshop Associated with RANLP 2019*. (Varna: Incoma Ltd.), 16–20.

Errattahi, R., El Hannani, A., and Ouahmane, H. (2018). Automatic speech recognition errors detection and correction: a review. *Proc. Comput. Sci*. 128, 32–37. doi: 10.1016/j.procs.2018.03.005

Feng, S., Kudina, O., Halpern, B. M., and Scharenborg, O. (2021). *Quantifying Bias in Automatic Speech Recognition*. arXiv [eess.AS]. Available online at: http://arxiv.org/abs/2103.15122

Finnegan, K., and Hickey, R. (2015). *Sheffield. Researching Northern English*. (Amsterdam; Philadelphia, PA: John Benjamins Publishing Company), 227–250. doi: 10.1075/veaw.g55.10fin

Fong, J. Y., Borsky, M., Helgadóttir, I. R., and Gudnason, J. (2018). *Manual Post-editing of Automatically Transcribed Speeches from the Icelandic Parliament - Althingi*. arXiv [eess.AS]. Available online at: http://arxiv.org/abs/1807.11893

Fraser, H. (2003). Issues in transcription: factors affecting the reliability of transcripts as evidence in legal cases. *Int. J. Speech Lang. Law* 10, 203–226. doi: 10.1558/sll.2003.10.2.203

Fraser, H. (2020). "Forensic transcription: the case for transcription as a dedicated branch of linguistic science," in *The Routledge Handbook of Forensic Linguistics*. Available online at: taylorfrancis.com (accessed October, 2022).

Fraser, H., and Kinoshita, Y. (2021). Injustice arising from the unnoticed power of priming: how lawyers and even judges can be misled by unreliable transcripts of indistinct forensic audio. *Crim. Law J*. 45, 142–152. Available online at: https://search.informit.org/doi/abs/10.3316/agispt.20210923053902

Fraser, H., Stevenson, B., and Marks, T. (2011). Interpretation of a Crisis Call: persistence of a primed perception of a disputed utterance. *Int. J. Speech Lang. Law* 18, 261. doi: 10.1558/ijsll.v18i2.261

Gaur, Y., Lasecki, W. S., Metze, F., and Bigham, J. P. (2016). "The effects of automatic speech recognition quality on human transcription latency," in *Proceedings of the 13th International Web for All Conference*. (New York, NY: Association for Computing Machinery), 1–8. doi: 10.1145/2899475.2899478

Gillick, L., Baker, J., Baker, J., Bridle, J., Hunt, M., Ito Y., et al. (1993). "Application of large vocabulary continuous speech recognition to topic and speaker identification using telephone speech," in *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2*, 471–474. doi: 10.1109/ICASSP.1993.319343

Godfrey, J., and Holliman, E. (1993). *Switchboard-1 Release 2 LDC97S62*. Philadelphia, PA: Linguistic Data Consortium.

Gold, E., Ross, S., and Earnshaw, K. (2018). "The 'west Yorkshire regional English database': investigations into the generalizability of reference populations for forensic speaker comparison casework," in *Interspeech 2018*. (Hyderabad: ISCA), 2748–2752. doi: 10.21437/Interspeech.2018-65

Hain, T., Woodland, P. C., Evermann, G., Gales, M. J. F., Liu, X., Moore, G. L., et al. (2005). Automatic transcription of conversational telephone speech. *IEEE Trans. Audio Speech Lang. Process*. 13, 1173–1185. doi: 10.1109/TSA.2005.852999

Harrington, L., Love, R., and Wright, D. (2022). "Analysing the performance of automated transcription tools for covert audio recordings," in *Conference of the International Association for Forensic Phonetics and Acoustics, July* (Prague).

Harrison, P., and Wormald, J. (in press). "Forensic transcription and questioned utterance analysis," in *Oxford Handbook of Forensic Phonetics*, eds F. Nolan, T. Hudson, and K. McDougall (Oxford: OUP).

Haworth, K. (2018). Tapes, transcripts and trials: The routine contamination of police interview evidence. *Int. J. Evid. Proof* 22, 428–450. doi: 10.1177/1365712718798656

Haworth, K. (2020). "Police interviews in the judicial process: police interviews as evidence," in *The Routledge Handbook of Forensic Linguistics*, eds M. Coulthard, A. May, and R. Sousa-Silva (London: Routledge), 144–158. doi: 10.4324/9780429030581-13

Hickey, R. (2015). "Researching northern English," in *Varieties of English Around the World, G55*, ed R. Hickey (Amsterdam: John Benjamins Publishing), 1–493.

Hughes, A., Trudgill, P., and Watt, D. (2005). *English Accents and Dialects: An Introduction to Social and Regional Varieties in the British Isles*. London: Atlantic Publications, Inc.

Jenks, C. J. (2013). Working with transcripts: an abridged review of issues in transcription. *Lang. Linguist. Compass* 7, 251–261. doi: 10.1111/lnc3.12023

Johnson, M., Lapkin, S., Long, V., Sanchez, P., Suominen, H., Basilakis, J., et al. (2014). A systematic review of speech recognition technology in health care. *BMC Med. Informat. Decision Mak*. 14, 94. doi: 10.1186/1472-6947-14-94

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., et al. (2020). Racial disparities in automated speech recognition. *Proc. Natl. Acad. Sci. U. S. A*. 117, 7684–7689. doi: 10.1073/pnas.1915768117

Kowal, S., and O'Connell, D. C. (2014). "Transcription as a crucial step of data analysis," in *The SAGE Handbook of Qualitative Data Analysis*, ed U. Flick (Thousand Oaks, CA: Sage), 64–79. doi: 10.4135/9781446282243.n5

Lima, L., Furtado, V., Furtado, E., and Almeida, V. (2019). "Empirical analysis of bias in voice-based personal assistants," in *Companion Proceedings of The 2019 World Wide Web Conference* (New York, NY: Association for Computing Machinery), 533–538. doi: 10.1145/3308560.3317597

Lindsey, G. (2019). *English After RP: Standard British Pronunciation Today*. Berlin: Springer.

Lippmann, R. P. (1997). Speech recognition by machines and humans. *Speech Commun.* 22, 1–15. doi: 10.1016/S0167-6393(97)00021-6

Littlefield, J., and Hashemi-Sakhtsari, A. (2002). *The Effects of Background Noise on the Performance of an Automatic Speech Recogniser. Defence Science and Technology Organisation Salisbury (Australia) Info.* Available online at: https://apps.dtic.mil/sti/citations/ADA414420 (accessed October, 2022).

Loakes, D. (2022). Does automatic speech recognition (ASR) have a role in the transcription of indistinct covert recordings for forensic purposes? *Front. Commun.* 7, 803452. doi: 10.3389/fcomm.2022.803452

Markl, N. (2022). "Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition," in *2022 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY: Association for Computing Machinery), 521–534. doi: 10.1145/3531146.3533117

Meyer, J., Rauchenstein, L., Eisenberg, J. D., and Howell, N. (2020). "Artie bias corpus: an open dataset for detecting demographic bias in speech applications," in *Proceedings of the Twelfth Language Resources and Evaluation Conference* (Marseille: European Language Resources Association), 6462–6468.

Microsoft Azure Cognitive Services (2022). *Test Accuracy of a Custom Speech Model*. Microsoft Azure Cognitive Services. Available online at: https://learn.microsoft.com/en-us/azure/cognitive-services/speech-service/how-to-custom-speech-evaluate-data?pivots=speech-studio (accessed February 2, 2023).

Mishra, A. (2021). *What is Rev AI's Accuracy?* Available online at: https://help.rev.ai/en/articles/3813288-what-is-rev-ai-s-accuracy (accessed January 23, 2023).

Mošner, L., Wu, M., Raju, A., Parthasrathi, S. H. K., Kumatani, K., Sundaram, S., et al. (2019). "Improving noise robustness of automatic speech recognition via parallel data and teacher-student learning," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* (Brighton), 6475–6479. doi: 10.1109/ICASSP.2019.8683422

Nolan, F., McDougall, K., de Jong, G., and Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *Int. J. Speech Lang. Law.* 16, 31–57. doi: 10.1558/ijsll.v16i1.31

O'Shaughnessy, D. (2008). Invited paper: automatic speech recognition: history, methods and challenges. *Pat. Recogn.* 41, 2965–2979. doi: 10.1016/j.patcog.2008.05.008

Papadopoulou, M. M., Zaretskaya, A., and Mitkov, R. (2021). "Benchmarking ASR systems based on post-editing effort and error analysis," in *Proceedings of the Translation and Interpreting Technology Online Conference* (Ashburn, VA: INCOMA Ltd.), 199–207. doi: 10.26615/978-954-452-071-7_023

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

Punch, K. F., and Oancea, A. (2014). *Introduction to Research Methods in Education.* Newcastle upon Tyne: SAGE.

Siniscalchi, S. M., and Lee, C.-H. (2021). "Automatic speech recognition by machines," in *The Cambridge Handbook of Phonetics*, *Cambridge Handbooks in Language and Linguistics*, eds R. A. Knight and J. Setter (Cambridge: Cambridge University Press), 480–500. doi: 10.1017/9781108644198.020

Sperber, M., Neubig, G., Nakamura, S., and Waibel, A. (2016). "Optimizing computer-assisted transcription quality with iterative user interfaces," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16).* (PortoroŽ: European Language Resources Association), 1986–1992.

Sperber, M., Neubig, G., Niehues, J., Nakamura, S., and Waibel, A. (2017). Transcribing against time. *Speech Commun.* 93, 20–30. doi: 10.1016/j.specom.2017.07.006

Stoddart, J., Upton, C., and Widdowson, J. D. A. (1999). *Sheffield Dialect in the 1990s: Revisiting the Concept of NORMs. Urban Voices: Accent Studies in the British Isles* (London: Longman), 72–89.

Stolcke, A., and Droppo, J. (2017). *Comparing Human and Machine Errors in Conversational Speech Transcription*. arXiv [cs.CL]. Available online at: http://arxiv.org/abs/1708.08615

Strycharczuk, P., López-Ibáñez, M., Brown, G., and Leemann, A. (2020). General Northern English. Exploring regional variation in the North of England with machine learning. *Front. Artif. Intell.* 3, 48. doi: 10.3389/frai.2020.00048

Szymański, P., Zelasko, P., Morzy, M., Szymczak, A., Zyła-Hoppe, M., Banaszczak, J., et al. (2020). *WER We Are and WER We Think We Are*. arXiv [cs.CL]. Available online at: http://arxiv.org/abs/2010.03432

Tatman, R. (2017). "Gender and dialect bias in YouTube's automatic captions," in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (Stroudsburg, PA: Association for Computational Linguistics), 53–59. doi: 10.18653/v1/W17-1606

Tompkinson, J., Haworth, K., and Richardson, E. (2022). "For the record: assessing force-level variation in the transcription of police-suspect interviews in England and Wales," in *Conference of the International Investigative Interviewing Research Group* (Winchester).

Tschäpe, N., and Wagner, I. (2012). "Analysis of disputed utterances: a proficiency test," in *Conference of International Association for Forensic Phonetics and Acoustics, August* (Santander).

Tüske, Z., Saon, G., and Kingsbury, B. (2021). *On the Limit of English Conversational Speech Recognition.* arXiv [cs.CL]. Available online at: http://arxiv.org/abs/2105.00982

Van Petten, C., and Kutas, M. (1991). Influences of semantic and syntactic context on open- and closed-class words. *Mem. Cogn.* 19, 95–112. doi: 10.3758/BF03198500

Vergyri, D., Lamel, L., and Gauvain, J.-L. (2010). *Automatic Speech Recognition of Multiple Accented English Data*. Available online at: www-tlp.limsi.fr; http://www-tlp.limsi.fr/public/automatic_speech_recognition_of_multiple_accented_english_data_vergyri.pdf (accessed January 20, 2023).

Walford, G. (2001). *Doing Qualitative Educational Research.* Bloomsbury: Bloomsbury Publishing.

Walker, A. G. (1990). "Language at work in the law," in *Language in the Judicial Process*, eds J. N. Levi and A. G. Walker (Boston, MA: Springer US), 203–244.

Wang, D., Wang, X., and Lv, S. (2019). An overview of end-to-end automatic speech recognition. *Symmetry* 11, 1018. doi: 10.3390/sym11081018

Wassink, A. B., Gansen, C., and Bartholomew, I. (2022). Uneven success: automatic speech recognition and ethnicity-related dialects. *Speech Commun.* 140, 50–70. doi: 10.1016/j.specom.2022.03.009

Watt, D., and Tillotson, J. (2001). A spectrographic analysis of vowel fronting in Bradford English. *Engl. World-Wide* 22, 269–303. doi: 10.1075/eww.22.2.05wat

Wells, J. C. (1982). *Accents of English*. Cambridge: Cambridge University Press.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., et al. (2016). *Achieving Human Parity in Conversational Speech Recognition.* arXiv [cs.CL]. Available online at: http://arxiv.org/abs/1610.05256

Zayats, V., Tran, T., Wright, R., Mansfield, C., and Ostendorf, M. (2019). Disfluencies and human speech transcription errors. *Proc. Interspeech* 2019, 3088–3092. doi: 10.21437/Interspeech.2019-3134

frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 24 July 2023
DOI 10.3389/fcomm.2023.1047799

OPEN ACCESS

EDITED BY
Helen Fraser,
The University of Melbourne, Australia

REVIEWED BY
Annina Heini,
Aston University, United Kingdom
Fabio Ferraz De Almeida,
University of Jyväskylä, Finland

*CORRESPONDENCE
Eero Voutilainen
✉ eero.voutilainen@helsinki.fi

SPECIALTY SECTION
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

RECEIVED 18 September 2022
ACCEPTED 28 February 2023
PUBLISHED 24 July 2023

CITATION
Voutilainen E (2023) Written representation of spoken interaction in the official parliamentary transcripts of the Finnish Parliament. *Front. Commun.* 8:1047799.
doi: 10.3389/fcomm.2023.1047799

Written representation of spoken interaction in the official parliamentary transcripts of the Finnish Parliament

Eero Voutilainen*

Department of Finnish, Finno-Ugrian and Scandinavian Studies, Faculty of Arts, University of Helsinki, Helsinki, Finland

In this article, I will analyze the written representation of spoken interaction in the official plenary session transcripts of the Finnish Parliament. The official parliamentary transcripts are not—and cannot be—identical copies of the original speech event. Instead, they are linguistically and textually edited in many ways. I will examine the different types of editorial changes that are made in the official Finnish parliamentary transcripts. These include phonological, morphological, and syntactic alterations, editing out of self-repairs, planning expressions, stuttering and slips-of-tongue, and finding written ways of expression for phenomena such as pauses, prosody, gestures, and non-verbal events. I will also discuss how the editorial changes affect the written representation of plenary session interaction.

KEYWORDS

parliamentary transcript, official transcript, parliamentary report, official report, verbatim report, entextualization, chronotope, conversation analysis

1. Introduction

The parliamentary plenary session is the highest decision-making organ in Finland where the Members of Parliament (MPs) oversee the acts of the government and discuss and decide on legislation, the national budget, and international agreements, among other topics (Finnish Parliament, 2023).[1] Speech has a central role in parliamentary democracy. Even the word *parliament* derives from the Latin communicative verb *parabolare* "to speak" (Etymonline, 2023, s.v. *parliament*). In Finland, the freedom of parliamentary speech is guaranteed in the constitution (§ 31). Since the very first sessions in the late nineteenth century, the discussion in the plenary session has been reported "verbatim" in the official plenary record.

In this article, I will analyze the written representation of spoken discourse in the official transcripts of the Finnish parliamentary plenary session. As is well known, the official parliamentary transcripts are not—and cannot be—identical copies of the original speech event. Instead, they are linguistically and textually edited in many ways. I will examine the central editorial changes which are made in the official Finnish parliamentary transcripts. I will focus on the changes that are foregrounded explicitly in the written guidelines of the

[1] The plenary session is the decision-making meeting of the parliament where the MPs debate publicly on political issues and decide on parliamentary matters in the plenary hall. The other major meetings include committee meetings, which prepare the matters for the plenary session, and meetings of parliamentary groups, where the political activities of the groups are planned. The meetings of committees and parliamentary groups are usually not open to the public (Finnish Parliament, 2023).

Records Office of the Finnish Parliament (Kirjo, 2021). I will also discuss how the editorial changes affect the written representation of plenary session interaction.[2]

This article proceeds as follows. In section 2, I present the data and methods of the study. In section 3, I introduce some key theoretical and practical perspectives in the making of the official parliamentary transcript. I focus on the genre of the parliamentary transcript, as well as the three central tensions which I consider prominent in the making and editing of the transcript. In section 4, I describe the process of producing the official Finnish parliamentary transcript. In section 5, I analyze the central linguistic and editorial practices in the making of the transcript. Finally, section 6 provides an overview and some discussion of the results of the study.

2. Data and methods

The data collected consist of digital video recordings of plenary sessions from 2008 to the present day and the official written records of the same period. During this time, the Finnish Parliament openly published all the plenary sessions online. To navigate the majority of the large dataset, I have used the annotated parliamentary corpus provided by the Language Bank of Finland (2019). The corpus includes the transcriptions of the plenary sessions from 10 September 2008 to 1 July 2016, aligned with the corresponding video recordings of the sessions with Automatic Speech Recognition (ASR) technology. From the newest material of the corpus to the present day, I have used public video recordings and the official transcripts published on the public website of the Finnish Parliament. To identify the most central, systematic practices, I have consulted the professional transcription manuals used in the Records Office of the Finnish Parliament (Kirjo, 2021), and my field notes which I have made since I began working in the Records Office of the Finnish Parliament at the beginning of 2010 (see below).

As my main method, I use conversation analysis (CA) which has been developed for analyzing the organization of social interaction in naturally occurring recorded data (Sacks et al., 1974; Heritage, 1984; see Sidnell, 2010; Sidnell and Stivers, 2013). More specifically, this study is contextualized with conversation analytical research on institutional interaction (Drew and Heritage, 1992; Heritage and Clayman, 2010). In my analysis of written transcripts, I also draw from genre analysis (Martin and Rose, 2008) and participant observation (Blevins, 2017). I have been a public servant in the Records Office of the Finnish Parliament since 2010, editing the official transcript according to the standing editorial principles and practices and deciding on those practices together with my colleagues.[3] My first-hand experience as a parliamentary editor has also helped me choose representative examples from the large corpora of written records and video recordings.

3. Making an official parliamentary transcript: Theoretical and practical perspectives

An official record, to which the official transcript belongs, is a formal account of what has taken place in an official chain of events, such as a meeting. In principle, a record resembles a memo in the sense that they both save and share information that is deemed important for an institution (see Guillory, 2004). Similarly to a memo, the record forms an essential part of the "official memory" of the organization (cf. Yates, 1989). The record serves the organization by choosing what to include and how to discursively formulate it. Practically, different records vary considerably, in terms of, for example, function (social aims), content (what is included in the record), structure (how the content is textually organized), and style (how the content is formulated and what kind of linguistic choices are made).

The plenary session record of the Finnish Parliament consists, roughly, of (1) technical sections which include presenting and declaring the conclusion of each matter on the agenda, and (2) a discussion under each topic. In this article, I focus on the transcript of the discussion excluding the technical sections. Quantitatively, records comprise ∼4,000–10,000 pages of technical sections and transcripts per year both online and in 4 to 10 large, printed volumes. The length of the transcripts depends on the discussion, which varies considerably between election periods, sessions, topics, and other factors.

The principles of transcription depend largely on how the genre of the transcript, and the record, is understood by its authors. *Genre* is usually seen as a schematic interactional category that directs both the production and interpretation of single texts. Genres are constantly evolving, as new texts are created (see Martin and Rose, 2008). Approaches to genre vary in whether they emphasize, for example, the macro-structure and linguistic features (Eggins and Martin, 1997) or social actions which are implemented in the texts (Devitt, 2004). In this section, I will focus on the social aims, target audiences, and key editorial principles of the transcript. In the next sections, I will concentrate on the transcription process (section 4) and the linguistic differences between the transcript and the plenary session (section 5).

The official parliamentary transcript in Finland serves at least three central **social aims**: (1) open mediation of public information (what the MPs say and how), (2) legitimatization of parliamentary decision-making (how the proposals are debated), and (3) preservation of nationally vital information for current and future generations. Plenary session transcripts may also be analyzed as official documents where public servants report parliamentary activities with official responsibility. The **target audiences** of the official parliamentary transcript may include, for example, citizens as a generalized group with supposed characteristics and requests, MPs, public servants who write and apply legal texts, researchers, and the media. The principles and practices of transcription are considerably affected by what target audiences are seen as primary.

[2] This article is partly based on my previous work in articles Tiittula and Voutilainen (2016) and Voutilainen (2016) in Finnish. However, the contents are thoroughly updated, with unpublished examples and analysis.

[3] It should be noted that I, as a member of the linguistic team in the Records Office, have also been involved in writing the editorial manual which I frequently cite in this article (Kirjo, 2021). However, the norms in the manual have been agreed on collectively, and they have general approval in the office.

For example, the treatment of the MPs as the primary audience might lead to editing transcripts heavily so that they meet the supposed or actual requests and intentions of the MPs. This could weaken the indexical connection between the transcript and the plenary session, or in other words, the authenticity and accuracy of the transcript. On the other hand, treating the citizens and the media as the primary target audience may lead to, for example, editing transcripts more lightly so that they convey both the content and the style of the speeches reliably and transparently to the reader because these matters are frequently focused on in public discussion.

According to the transcription manual of the Finnish Parliament, the transcription and editing practices have been consciously designed so that they mediate the plenary session to the reader as openly as possible and consider the many different purposes and target audiences of transcripts (Kirjo, 2021, p. 7). The official transcript has been regulated quite lightly from outside of the Records Office. The Parliament's rules and procedures state only as follows: "A record will be made of the plenary session, into which the proceedings of matters and discussions in the plenary session will be recorded. The speech transcribed in the record will be given to the speaker for verification. There can be no changes in the content of the speech." (§ 69). These guidelines are quite short and abstract, which means that the making of the transcript is largely based on self-regulation within the Records Office. This self-regulation is closely related to and affected by, among other things, the genre of parliamentary record with its social aims and conventions; the expected needs of the target audiences; the values, goals, guidelines, training, and culture of the transcribing community; and the personal preferences, ideals, and linguistic ideologies of the transcribers and editors of the transcript (Voutilainen, 2017).

The conversation in the plenary session is heavily regulated institutional interaction where the participants orient to the key features of the institution. These include (see Drew and Heritage, 1992; also Heritage and Clayman, 2010) as follows:

1) Institutional goals (e.g., deciding on the legislation, budget, and contracts of the country) and identities (e.g., the roles of MPs, ministers, chairpersons of committees, and government and opposition groups).
2) Social constraints (e.g., the chairperson of the parliament as the regulator of turn-taking, turn-types, and overall structure of the interaction).
3) Inferential frameworks and procedures (e.g., the institutional consequences of making proposals and seconding them in the conversation).

These institutional features of the plenary session are considerably reflected in the official transcript. The nature of the transcription is also naturally affected by the fact that the discourse in the plenary session is very heterogeneous. There are many genres of conversation (e.g., discussion about a law proposal, budget discussion, and question time) and several institutionally regulated turn-types (e.g., representation speech, group speech, "regular speech," comment, and interruption) with their own norms and expectations. Moreover, the topics, purposes, and target audiences of the speeches are manifold (see Bayley, 2004; Ilie, 2015, 2016, 2018). As a consequence, the transcribed discourse material is linguistically very diverse.

Transforming speech into written text is a highly complex linguistic activity. Spoken and written interaction are in many ways different as semiotic channels, concerning production, the product, and the reception. From a linguistic perspective, reproducing and mediating linguistic material from speech to writing may be analyzed from various angles, such as diamesic translation (Gottlieb, 2018) and entextualization (Park and Bucholtz, 2009), which are discussed later, and also recontextualization (Linell, 1998), reported speech (Holt and Clift, 2010), representation (Goodwin, 1994), repetition (Johnstone, 1994), replay (Merritt, 1994), recurrence (Gault, 1994), reformulation (Merritt, 1994), reanimation (Fairclough, 1992), paraphrase (Steiner, 1975), transformation (Eades, 1996), versioning (Potter and Wetherell, 1987), accounting (Rapley, 2001), and quoting (Haapanen, 2017) (on related concepts, see Rock, 2007).

Aside from mediation, the principles followed in the transcription and editing process may be approached analytically, for example, as genre-conscious language regulation (Tiililä, 2012). In genre-conscious language regulation, the transcribed plenary speeches are edited so that their original nature is preserved when presented in the official written record. The norms, expectations, and interpretational frames of the genre are treated as essential when editing the text, even though it might mean deviating from, for example, the norms of the written standard language. The parliamentary speeches include a considerable amount of regional, social, and situational linguistic variation which activates certain rhetorical and stylistic meanings. If all this variation were to be removed from the transcript, it would change the nature of speeches, and thus heavily loosen the indexical connection between the speech and the transcript. According to the transcription and editing guidelines of the Finnish Records Office, this would be seen as contradictory to the ideals of openness and transparency which are expected from the transcript (Kirjo, 2021, p. 8).

In practice, the genre-conscious language regulation of plenary session transcripts requires consideration of the normative expectations of two genres: the plenary session and the plenary record. These results in at least three central tensions in the transcription and editing process: (1) speech vs. writing, (2) authenticity vs. readability, and (3) linguistic variation vs. written standard. The first tension lies between speech and writing which are two different semiotic channels or modes: speech is acoustic sound waves in the air, whereas writing is a visual artifact. These semiotic channels have numerous considerable differences concerning, for example, communicative resources, production, reception, social status, and expectations (Ong, 1982; Biber, 1988; Halliday, 1989). Theoretically, in this article, I approach transcending this barrier as an *entextualization* process (see Bauman and Briggs, 1990; Park and Bucholtz, 2009): the individual turns-at-talk are decontextualized from their original context—face-to-face interaction in the plenary session—and recontextualized into the official plenary record, a written text artifact with its own institutional goals. This has some unavoidable consequences for the nature of the transcript: the speech is necessarily changed when transformed into written form. This creates tension in the connection between the parliamentary session and the parliamentary transcript. For example, since speech is received differently in written form, it may activate different interpretations, values, and attitudes in the reader. Essentially, transcription as a profession can be approached as a form of intersemiotic or diamesic translation (see Jakobson, 1959; Gottlieb, 2018) between two modalities within a single language.

The second tension in the transcription process is caused by the fact that the official parliamentary transcript aims to be both authentic—or reliable or accurate—as a report of a spoken interactional event and readable as a written text. Even though editorial changes in speech risk harming authenticity, some editing is usually treated as necessary so that the texts are easily readable and understandable for the readers, many of whom are most likely not trained in reading accurate scientific transcripts (such as conversation analytic transcripts which are used in this article). Some changes are also, paradoxically, necessary to avoid speeches and their reception from changing in the transcription process and to keep the experience as authentic as possible. This is demonstrated in section 5 of this article (e.g., changes in word order to compensate for the loss of prosody and tone in the transcription process). Authenticity means, in the Finnish Records Office, that the position of the reader is as similar as possible to the position of the member of the audience who is listening to the discussion in the plenary hall—the transcript should not be less understandable or less fluent than the speech event but also not more so (Kirjo, 2021, p. 7–8). Because of this, the removal of the multimodal situation, intonation, tone, pauses, and other non-verbal features frequently requires some intervention for the transcript to be readable. On the other hand, the complexity and ambiguity of the speech are usually left largely unchanged so that the overall experience of the speech is not harmed (see section 5 below).

The third tension between the session and the transcript lies between the naturally occurring linguistic variation in the speech and the norms of the written standard language. Because the plenary session transcript aims to represent the spoken interaction authentically, editing speech must consciously detach itself from the commonly written language bias—the view where written language is treated automatically as primary to spoken language and where linguistic features typical of spontaneous speech are often seen as secondary when they differ from their equivalents in written language (see Linell, 2005). Consciously breaking away from the written language bias means that the properties of spontaneous speech, such as regional, social, and situational variation, are not treated as mistakes from the perspective of written codified texts. It must be noted, however, that spoken face-to-face interaction and official written texts are generally met with different attitudes and expectations in the language community (Tiittula and Nuolijärvi, 2013). Because of this, the Records Office of the Finnish Parliament has decided to follow some conventions of the written standard. For example, self-corrections, stuttering, and planning expressions, which are typical of spontaneous speech, would probably draw more attention among the readers of the transcript than they would among the audience of the plenary session (Kirjo, 2021, p. 10). By standing out in the transcript, they might also activate different interpretations about the speech and the speaker, such as insecurity or incompetence, by readers who are not accustomed to reading transcripts of spontaneous speech (Kirjo, 2021, p. 10.).

Because the speeches are unavoidably changed in the transcript, they also evoke ethical and political considerations. All speeches discursively construct the social and political identity of the MPs. They are public performances that affect the way the MPs, as well as the views and groups that they represent, are interpreted by the recipients. If the transcribers or editors change the social reception and interpretation of the speeches, they change the relationship between the MP and the audience, which has possible political consequences for the MP. Because of this, the Finnish Records Office has made systematic and detailed guidelines for parliamentary transcription and editing, in order to treat all speeches systematically and equally regardless of who is speaking and who is transcribing the speech (Kirjo, 2021).

4. The transcription process

In the Records Office of the Finnish Parliament, there are 21 public servants who make the plenary session record.[4] The roles and activities of these people are as follows:

1. The document secretaries act as **transcribers** who produce initial drafts of the transcripts, listening to the audio record and using a regular keyboard and automatic speech recognition (ASR) software.[5]
2. The senior specialists act as **editors** who edit the initial transcripts based on the linguistic and editorial principles of transcription while listening to the record. The senior specialists also prepare the technical sections of the written record, such as the openings and closings of agenda items by the chairperson, vote results, and the decisions by the parliament.
3. The head of office or the leading specialist acts as the **responsible official** who revises the technical sections and publishes the finished transcripts online.
4. After the sessions, the senior specialists act as **post-editors** who revise whole transcripts, correct mistakes, and make systematic decisions in, for example, cases where the individual editors have made different orthographic choices concerning the same expression. After post-editing, the revised versions of the transcripts are published online.
5. Two of the document secretaries act as **desktop publishers** who prepare the layout and deliver the finished copies to the printing house.

Moreover, during the session, a senior specialist works in the plenary hall as a plenary session secretary. In this role, they make the necessary corrections to the automatically reported metadata of the session (e.g., names, turn types, and starting times of the

[4] In addition to this, two public servants in the Swedish Office make the transcripts of the Swedish plenary session speeches.

[5] The basic transcription work is done by listening to the audio which is transmitted through the microphones in the plenary hall. When needed, the editors also use the public video recording to access non-verbal activities in the plenary hall.

speakers) and report all the important activities in the session which are not automatically recorded by the microphones. These include interruptions and essential non-verbal communication by the MPs (see section 5).

The continuously updated metadata of the session form a roster for the transcribers and editors in the office with each speaking turn listed chronologically on individual lines. Transcribers (document secretaries) use this roster to reserve up to 10-min long shifts of whole speeches for transcription. They make an initial transcript where they apply some conventional orthographic and language-regulatory practices, following office guidelines. When the initial draft transcript is finished, the editor (senior specialist) edits it for publication by using the office's linguistic and editorial standards while listening to the audio record. The responsible official (the head of office or the leading specialist) publishes the transcript when single speeches and matters on the agenda have been edited. After the initial publication, the MPs have the right to suggest small alterations in their speech. The basic principle is that the alterations should be corrections to observable mistakes in the transcript, and they should not affect the content or overall style of the speech (Peltola, 2015).

The transcripts are edited delicately at the different stages. The initial transcript by the document secretary mediates the speech from the spoken to written mode of communication. Here, intonation, tone, pauses, inhalations and exhalations, slowly and quickly uttered words, quiet and creaky voices, and other paraverbal elements are removed, and orthographic features, such as punctuation and capital letters, are added. The mistakes made by the ASR program are corrected. Some features caused by the time-boundedness of the production of spontaneous spoken language are not transcribed, such as clear cases of planning expressions and self-corrections.[6] In addition, some elements of everyday speech, such as phonological features of dialect or everyday talk, are standardized in this phase (about all the changes mentioned here, see section 5). However, most of the changes are carried out in the editing phase.

Earlier in history, before audio recordings, the plenary session speeches were reported with pen shorthand (Kallioniemi, 1946). At that time, the speeches were changed considerably more both in the transcription and editing stages. The changes in transcription and editorial principles have probably been caused by at least the following key factors: First, the audio recordings and then the direct video broadcasts online have made it easy to compare the original session with the written transcript. Second, the parliamentary speech culture and language attitudes in parliament and the speech community have changed during recent decades in such a way that now documented spoken discourse by MPs does not have to be, or should not be, mechanically turned into standard written language. Following the same line of thinking, the speeches should not be stylized to be more of "higher style" or "better language" because it could remove socially or rhetorically relevant phenomena and thus harm the openness, accuracy, and authenticity of the transcript. Third, due to improvements in linguistic research, the editors of the transcript have more information about linguistic variation and its meanings in social interaction. The strong interest that the media and citizens have frequently shown toward parliamentary speeches in public discourse has reinforced the idea that the language of the transcripts should not be altered too much in the editing process. I will discuss these principles and practices in more detail in the following section.

5. Linguistic and editorial practices in the Finnish parliamentary transcripts

In this section, I will examine the linguistic and editorial practices in the Finnish Records Office by comparing the video recordings of the plenary sessions and the official written transcripts of the speeches (for the analysis of other parliamentary transcripts, see Cortelazzo, 1985; Slembrouck, 1992; Hughes, 1996; Mollin, 2007; Gardey, 2010, 2013; Treimane, 2011; Cucchi, 2013). I will focus particularly on the features which have been discussed in the guidelines of the Record Office (Kirjo, 2021) and the public presentations of the office (Voutilainen et al., 2013). In addition, I will analyze some other linguistic and interactional features which I see as central to the transcripts based on my comparative analysis.

The linguistic practices of the Finnish Parliament can be roughly divided into phonological, morphological, and syntactic transcription and editing strategies. In addition to this, there are explicit guidelines about transcribing and editing many elements of spoken language, such as self-corrections, planning expressions, slips-of-the-tongue, multimodal elements, interruptions, and administrative metadiscourse by the chairperson between official speeches.

Concerning **phonology**, the main practice followed in Finnish parliamentary transcripts is that non-standard regional, social, and situational variations are standardized (e.g., *sie* → *sinä* "you"; *kun* → *kuin* "than"). According to the editorial manual, this decision is based on the observation that this type of non-standard variation in Finnish is usually much more noticeable and is likely to draw more attention in writing (Kirjo, 2021, p. 9). Non-standard phonetic features might also make the transcript harder to read for people who are not accustomed to reading unedited transcripts (Tiittula and Nuolijärvi, 2013). An important exception to this rule is the retention of non-standard features which carry apparent rhetorical or stylistic meaning in the context. This can occur, for example, when an MP clearly uses single dialectal features as a rhetorical resource, such as the dialectal *Pyrsseli* instead of the standard Finnish *Bryssel* for "Brussels," to highlight the foreignness of the European Parliament to "ordinary" citizens in the provinces. A similar phenomenon occurs when an MP changes their style from formal to everyday style within the same speech when addressing a new audience.

Lexical choices are not usually changed in the transcript, even though the MPs might use rare, low-register, or slang words. The reason for this principle is that these words are neither particularly hard to read nor do they, arguably, activate different interpretations

[6] By time-boundedness, I mean that the final product of spontaneous speech and the temporal production process of speaking are inseparably intertwined. This means that, for example, traces of real-time planning and self-corrections are observable in the speech (see Hakulinen et al., 2004, p. 24–25). In this respect, spontaneous speech differs essentially from much of the written communication where such features are not visible in the final text.

in the transcript. This principle is not followed in the transcripts of all parliaments. In the House of Commons in the UK, for example, it has been a conventional practice to change certain everyday compound verbs to their high-prestige, single-word equivalents (e.g., *look at* → *consider; make sure* → *ensure; have to* → *must*; see Mollin, 2007). In a similar fashion, some hedges concerning the certainty of a statement have been removed in the transcripts of the European Parliament (e.g., *I think, of course*; Cucchi, 2013).

Morphological elements are met with situational consideration. Some morphological features typical of spoken language are systematically changed into written standard language, such as governance in nouns and verb forms (e.g., *merkitys johonkin* "meaning to something or someone" → *merkitys jollekin* "meaning for something or someone"; *vaikuttaa jollekin* "make an impact in something" → *vaikuttaa johonkin* "make an impact on something"). Many others are transcribed as they are, such as morphological passive in the second person plural which is a non-standard feature in spoken Finnish [e.g., *me mennään* (cf. *me menemme*) "we go"]. Changes are made in cases where the non-standard variant would, according to the editor, activate interpretations and attitudes in the transcript that would not arise in spoken communication (Kirjo, 2021, p. 9). Otherwise, non-standard variants are left to indicate different rhetorical and stylistic choices in the official record.

In **syntactic structures,** the editors favor relatively light editing. For example, cases of atypical word order or complex clauses which would draw attention in standard prose are often left unedited. Generally, they are edited if they are seen as considerably harmful to readability or they give rise to a stylistically different interpretation in writing (Kirjo, 2021, p. 10). A clear exception is formed by different processive structures which are most likely caused by the time-boundedness of spontaneous speech (Hakulinen et al., 2004, p. 25–25; see footnote 5 mentioned earlier). These types of syntactic structures are usually edited in the transcript as follows (1)[7]:

(1) 19th December 2019; 5:44 pm
Original speech[8]

ja tämä on se linja =ja tänä vuonna tulee
and this is the line =and this year will
tuo .hh todennäkösesti olemaan
that .hh probably be
aika paljon alhaisempi tuo käyttö .hh aste
pretty much lower that usage .hh rate
kun kun (.) viime vuonna oli.
than than (.) was last year.
eli (0.4) kyllä me kestäviä (0.4) olemme.
so (0.4) indeed we are (0.4) sustainable.

Official transcript

Tämä on se linja, ja tänä vuonna tuo käyttöaste tulee
This is the line, and this year that usage rate will
todennäköisesti olemaan aika paljon alhaisempi
probably be pretty much lower
kuin viime vuonna oli. Eli kyllä me kestäviä olemme.
than it was last year. So indeed we are sustainable.

In example 1, the syntactic cleft structure *tänä vuonna tulee* ***tuo*** *todennäköisesti olemaan aika paljon alhaisempi* ***tuo käyttöaste*** "**that** will probably be pretty much lower **that usage rate**" has been edited so that the initial pronominal noun phrase (NP) is removed, and the latter, lexical NP is moved in its place as the subject of the clause: *tänä vuonna tuo käyttöaste tulee todennäköisesti olemaan aika paljon alhaisempi* "this year that usage rate will probably be pretty much lower." This way, the utterance no longer has a subject which is split and placed at the beginning and end of the clause. The first part of the finite verb form (*tulee olemaan* "will be") has also been moved after the subject, which is seen as the neutral, non-emphatic word order in written standard Finnish (see Hakulinen et al., 2004, § 1366). Without this editorial choice, this part of written speech might appear more scattered and sporadic than it does in spoken language where it is quite common, unlike most written genres (Hakulinen et al., 2004, § 1064). However, the edited version might seem more polished and straightforward to some readers. Here, the editorial choice can be seen as favoring readability and the usual conventions of written standard prose.

In transcripts, **self-corrections and planning expressions** are usually edited out, unless they are commented on in the session by the speaker or by another participant (Kirjo, 2021, p. 10). In self-corrections (see Schegloff et al., 1977), the corrected expression is removed, and the final linguistic choice by the speaker is left in the transcript. The reason for this is that, while in speech, the corrected elements cannot be removed afterward and corrections are frequent, in text, they would attract more attention and possibly activate social interpretations about the speaker that would not be made while listening to the speech, such as unfocused, uncertain, or unskilled in the matter at hand. This is illustrated in example 2, where the word form *työllistämiskorvauksiin* "employing benefits" and the following repair initiator *tai* "or" are edited out and the following word form *työttömyyskorvauksiin* "unemployment benefits" is left in the transcript. The word searches (see Schegloff et al., 1977, p. 363) before the self-correction (*työttömyys- työ-* "unemployment- unemp-") are also excluded from the transcript as follows:

(2) 7th February 2018; 2:13 pm
Original speech

ja tässäkin kuten ministeri Lindström
and here too as minister Lindström
omassa puheenvuorossa lopetti tämän
in his own speech ended this
esityksen tähän että (0.4) kahdeksantoista
presentation in this that (0.4) eighteen
tuntia siellä on se raja (0.4) ja
hours is the limit there (0.4) and
siitä rupeaa kertymään sitten se (.) kaikki
from there will then start building up (.) all the
la- vaa- vaadittavat työttömyys- (.)
le- re- required unemployment- (.)
yhh. työ- tähän (0.6) öö
yhh. unemp- into this (0.6) uhm
työllistämiskorvauksiin tai

[7] The Conversation Analysis transcription conventions used here are based on Jefferson (2004); see also Hepburn and Bolden (2013).

[8] The English translations of the Finnish speech examples are intentionally literal, so that the reader gets an accurate view of the data.

employing benefits or

`(0.2) työttömyyskorvauksiin`

(0.2) unemployment benefits

`(.) öö tulevat öö nämä (0.2) rahamäärät`

(.) uhm incoming uhm these (0.2) sums of money

`alkavat sieltä kertyä.`

will start to build up from there.

Official transcript

Kuten ministeri Lindström omassa puheenvuorossaan lopetti
As minister Lindström, in his speech, ended
tämän esityksen, niin 18 tuntia siellä on se raja, ja siitä
this presentation, 18 hours is the limit there, and from there
rupeaa kertymään sitten se kaikki vaadittava, nämä
will then start building up all that is required, these
työttömyyskorvauksiin tulevat rahamäärät alkavat
sums of money that come to the unemployment benefits start to
sieltä kertyä.
build up from there.

At the beginning of the example above, the expression *ja tässäkin* "in here too" is also excluded from the transcript, possibly as a so-called "false start" where the speaker is interpreted as discontinuing the initial formulation and replacing it with another (here, *kuten ministeri Lindström...* "like minister Lindström..."). In the Records Office of the Finnish Parliament, these instances are often also treated as self-corrections, and the latter formulation is included in the official transcript.

Sometimes interpreting an expression as a "false start" might be open for debate. For example, the aforementioned expression *ja tässäkin* "in here too" might be interpreted in some contexts as connecting the utterance to something prior in the speech. Nonetheless, in example 2, the editor has interpreted it as self-correction. On the other hand, the difference between "false starts" and other expressions that are left incomplete is sometimes difficult to make. As a rule of thumb, short expressions that the speaker leaves incomplete and which do not carry much meaning according to the editor are interpreted as "false starts" and thus self-corrections. If the discontinued expression is longer and is interpreted as relevant to the speech, it is included in the transcript. If the unfinished utterance cannot be completed or connected to a neighboring utterance with very light and neutral editing (e.g., by changing the word order, adjusting inflection, or adding a grammatical word without changing the meaning of the utterance), its ending is marked with an ellipsis (...) (Kirjo, 2021, p. 135–136).

In planning expressions, the evident cases are removed based on the same practice. In plenary speeches, these include, for example, particles *niinku* "like" and *tota* "kind of" and hesitation markers such as *mm, öö,* and *ee.* In example 3, the planning particle *tota* "like" is left out from the transcript.

(3) 12th February 2020; 2:38 pm

Original speech

`myös vasemmistoliitto .hh öö(0.2)`
`tervehtii`

also the Left Alliance `.hh` uhm (0.2) greets

`ilolla tätä hallituksen (.)`

with joy this government's (.)

`esitystä =ja (.) ja antaa kaiken tukensa`

proposal =and (.) and gives all its support

`ministeri Kiurulle ja hallitukselle`

to minister Kiuru and the government

`siihen että tämä työ saadaan .hh hyvin`

for that this work is get `.hh` well

`tehtyä loppuun ja .hh ja tota`

finished up and `.hh` and like

`henkilöstömitotus nolla pilkku seitsemän`

the personnel requirement zero point seven

`sitovaksi lakiin`

as binding in the law

Official transcript

Myös vasemmistoliitto tervehtii ilolla tätä hallituksen esitystä ja
Also Left Alliance greets with joy this government's proposal and
antaa kaiken tukensa ministeri Kiurulle ja
gives all its support to minister Kiuru and
hallitukselle siihen, että tämä työ saadaan hyvin tehtyä loppuun
for that this work is get well finished up
ja henkilöstömitoitus 0,7 sitovaksi lakiin.
and the personnel requirement 0,7 as binding in the law.

In addition to the planning particle *tota* "like," there are also a hesitation marker (*öö* "ehm") and two instances of non-emphatic repetition (*ja ja* "and and") in the example. Both have been edited out of the transcript so that these frequent processing expressions of spontaneous speech do not attract special attention or activate different interpretations in written form.

The practices that self-corrections and planning expressions by the MPs are not included in the transcript may have a few effects on the official record. First, it can be said that these features which are probably caused by the time-boundedness of speech (see footnote 5 above) would be likely to evoke a different, possibly less formal impression of the speech and the speaker. Second, the exclusion of self-corrections and planning expressions may make the transcribed speeches more prepared and literal in style (Slembrouck, 1992). Third, both self-corrections and planning expressions make visible how the speaker constructs a turn-at-talk. By removing them, the editors of the transcript exclude features of real-time turn-design and linguistic processing of the speaker. Moreover, self-corrections, in particular, reveal the norms of the institution by correcting non-normative linguistic actions and formulations (see Drew, 2013). When editors exclude self-corrections, they remove traces of possible non-normative actions and formulations which the speaker corrected in the session. However, they are not erased in the transcript when someone reacts to corrections or corrected parts of speech (Kirjo, 2021, p. 10). Removing self-corrections and the corrected elements would make the reactions of another speaker impossible to understand for the reader. On the other hand, if the reactions were removed, it would considerably harm the reliability of the transcript.

A similar convention has been extended to so-called innocent **blunders**, or **slips-of-the-tongue**, which are usually corrected in the transcripts. This is presented in example 4 where the apparent blunder *ilmastointimuutos* "air-conditioning change" has been corrected to *ilmastonmuutos* "climate change."

(4) 11th October 2012; 4:56 pm

Original speech

arvoisa puheenjohtaja (.) on muistettava
honorable chairperson (.) it must be remembered
että rannikko .hh kunnissamme ei tulvi
that in our archipelago .hh municipalities it doesn't flood
mitään s- tsunamia(0.8) .hh vaan tavallinen
any s- tsunami (0.8) .hh but ordinary
vesi (.) josta (0.2) .hh joista ja
water (.) where (0.2) .hh from rivers and
jokien valuma-alueilta (.) yläjuoksuilta.
the catchment areas of rivers (.) from upper reaches.
(0.6) .hh ilmastointimuutos ei ole (.)
(0.6) .hh air-conditioning change is not (.)
myöskään mikään rannikkoväestön syytä .hh
the blame of archipelago people either .hh
tulisiko hallitus näin ollen pyrkiä
should the government therefore strive
kustannusten jakamiseen .hh
to the division of costs .hh
niiden osapuolten välillä jotka johtavat
between those parties who lead
valuam- valumavesi- vesien vesistöihin
draining- drainagewater- waters' water systems

Official transcript

Arvoisa puheenjohtaja! On muistettava, että
Honorable chairperson! It must be remembered that
rannikkokunnissamme ei tulvi mitään tsunamia
in our archipelago municipalities it won't flood any tsunami
vaan tavallinen vesi joista ja jokien valuma-alueilta
but ordinary water from rivers and catchment areas of rivers
yläjuoksuilta. Ilmastonmuutos ei ole
from upper reaches. Climate change is not
myöskään mikään rannikkoväestön syy.
the blame of archipelago people either.
Tulisiko hallituksen näin ollen pyrkiä
Should the government therefore strive
kustannusten jakamiseen niiden osapuolten välillä,
to the division of costs between those parties
jotka johtavat valumavesiä vesistöihin?
who lead drainage waters to water systems?

In Finnish, the difference in the formulation is small, and the forms can be easily mixed. The difference in meaning, however, is considerable and may be a cause of unintended humor. Moreover, the MP in the example frequently uses both Finnish and Swedish in speeches, and it is apparent that Finnish is not his mother tongue. Slip-of-tongue is a good example of a phenomenon that is emphasized in the written text but might even pass unnoticed by the participants of the speech event. The same can be said to apply to **stuttering** and **word searches** which are also by rule edited out of the official Finnish parliamentary transcript. In example 4, there are a few cases of these phenomena (*s- tsunamia* "s- tsunami," *valuam- valumavesi- vesien* "draing- drainage water- waters").

In Finnish parliamentary transcription, the same principle that applies to self-corrections and planning expressions is applied to stuttering and slips-of-the-tongue: they are corrected only if the participants do not explicitly react to them in the session (see Kirjo, 2021, p. 10). However, it should be noted that the difference between slips-of-the-tongue and incorrect knowledge might be hard to distinguish. This is apparent with, for example, wrong figures, names, and citations which might be the cause of either a slip-of-the-tongue or wrong information. The editorial guidelines (Kirjo, 2021, p. 10) state that if the mistake is clearly caused by wrong or incomplete information, there will be no correction in the transcript because the mistake is the MP's responsibility. Correcting MPs' wrong information could be easily seen as contradictory with an openness which is mentioned as a key value in the strategy of the Finnish Parliamentary Office (Parliamentary Office, 2019).

One essential category of editorial changes in the transcript is formed by different non-verbal features of parliamentary speech. The removal of **prosody**, for example, which unavoidably happens in the written transcript, might lead to a change of meaning in the range of certain particles and adverbs if the word order remains unchanged. This is illustrated in example 5 as follows:

(5) 7th September 2021; 6:38 pm

Original speech
rajoitusten purkaminen on mielestäni
Dismantling restrictions is in my view
myös perusteltua (0.4)
also justified (0.4)
rokotuskattavuuden kannalta.
considering vaccination coverage.

Official transcript

Rajoitusten purkaminen on mielestäni perusteltua
Dismantling restrictions is in my view justified
myös rokotuskattavuuden kannalta.
considering also vaccination coverage.

In example 5, the particle *myös* "also" refers, by virtue of the emphasis and the pause, to the phrase *rokotuskattavuuden kannalta* "considering vaccination coverage" and not to the word *perusteltua* "justified" which immediately follows. This emphasis is removed when the speech is transcribed, which directs the reference incorrectly to the word *perusteltua* "justified." To preserve the original reference, the editor has changed the word order in the sentence; the particle has been moved right before the phrase to which it refers.

Pauses are usually not explicitly marked in the transcript. However, where they have been identified as having rhetorical significance pauses in the speech have been indicated, for example, with a dash or a full stop and a change of sentence in the transcript. The use of typography with dash is presented in example 6 as follows:

(6) 31st March 2022; 4:53 pm
Original speech

olisi hienoa (0.4) että ottaisimme sen
it would be great (0.4) that we would take the
kannan (0.2) että ihan <oikeasti> (0.6)
stance (0.2) that quite <really> (0.6)
arvostamalla hoitajaa ja antamalla hänelle
by appreciating the nurse and by giving them
kunnon ((puhemies koputtaa nuijalla))

a decent ((the chairman knocks with the gavel))
palkan (.) me saamme heitä lisää
salary (.) we get more of them

Official transcript

Olisi hienoa, että ottaisimme sen kannan, että
It would be great that we would take the stance that
—ihan oikeasti—arvostamalla hoitajaa ja antamalla hänelle
— quite really — by appreciating the nurse and by giving them
kunnon palkan [Puhemies koputtaa] me saamme heitä lisää.
a decent salary [The chairman knocks] we get more of them.

In the above mentioned example, the dash is used to indicate two rhetorically relevant pauses which, together with pronouncing the keyword noticeably slower than surrounding speech, form an emphatic parenthetical structure inside the ongoing subordinate clause. Following this structural interpretation, the particle *että* "that" is moved to precede the parenthesis, even though the MP utters it after the first pause. This editorial decision highlights the rhetorical choice by typographical means.

Similarly to prosody, **multimodal elements** of the interaction, such as gestures, gazes, and movements, as well as non-verbal actions and events, are unavoidably erased when the speech in face-to-face interaction is represented in writing. Because of this, the editor includes the multimodal elements in the transcript that the plenary session secretary has made a note of during the session. The editor, in the next work phase, removes the ones that they do not consider necessary for comprehending the speech in the same way as the participants do in the plenary hall. These elements are included in square brackets within the transcript in the place where they occur (Kirjo, 2021, p. 53–54). This is demonstrated in example 7 as follows:

(7) 2nd June 2015; 2.51 pm
Official transcript

Arvostamani pääministeri Sipilä, te olette tässä kuvassa
prime minister Sipilä who I appreciate, you are in this picture
opiskelijan kanssa, tekstinä "Koulutuksesta ei leikata",
with a student, with text "No cuts from education",
vieressä ministeri Stubb. [Puhuja näytti kuvaa][9]
next to minister Stubb. [The speaker showed a picture]

In the example, the MP refers to an artifact with an NP *tässä kuvassa* "in this picture." The pronoun *tässä* "in this," in this case, refers to the material context of the session. The editor has interpreted this deictic reference as an expression that requires an explanation for the reader of the transcript. To address this issue, the editor has added a description in brackets.

In addition to the features that have been described earlier, there have traditionally been a few other editorial decisions that have had a noticeable effect on how the interaction is presented in the official transcript. The **interruptions**, or **interjections**, made by the MPs are transcribed in square brackets in the transcript in a similar way to the multimodal elements that were described earlier. The interruptions are not an official part of the plenary session discussion, but they are passively tolerated in the session and routinely included in the transcript when someone reacts to them or when they are otherwise seen by the editors as essential to the session (Kirjo, 2021, p. 34–35). Example 8 shows an interruption that has been included in the official transcript as follows:

(8) 19th September 2017; 2:31 pm
Original speech

MP: joo =arvoisa (0.4) puheenjohtaja (1.0)
yeah =honorable (0.4) chairperson (1.0)
täällä on käyny esille
it has turned out here
[se että nuorten (1.4)
[that young peoples' (1.4)
I[10]: [puhemies
[chairman
MP: puhemies (.) nii vielä toistaseks
chairman (.) yes still for the time being
=kiitos (0.8) elikkä tota (0.4)
=thank you (0.8) so, like, (0.4)
tääll on käyny ilmi se
it has turned out here
että nuorten alkoholinkäyttö
that alcohol consumption by young people
on vähentyny (.) tuo- näinä vuosina
has decreased (.) tu- during these years

Official transcript

Arvoisa puheenjohtaja! [Eduskunnasta: Puhemies!]
Honorable chairperson! [From the parliament: Chairman!]
—Niin, puhemies vielä toistaiseksi, kiitos!—
—Yes, chairman still for the time being, thank you!—
Täällä on käynyt ilmi se,
It has turned out here
että nuorten alkoholinkäyttö
that the alcohol consumption by young people
on vähentynyt näinä vuosina.
has decreased during these years.

In the example, the MP starts his speech with the form of address *arvoisa puheenjohtaja* "honorable chairperson," which is common as an official form of address in other meetings in Finland, but in parliament, the official formulation is *arvoisa puhemies* "honorable chairman." After the form of address, the MP manages to utter a few words before there is an interruption in the overlapping speech by another MP from the plenary hall, correcting him with the official formulation. The MP who has the floor interrupts his speech and reacts by repeating the official formulation, confirming that it is the correct formulation at the moment (*nii vielä toistaseks* "yeah still for the time being") and thanking the other MPs (*kiitos* "thank you"). He then, after a pause, continues with the speech by repeating the utterance which was interrupted. In the official transcript, the interruption has been included in the transcript in square brackets. The plenary session secretary has not confirmed the identity of the MP who made the interruption, so it has been marked with the source expression

[9] In the earlier data, the explanations in the brackets are in past tense, whereas in the current data they are in present tense.

[10] I, in this transcript, stands for an interruption from another MP in the plenary hall.

Eduskunnasta "from the parliament." The reaction by the MP who has the floor is separated from the surrounding speech with dashes. The interruption has been moved to directly follow the form of address that it comments upon, and the interrupted talk is removed probably because the MP repeats it after his reaction to the interruption. In other cases, interrupted talk is usually marked in the transcript with an ellipsis (. . .).

In the Finnish official transcript, interruptions that are not heard properly by the plenary session secretary, are not audible on the digital record, and are not reacted to by any MP, are often excluded from the transcript, unless it is seen as important to convey that the speech caused a commotion in the session (Kirjo, 2021, p. 34–35). Moreover, especially when there is a considerable number of interruptions, an interruption is often left out when an editor has interpreted it as of little importance to the session (e.g., supporting chants like *hyvä* "good" or *juuri näin* "just like that" from the same parliamentary group). In these cases, the editor decides that not including all the interruptions and the harm that it does to the authenticity, or accuracy, of the transcript is "a lesser evil" than the harm that would otherwise occur to the readability of the transcript. Moreover, since these types of interruptions are quite frequent in the session and only some of them are caught by the plenary session secretary, the ones that are caught could be seen as getting a disproportionate weight when transcribed in the report (Kirjo, 2021, p. 34–35). Similarly, the **parallel discussions** which take place in the plenary hall between MPs at the time of the session are left out of the transcript as a matter of routine.

For the sake of readability and to retain the authentic impression when mediating speech into writing, some of the **utterance-initial particles** have been removed when they have been considered as not having a special function, rhetorical weight, or stylistic significance (e.g., *ja* "and," *mutta* "but," *no* "well," and *eli* "so") (Kirjo, 2021, p. 147). These discourse markers often connect utterances in spontaneous speech. If all of them were included in the transcript, they would create considerably long compound clauses which would most likely negatively affect readability and create a very different impression for the reader than they do in speech. However, research on particles (Sorjonen, 2001; Heritage, 2015) shows that they can have a large array of significant interactional functions. When editors identify an utterance-initial particle as having functional relevance in the transcript, they do not remove them. A case where utterance-initial particles have been included in the transcript can be seen in example 9 as follows:

(9) 14th September 2016; 4.07 pm
Original speech

arvoisa puhemies (1.2) vastaan (0.8) hh.
honorable chairman (1.2) I answer (0.8) hh.
yhteen kysymykseen joka tuli usealta
one question which came from multiple
(0.4) taholta tässä =elikkä (0.6) ja yritän
(0.4) sources here =so (0.6) and I try
olla lyhytsanainen (2.2) kysyttiin
to be brief (2.2) it was asked
sitä että miten lainvalmistelussa
how in legislation could one
paremmin voitasiin perustuslaki- (0.4) ja
better the matters concerning the constitution (0.4) and
säätämisjärjestysnäkökohdat ottaa huomioon
the legislative proceedings take into account
(1.6) no (0.4) ensinnäkin näen niin
(1.6) well (0.4) first I see so
että (0.6) ministeriön (1.2) omat
that (0.6) ministry's (1.2) own
lainvalmistelijat (.) omat virkamiehet
legislators (.) own officials
on avainasemassa tässä (.) että
are in key position here (.) so that
ministeriöissä on riittävä (0.6)
the ministries have sufficient (0.6)
perustuslain tuntemus
knowledge of the constitution

Official transcript

Arvoisa puhemies! Vastaan yhteen kysymykseen, joka tuli
Honorable chairman! I answer one question which came
usealta taholta tässä—ja yritän olla lyhytsanainen.
from multiple sources here—and I try to be brief.
Elikkä kysyttiin sitä, miten lainvalmistelussa
So it was asked how in legislation
paremmin voitaisiin
could one better
perustuslaki- ja säätämisjärjestysnäkökohdat
constitution and legislative proceedings matters
ottaa huomioon. No, ensinnäkin näen niin, että
taken into account. Well, first I see that
ministeriön omat lainvalmistelijat, omat virkamiehet,
ministry's own legislators, own officials,
ovat avainasemassa tässä, että ministeriöissä
are in key position here, so that the ministries
on riittävä perustuslain tuntemus. [. . .]
have sufficient knowledge of the constitution. [. . .]

In this example, utterance-initial particles *elikkä* "so" and *no* "well" have been included in the transcript. This means that the editor has interpreted them both as having relevant functions in the speech, besides connecting utterances. First, after saying that he will answer one question that was posed by many people, the speaker proceeds with *elikkä* to report the question. Here, the particle marks the following utterance as a conclusion from the previous utterance and also provides a shift to the next action (see Hakulinen et al., 2004, § 1031). In the example, the editor has not only included the particle but also moved it to directly precede the question, interpreting the first-person performative after it (*ja yritän olla lyhytsanainen* "and I try to be brief") as a metapragmatic increment that is actually supposed to target the previous utterance and thus be located before the particle. After reporting the question, the speaker then begins the answer with a pause and a particle *no* "well." With the particle, the speaker indicates that he acknowledges the project behind the reported question and starts to process it in the utterance that follows. Simultaneously, the particle might also stress that the question presents a problem that requires a solution (Vepsäläinen, 2019). This interpretation is in line with the relatively lengthy answer that follows the particle. In both cases, the utterance-initial particle serves a distinct function in the speech, and the editor has included them in the utterance as rhetorically significant.

Similarly to some utterance-initial particles, some of the **metadiscursive expressions** (such as *sitten* "then," *ja se että* "and the fact that"), which would draw more attention or activate more literal interpretations in writing, are not included in the transcript. The same goes for **mannerisms** that some of the MPs use frequently (e.g., *todella* "really," *myöskin* "also" several times in a sentence). In these instances, only some, or none, of the cases are left in the transcript as stylistic markers, on the grounds that an expression, which is repeated extensively without rhetorical weight, is emphasized more in the transcript than in speech (Kirjo, 2021, p. 11).

6. Discussion

In Finland, the making of the official parliamentary transcript involves many types of editorial changes. These include certain phonological, morphological, and syntactic features, self-corrections and planning expressions, stuttering, and slips-of-the-tongue, as well as several prosodic and non-verbal features of interaction. The linguistic and editorial practices have been documented in the internal guidelines of the Records Office of the Finnish Parliament. Editing the parliamentary transcript is a form of genre-conscious language regulation where the editor operates among several interconnecting tensions. These include tensions between speech and writing as semiotic channels, authenticity and readability as competing ideals of the transcript, and naturally occurring spoken language variation and the norms of the written standard language.

Making an official transcript can be observed as a process of entextualization (Bauman and Briggs, 1990; Park and Bucholtz, 2009; see section 3 above), which decontextualizes individual turns-at talk from the original speech event and recontextualizes them into another semiotic mode with partially different communicative resources (written text) and into another genre with its own goals and expectations (e.g., an official parliamentary plenary session record). As Komter (2022) highlights regarding this issue, this might have a profound effect on the transcribed speech event, depending on, for example, what the purpose and status of the transcript are, how the transcript is used, what is selected to be included as relevant in the transcript, and how the participants are presented. Moreover, the goals of the original speeches are unavoidably intertwined with the goals of the parliamentary record where the transcripts are included. This new context inevitably affects how the transcribed speeches are received and interpreted (see also Holder et al., 2022). Finally, all the different editorial decisions that are made in the transcription process always have an impact on how speeches are presented in the transcript. Whether these decisions are successful or not depends on the context and purpose of the transcript. As Fraser (2022) showed, no transcript is valid for all purposes: transcription choices that work well in one context might be unacceptable in another.

In addition to the explicit principles and practices of Finnish parliamentary reporting, the most central of which have been analyzed in this article, there are undoubtedly other differences between the parliamentary session and the written transcript which are based on the individual decisions of the editors in different contexts. The detailed analysis of these phenomena is outside the scope of this study and is left for future research on Finnish parliamentary transcripts.

The editorial changes in official transcripts affect the mediation of Finnish parliamentary interaction in a number of ways. First, the standardization of linguistic variation affects the tone of transcribed speeches. Removing mostly phonological but also some morphological and syntactic variations can be seen as preserving readability and preventing over-emphasis on some spoken language features. Having said that, removing this variation might turn the register of the speeches toward a more formal direction. Second, editing some gradually emerging structures into more coherent ones, as well as removing elements, such as self-corrections, planning expressions, stuttering, and slips-of-the-tongue, affects how speakers' ways of processing their thoughts are conveyed to the readers. It might, for example, make the transcripts appear more controlled or deliberate than the original speech. Applying the linguistic metafunctions introduced by Halliday (2003), the ideational meanings that deal with describing reality are emphasized, but the interpersonal and textual meanings are often affected by editing (see also Slembrouck, 1992).

The principles of creating and editing parliamentary transcripts have changed considerably during the past few decades. This can be observed when comparing the current practices with how Kallioniemi (1946, p. 147) describes parliamentary transcription in the late 19th century. According to him, the stenographer should edit "lousy" speeches so that they became "exemplary in terms of content and language" and confusing statements became clear. In fact, according to the experienced officials in the Records Office, this type of orientation to transcription prevailed well into the 1980s, when the editors began to develop more authentic linguistic and editorial principles. The old ideals and practices were occasionally criticized for changing the transcribed speeches so much that it gave the impression that they were spoken by the same person (Kallioniemi, 1946). This means that the old transcripts might differ considerably from the original sessions (cf. Harvard, 2011, on the old transcripts of the Swedish Parliament). Some old editorial principles that are no longer followed are the correction of false statements (e.g., figures, names, and other information), the correction of false citations (e.g., unclear formulations and missing words), and changing inappropriate behavior (e.g., informal forms of address and improper words).

It is left for later research to give a more comprehensive picture of the development of transcription and editing principles in the Finnish Parliament. However, I will illustrate a recent significant change in the editorial principles here. Between the late nineteenth century and early 2021, most short routine turns by the chairperson were excluded from the transcript. Most of these were turns where the chairperson gave the floor to the next speaker (e.g., *Seuraavaksi edustaja Meri* "Next MP Meri"). The reason given for the exclusion of these administrative turns in the Records Office was that they were seen as unnecessary for the reader between MPs' speeches. The same argument was made for the exclusion of different metadiscursive and technical remarks which refer to the organization of the session, such as comments on the microphone (*anteeksi onko mikrofoni päällä* "excuse me is the microphone on"). Another reason for the exclusion of routine turns by the chairperson and different metadiscursive remarks was that they were not seen as substantial parts of the session but rather as its technical administration (Voutilainen, 2016). However, since the beginning of 2021, all these have been included in the transcript—except for some of the simple cancellations of taking the floor when they are not commented on in the session. The reason for including all the chairpersons' routine turns, as well as administrative and technical remarks, is to convey the nature of the parliamentary plenary session as institutional interaction. A further reason for their inclusion is to reduce the monologization of the session in the transcript, which was seen to happen during the earlier practice when the administrative turns had been edited out (Records Office, 2021; Voutilainen, 2021).[11]

The inclusion of chairpersons' turns has had a considerable impact on how the plenary session interaction is conveyed in the official transcript. The chairperson has a significant administrative role in every official speech by giving the floor to the MPs and managing the technical details of the session. When these administrative and technical turns were largely edited out, the focus of the official transcript was almost exclusively on the individual speeches, whereas the nature of the plenary session as institutional interaction was faded out. The inclusion of chairpersons' turns and other technical talk has increased, mediating the nature of plenary session conversation as a whole. In other words, it has significantly affected the chronotope of the transcript, i.e., how the sense of time and space in the institution is communicated to the reader (see Bakhtin, 1981; see also De Fina and Perrino, 2020).

The relationship between parliamentary sessions and parliamentary transcripts is especially important to consider when using parliamentary transcripts as data for scientific research. There might not be serious validity problems for analyzing the content of the speeches, but when studying the interactional details, discourse processing, or linguistic variation, for example, the researcher should consult the original audio and video recordings of the sessions. Even when analyzing the content of the speeches, it is important to note that the form and content of the speeches might be difficult to keep separate in practice (see Semino, 2011). For example, the regional phonological features that are removed in the transcript might carry important weight in practically profiling the MP as a politician with regional issues at heart. In addition to this, there is always the possibility that some parts of speech have been misheard or misinterpreted by the transcriber and the editor. For a parliamentary researcher, it is nonetheless important to familiarize oneself with the transcription practices and editorial principles of the parliament in question. Preliminary studies suggest that there are vast differences between these practices and principles in different parliaments (Voutilainen, 2019a,b). To shed more light on this, it is important to provide systematic comparisons between transcription cultures, working methods, and linguistic ideologies of different parliamentary reporting offices in the future.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.eduskunta.fi/FI/search/Sivut/Vaskiresults.aspx#Default=%7B%22k%22%3A%22%22%2C%22o%22%3A%5B%7B%22d%22%3A1%2C%22p%22%3A%22Laadintapvm%22%7D%5D%2C%22r%22%3A%5B%7B%22n%22%3A%22Asiakirjatyyppinimi%22%2C%22t%22%3A%5B%22%5C%22%C7%82%C7%8250c3b67974c3a46b69726a61%5C%22%22%5D%2C%22o%22%3A%22AND%22%2C%22k%22%3Afalse%2C%22m%22%3A%7B%22%5C%22%C7%82%C7%8250c3b67974c3a46b69726a61%5C%22%22%3A%22P%C3%B6yt%C3%A4kirja%22%7D%7D%2C%7B%22n%22%3A%22Toimija%22%2C%22t%22%3A%5B%22%5C%22%C7%82%C7%8254c3a47973697374756e746f%5C%22%22%5D%2C%22o%22%3A%22AND%22%2C%22k%22%3Afalse%2C%22m%22%3A%7B%22%5C%22%C7%82%C7%8254c3a47973697374756e746f%5C%22%22%3A%22T%C3%A4ysistunto%22%7D%7D%5D%7D.

Ethics statement

Written informed consent was not obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Acknowledgments

EV wishes to thank his colleagues in the Records Office of the Finnish Parliament for the rich and enlightening discussions about the making of the official transcript over the past 13 years. EV especially wants to thank Annamari Koskelo, Maarit Peltola, and Veli Tolvanen for commenting on a draft of this article. EV also thanks the editors and reviewers for their insightful and constructive comments and suggestions.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher,

[11] Regarding the same phenomenon in the Hansard of the House of Commons, UK, see Slembrouck (1992). Regarding monologisation in quoting, see Haapanen (2017).

the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bakhtin, M. (1981). *The Dialogic Imagination: Four Essays.* Austin, TX: University of Texas Press.

Bauman, R., and Briggs, C. (1990). Poetics and performance as critical perspectives on language and social life. *Annu. Rev. Anthropol.* 19, 59–88. doi: 10.1146/annurev.an.19.100190.000423

Bayley, P. (2004). "Introduction: the whys and wherefores of analysing parliamentary discourse," in *Cross-cultural Perspectives on Parliamentary Discourse*, ed. P. Bayley (Amsterdam: John Benjamins), 1–44. doi: 10.1075/dapsac.10.01bay

Biber, D. (1988). *Variation Across Speech and Writing.* Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511621024

Blevins, M. D. (2017). "Participant observation," in *Encyclopedia of Communication Research Methods*, ed. M. Allen (London: Sage), 1188–1190.

Cortelazzo, M. (1985). "Dal parlato al (tra)scritto: I resoconti stenografici dei discorsi parlamentari," in *Gesprochenes Italienisch in Geschichte und Gegenwart*, eds. G. Holtus and E. Radtke (Tübingen: Narr), 86–118.

Cucchi, C. (2013). Dialogic features in EU non-native parliamentary debates. *Rev. Air Force Acad.* 22, 5–14. Available online at: https://www.afahc.ro/ro/revista/Nr_3_2012/Articol_Cucchi.pdf

De Fina, A., and Perrino, S. (2020). Introduction: chronotopes and chronotopic relations. *Lang. Commun.* 70, 67–70. doi: 10.1016/j.langcom.2019.04.001

Devitt, A. (2004). *Writing Genres.* Carbondale: South Illinois University Press.

Drew, P. (2013). "Turn design," in *The Handbook of Conversation Analysis*, eds. J. Sidnell, and T. Stivers (Malden: Wiley-Blackwell), 131–149. doi: 10.1002/9781118325001.ch7

Drew, P., and Heritage, J. (1992). "Analyzing talk at work: An introduction," in *Talk at Work*, eds. P. Drew, and J. Heritage (Cambridge: Cambridge University Press), 3–65.

Eades, D. (1996). "Verbatim courtroom transcripts and discourse analysis," in *Recent Developments in Forensic Linguistics*, ed. H. Kniffka (Frankfurt: Lang), 241–254.

Eggins, S., and Martin, J. (1997). "Genres and registers of discourse," in *Discourse as Structure and Process*, ed. T. A. van Dijk (London: Sage), 230–256. doi: 10.4135/9781446221884.n9

Etymonline (2023). *Online Etymology Dictionary*. Compiled by Douglas Harper. Available online at: https://www.etymonline.com/ (accessed March 24, 2023).

Fairclough, N. (1992). *Discourse and Social Change.* Oxford: Polity.

Finnish Parliament (2023). *The Public Web Page of the Finnish Parliament in English.* Available online at: https://www.eduskunta.fi/EN/pages/default.aspx (accessed March 24, 2023).

Fraser, H. (2022). A framework for deciding how to create and evaluate transcripts for forensic and other purposes. *Front. Commun.* 7, 898410. doi: 10.3389/fcomm.2022.898410

Gardey, D. (2010). Scriptes de la démocratie. Les sténographes et rédacteurs des Débats (1848–2005). *Sociol. Travail* 2, 195–211. doi: 10.1016/j.soctra.2010.03.001

Gardey, D. (2013). "Enregistrer' et render les débats publics en Grande-Bretagne et en France. La sténographie comme exigence et révélateur de la démocratie parlamentaire?," in *Faire parler leParlement. Methodes et enjeux de l'analyse des débats parlamentaires pour les sciences sociales,* eds. C. De Galembert, O. Rozemberg, and C. Vigour (Paris: Libraire Générale de Droit et Jurisprudence), 73–89.

Gault, R. (1994). "Education by the use of ghosts: strategies of repetition in Effi Briest," in *Repetition in Discourse. Interdisciplinary Perspectives*, ed. B. Johnstone (Ablex: Norwood), 139–151.

Goodwin, C. (1994). Professional vision. *Am. Anthropol.* 96, 606–633. doi: 10.1525/aa.1994.96.3.02a00100

Gottlieb, H. (2018). "Semiotics and translation," in *The Routledge Handbook of Translation Studies and Linguistics*, ed. K. Malmkjaer (London: Routledge), 45–63. doi: 10.4324/9781315692845-4

Guillory, J. (2004). The memo and modernity. *Critic. Inq.* 31, 108–132. doi: 10.1086/427304

Haapanen, L. (2017). *Quoting Practices in Written Journalism.* Helsinki: University of Helsinki.

Hakulinen, A., Vilkuna, M., Korhonen, R., Koivisto, V., Heinonen, T. R., and Alho, I. (2004). *Iso suomen kielioppi. [The Comprehensive Grammar of Finnish].* Helsinki: Finnish Literature Society.

Halliday, M. A. K. (1989). *Spoken and Written Language.* Oxford: Oxford University Press.

Halliday, M. A. K. (2003). *An Introduction to Functional Grammar.* 3th ed. Revised by M.I.M Matthiessen. London: Hodder Arnold.

Harvard, J. (2011). "Riksdagsprotokollen som medium," in *Dolt i offentligheten: Nya perspektiv pa traditionellt kallmaterial*, eds. J. Harvard, S. Förhammar, and D. Lindström (Lund: Sekel), 25–46.

Hepburn, A., and Bolden, G. B. (2013). "The conversation analytic approach to transcription," in *The Handbook of Conversation Analysis*, eds. J. Sidnell, and T. Stivers (Malden: Wiley-Blackwell), 57–76. doi: 10.1002/9781118325001.ch4

Heritage, J. (1984). *Garfinkel and Ethnomethodology.* Cambridge: Polity Press.

Heritage, J. (2015). Well-prefaced turns in English conversation: a conversation analytic perspective. *J. Pragmat.*, 88, 88–104. doi: 10.1016/j.pragma.2015.08.008

Heritage, J., and Clayman, S. (2010). *Talk in Action. Interactions, Identities and Institutions.* Chichester: Wiley-Blackwell. doi: 10.1002/9781444318135

Holder, A., Elsey, C., Kolanoski, M., Brooker, P., and Mair, M. (2022). Doing the organization's work—transcription for all practical governmental purposes. *Front. Commun.* 6, 797485. doi: 10.3389/fcomm.2021.797485

Holt, E., and Clift, R. (2010). *Reporting Talk. Reported Speech in Interaction.* Cambridge: Cambridge University Press.

Hughes, R. (1996). *English in Speech and Writing. Investigating Language and Literature.* London: Routledge.

Ilie, C. (2015). "Parliamentary discourse," in *The international Encyclopedia of Language and Social Interaction*, eds. K. Tracy, C. Ilie, and T. Sandel (New Jersey: John Wiley and Sons). doi: 10.1002/9781118611463.wbielsi201

Ilie, C. (2016). "Parliamentary discourse and deliberative rhetoric," in *Parliament and Parliamentarism. A Comparative History of a European Concept,* eds. P. Ihalainen, C. Ilie, and K. Palonen (New York: Berghahn Books). doi: 10.2307/j.ctvgs0b7n.13

Ilie, C. (2018). "Parliamentary debates," in *The Routledge Handbook of Language and Politics*, eds. R. Wodak, and B. Forchtner (London: Routledge). doi: 10.4324/9781315183718-24

Jakobson, R. (1959). "On linguistic aspects of translation," in *On translation*, ed R. A. Beuwer (Cambridge: Harvard University Press), 232–239.

Jefferson, G. (2004). "Glossary of transcript symbols with an introduction," in *Conversation Analysis: Studies from the First Generation*, ed. G. Lerner (Amsterdam: John Benjamins), 13–31. doi: 10.1075/pbns.125.02jef

Johnstone, B. (ed.) (1994). *Repetition in Discourse. Interdisciplinary Perspectives.* Ablex: Norwood.

Kallioniemi, K. (1946). *Pikakirjoitus ja säätyvaltiopäivät. [Stenography and the Parliament for the Estates.]* Helsinki: Otava.

Kirjo (2021). *Linguistic and Editorial Guidelines of the Finnish Parliament (in Finnish).* Helsinki: Finnish Parliament, Records Office.

Komter, M. (2022). Institutional and academic transcripts of police interrogations. *Front. Commun.* 7, 797145. doi: 10.3389/fcomm.2022.797145

Language Bank of Finland (2019). *Plenary Sessions of the Parliament of Finland, Kielipankki Korp Version 1.5.* Available online at: http://urn.fi/urn:nbn:fi:lb-2019101621 (accessed March 24, 2023).

Linell, P. (1998). *Approaching Dialogue. Talk, Interaction and Contexts in Dialogical Perspectives.* Amsterdam: Benjamins. doi: 10.1075/impact.3

Linell, P. (2005). *The Written Language Bias in Linguistics. Its Nature, Origins, and Transformations.* London: Routledge. doi: 10.4324/9780203342763

Martin, J., and Rose, D. (2008). *Genre Relations. Mapping Culture.* Sheffield: Equinox.

Merritt, M. (1994). "Repetition in situated discourse—exploring its forms and functions," in *Repetition in Discourse. Interdisciplinary Perspectives*, ed. B. Johnstone (Ablex: Norwood), 23–36.

Mollin, S. (2007). The Hansard hazard: gauging the accuracy of British parliamentary transcripts. *Corpora* 2, 187–210. doi: 10.3366/cor.2007.2.2.187

Ong, W. (1982). *Orality and Literacy: The Technologizing of the Word.* London: Methuen. doi: 10.4324/9780203328064

Park, J., and Bucholtz, M. (2009). Introduction. Public transcripts: entextualization and linguistic representation in institutional contexts. *Text Talk* 29, 485–502. doi: 10.1515/TEXT.2009.026

Parliamentary Office (2019). *Eduskunnan kanslian strategia 2020–2023. [The Strategy of Finnish Parliamentary Office 2020–2023. in Finnish]*. Helsinki: Finnish Parliament.

Peltola, M. (2015). *Pöytäkirjatoimisto. [Records Office. A Finnish presentation for the newly elected MPs in the Finnish Parliament (in Finnish)]*. Helsinki: Finnish Parliament.

Potter, J., and Wetherell, M. (1987). *Discourse and Social Psychology. Beyond Attitudes and Behaviour*. London: Sage.

Rapley, T. (2001). The art(fulness) of open-ended interviewing: some considerations on analysing interviews. *Qual. Res.* 1, 303–323. doi: 10.1177/146879410100100303

Records Office (2021). *Memos of the Editorial Meetings of the Records Office of the Finnish Parliament* (in Finnish). Helsinki: Parliament pf Finlans, Record Office.

Rock, F. (2007). *Communicating Rights. The Language of Arrest and Detention*. Basingstoke: Palgrave-Macmillan. doi: 10.1057/9780230286504

Sacks, H., Schegloff, E., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language* 50, 696–735. doi: 10.1353/lan.1974.0010

Schegloff, E., Sacks, H., and Jefferson, G. (1977). The preference of self-correction in the organization of repair in conversation. *Language* 53, 361–382. doi: 10.1353/lan.1977.0041

Semino, E. (2011). "Stylistics," in *The Routledge Handbook of Applied Linguistics*, ed. J. Simpson (London: Routledge), 541–553.

Sidnell, J. (2010). *Conversation Analysis. An Introduction.* Chichester: Wiley-Blackwell. doi: 10.1093/obo/9780199772810-0062

Sidnell, J., and Stivers, T. (2013). *The Handbook of Conversation Analysis.* Chichester: Wiley-Blackwell. doi: 10.1002/9781118325001

Slembrouck, S. (1992). The parliamentary Hansard 'verbatim' report: the written construction of spoken discourse. *Lang. Liter.* 1, 101–119. doi: 10.1177/096394709200100202

Sorjonen, M.-L. (2001). *Responding in Conversation: A Study of Response Particles in Finnish.* Amsterdam: John Benjamins. doi: 10.1075/pbns.70

Steiner, G. (1975). *After Babel. Aspects of Language and Translation.* Oxford: Oxford University Press.

Tiililä, U. (2012). Tekstilajit kielenhuollossa [Genres in language planning]. *Kielikello* 4, 8–10. Available online at: https://www.kielikello.fi/-/tekstilajit-kielenhuollossa

Tiittula, L., and Nuolijärvi, P. (2013). *Puheen illuusio suomenkielisessä kaunokirjallisuudessa* [The illusion of speech in Finnish literary fiction]. Helsinki: Finnish Literature Society.

Tiittula, L., and Voutilainen, E. (2016). "Puhe, kirjoitus ja puheen muuttaminen kirjoitukseksi [Speech, writing, and turning speech into writing]," in *Puheesta tekstiksi. Puheen kirjallisen esittämisen alueita, keinoja ja rajoja [From spoken word to written text. Domains, means and restrictions of the representation of spoken language in written form],* eds. L. Tiittula, and P. Nuolijärvi (Helsinki: Finnish Literature Society), 29–62.

Treimane, L. (2011). Analyzing parliamentary discourse: systemic functional perpective. *Kalbotyra* 63, 78–93. doi: 10.15388/Klbt.2011.7653

Vepsäläinen, H. (2019). *Suomen "no"-partikkeli ja kysymyksiin vastaaminen keskustelussa (Finnish "no" particle and answering questions in conversation).* University of Helsinki, Helsinki, Finland. Available online at: http://urn.fi/URN:ISBN:978-951-51-5189-6

Voutilainen, E. (2016). "Tekstilajitietoista kielenhuoltoa: puheen esittäminen kirjoitettuna eduskunnan täysistuntopöytäkirjoissa [Genre-conscious language planning: written representation of speech in plenary session records]," in *Puheesta tekstiksi. Puheen kirjallisen esittämisen alueita, keinoja ja rajoja [From spoken word to written text. Domains, means and restrictions of the representation of spoken language in written form]*, eds. L. Tiittula, and P. Nuolijärvi (Helsinki: Finnish Literature Society), 162–191.

Voutilainen, E. (2017). The regulation of linguistic quality in the Finnish parliamentary verbatim reporting. *CoMe* 2017, 61–73. Available online at: http://comejournal.com/wp-content/uploads/2019/06/5.-CoMe-II-1-2017.-Voutilainen.pdf

Voutilainen, E. (2019a). "Everyday linguistic and editorial choices in parliamentary reporting," in *Intersteno Parliamentary and Other Professional Reporters' Section Meeting* (Cagliari, Italy).

Voutilainen, E. (2019b). "Linguistic ideologies and editorial principles in parliamentary reporting," in *52nd Intersteno Congress* (Cagliari, Italy).

Voutilainen, E. (2021). *Reporting the Chairperson's Speech in the Official Parliamentary Report: Useless Noise or Essential Information?* Tiro 2/21.

Voutilainen, E., Peltola, M., Räty, T., and Varisto, N. (2013). "Rules of reporting: the principles of representing spoken discourse in the Records Office of the Finnish Parliament," in *49th Intersteno Congress: Intersteno Parliamentary and other Professional Reporters Section (IPRS)* (Gent, Belgium).

Yates, J. (1989). *Control Through Communication. The Rise of System in American Firms.* Baltimore: Johns Hopkins University Press. doi: 10.56021/9780801837579

frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 24 August 2023
DOI 10.3389/fcomm.2023.1062733

OPEN ACCESS

EDITED BY
Emma Richardson,
Aston University, United Kingdom

REVIEWED BY
Mary Macken-Horarik,
Australian Catholic University, Australia
Magnus Hamann,
Loughborough University, United Kingdom

*CORRESPONDENCE
Anna-Lena Carlsson
✉ anna-lena.carlsson@mdu.se

RECEIVED 06 October 2022
ACCEPTED 14 June 2023
PUBLISHED 24 August 2023

CITATION
Carlsson A-L and Svensson Harari N (2023) The complexity of situated text design: a negotiation between standardization and spoken language in a manufacturing company. *Front. Commun.* 8:1062733.
doi: 10.3389/fcomm.2023.1062733

The complexity of situated text design: a negotiation between standardization and spoken language in a manufacturing company

Anna-Lena Carlsson* and Natalia Svensson Harari

School of Information Design, School of Innovation, Design and Engineering, Mälardalen University, Eskilstuna, Sweden

In information design textbooks, text design is mostly understood as typography and layout. The meaning-making process of language, involving social interaction that affects language, is rarely acknowledged. Instead, texts are supposed to be "clearly" written. In this research article, we argue that the understanding of text design could benefit from also addressing text production and use situated amid social activity. This article presents a study on a text design process partly based on spoken language and owned by assembly operators in a workplace. Capturing the spoken dialogue and transforming it into instructive texts resembling transcripts are essential steps in securing the best practices for the smallest tasks in manual assembly, the minima of working, which is crucial for manufacturing. Our aim within the information design field is 2-fold: To underline the meaning-making process in language as a social phenomenon and to show that the situated design perspective, i.e., an outlook that highlights the uniqueness of the setting, can be important for the production and use of certain texts, such as instructions, and for affecting language. We asked ourselves: What are the consequences for the information design field when meaning-making in a language is understood as being socially situated in an activity? We have studied a design process and used observations, interviews, and text analysis to gather data. The result showed that the workers' ownership of text documents is crucial for the texts' use, yet the texts used do not meet the standard of information design textbooks. Moreover, the design of the text involves a continuous and non-linear collective negotiation that balances standardization in language and work procedures with the incorporation of operators' linguistic improvements. We unfold a case of text design where there is a closeness of designer and user roles, a non-linearity of the process, and an understanding of an information design product as becoming rather than having been finalized for use.

KEYWORDS
text design, spoken language, transcription, ownership, manual assembly instructions

1. Introduction

Information design is a field of practice that produces information to be as effective as possible for understanding communication or promoting an action, e.g., in instructions. Graphic design and visuals have historically been in focus. In information design textbooks, text design has mostly been understood as typography and layout, although semiotics can be briefly mentioned (e.g., Jacobson, 2000; Pettersson, 2002; Coates and Ellison, 2014). When commenting on language, the focus has frequently been on formal and vague tropes such as writing "clearly," using a "consistent writing style," and producing "concrete" texts (see, for example, Pettersson, 2002) without "embellishments" (Frascara, 2015). This is all very well when it comes to communicating with larger groups, and there is a necessary distance between the designer and the user of information; the user is still the "other" to be considered by the designer (Frascara, 2015, p. 5). However, there are areas where information is both being produced and used closer to the actions informed about, but this comes with consequences for the idea of, e.g., text design: the closeness of designer and user roles, the non-linearity of the process, and the understanding of an information design product as becoming rather than having been finalized for use. We highlight this through an illustrative case where social interaction in producing and using informative texts and other activities is crucial for meaning-making in communication.

Our aim in the field of information design is to underline the meaning-making process in language as social and sometimes as situated in a particular setting. Therefore, texts have traces of the uniqueness of the context and social interactions in their form. Texts even have features like those in transcriptions of talk, i.e., texts have captured the talk in the workplace, where they are both produced and used. The case highlights the consequences for text design in an organization that aims to standardize communication while also recognizing the need to capture the minima of tasks discussed and performed by operators in their activity. We asked ourselves: What are the consequences for the information design field when meaning-making in a language is understood as being socially situated in an activity? The significance for the information design field is to broaden the understanding of text design, mostly understood as typography, and recognize the lived interplay with form and content in a situated and social setting for meaningful communication.

By neglecting to incorporate insights from other disciplines, such as linguistics acknowledging the social nature of language, the discipline of information design runs the risk of developing inadequate theories that do not effectively address the complexities of real-world information design practices. The practical contribution lies in fostering a more favorable understanding of collaborative writing and using texts based on spoken interaction in information design.

It should be noted that social semioticians, sometimes referred to in information design, work with how multimodal communication functions during activity in education (e.g., Bezemer and Kress, 2016). Kress and van Leeuwen (2006) also started from a social and linguistic base, e.g., by promoting visuals in socio-semiotic communication. We are not claiming that our study is the first to assume the social perspective in communication and design; rather, we argue that it is still relevant to underline that the meaning-making in a *language* is social, as demonstrated by the case in our study, to enhance the understanding of informative *text* design. Moreover, it is worth noting that the consequences of factors such as the division between the designer and user roles, the categorization of texts as "instruction" or "personal letters", and the outlook on the processes involved in text design have ongoing significance and merit further discussion. The field of information design often falls short of accommodating a spectrum that encompasses both the design of generic information for broader audiences and the design of situated information that acknowledges the uniqueness of the context. Our position aligns with the latter outlook but with a focus only on language in text design. In addition, it should also be mentioned that technical communication is indeed a field where scholars have challenged the ideal of writing "clearly". We shall return to this subject later in the article. In this introduction, we now turn to the perspective of text design and situated design used, followed by a summary of the positions taken on language from linguistics, supporting our perspective.

The text design used in this article is defined in line with Schön's (1983) concept of the reflective design practitioner. Schön (1983) emphasizes the concept of continuous and reflective conversation with the design material, which is closely linked to practice and is relevant for situational design (e.g., Simonsen and Hertzum, 2012). The situated design perspective, as described by Simonsen and Hertzum (2012) and Simonsen et al. (2014), underscores the uniqueness of a situation into the design and has its roots in the understanding of knowledge as situated. A situated perspective can underline both the fact that different people come together in the design and also the collaborative, ongoing work improvements of a design (see Olsen and Heaton, 2012) as relevant for the organization's aim of capturing the minima in work tasks in our case. The term "minima" used in this article refers to the smallest entities in the lived practice of a situation, in contrasts to standards, for instance, in concepts. The smallest work entities make their mark in language. In our case, the philosophy of the company embraces a bottom-up perspective in the name of efficiency and quality work.

As we shall observe, the capture of operators' best practices regarding the smallest work task is accomplished through a specific type of text utilized on the production line, affecting the language employed within the text. Moreover, the user can be problematized from a situated design perspective: "The concept of the user relates to the appearance of many different actors on the stage of design" (McHardy et al., 2012, p. 99). Here, the users can be "makeshift users" involved in the design project (p. 96), i.e., participants in design. Concerning our design research perspective, we have taken the research-into-design approach, i.e., we studied a design process, rather than the research-through-design perspective, i.e., gaining knowledge through participation in the design process, which is more usual in situated design research (Baerenholdt et al., 2012).

Regarding the process of writing texts amid an ongoing activity, we will refer to Schön's (1983) insights concerning reflection on and in action during the design process, as well as his thoughts on capturing the knowledge acquired through reflection in the form of *description*. Schön (1983) links intuitive reflection in action with the difficulty of formulating what one knows; tacit knowledge does not easily transform into language. However, knowledge in action must be transformed into language for communication with others if a consensus is desired. The descriptions linked to action can be viewed as part of the frame of a practice affecting practitioners' reflection in action. Schön (1983) writes that the media cannot be separated from the language here; they make up the "'stuff' of inquiry in terms of how the practitioners move, experiment, and

Abbreviations: O, operator; TL, team leader; ATL, assistant team leader; PT, production technician; XPS, company x's production system; SOS, standard operation sheets.

explore" (p. 271). There is a "feel" for the activity of touch as well as a "feel" for the language (p. 271). The media and language, for example, the language on paper in the binder on a workshop floor, as we shall notice in our case, are also subject to change. Schön (1983) wrote that the reflection accomplished in action is not dependent on a description of the intuitive knowledge but that some descriptions can be appropriate: "Descriptions that are not very good may be good enough to enable an inquirer to criticize and restructure his intuitive understandings so as to produce new actions that improve the situation or trigger a reframing of the problem" (p. 277). As we shall observe, reflection on action is necessary, together with other roles representing the frame for creating the instructive texts with the character of transcripts in our case. The issue with transcripts, however, is mostly that, when they become decontextualized, understanding suffers, e.g., in transcripts from covert recordings later used in court (Gilbert and Heydon, 2021). We will come back to this in the findings and Discussion section.

Collaborative design, often associated with the situated design perspective, is frequently referred to as co-design. It involves many stakeholders engaging in various modes of design. For instance, Roth et al. (2017) examined dialogues encompassing both words and gestural interactions among collaborators and their engagement with materials throughout the co-designing processes. Because of the situatedness, this can mean that designers and users interact during production. Lee (2008, p. 33, with reference to Lefebvre, 1972) is of interest for the present article because of the proposal of a "realm of collaboration" instead of, on the one hand, an abstract space where designers and experts work, and, on the other hand, a concrete space where people, i.e., users, live. Lefebvre (2003, p. 182) wrote about a concrete space of "habiting, gestures and paths, bodies and memory, symbols and meanings … contradictions and conflicts between desires and needs", in contrast with an abstract space where designers "look down on their 'objects' … from above and afar". The image of a collaborative space can be contrasted with the linear process produced by the notion of a distance between the designer and the user. From the field of co-creation in designing, we can understand that collaborative writing in the situated text design has a close link with Schön's (1983) frame of media and language in that it is intertextual and verbal; it is in its meaning-making and consists of negotiations, talks about work, in a concrete space, or as Lee (2008) calls it, in a realm of collaboration.

Having addressed our design focus, we turned to how language is related to the social setting in which it obtains its meaning. Thinking in terms of linguistics is important here. The arbitrary character of linguistic signs, highlighted in linguistics by De Saussure's (2015) semiotics, underlines that there is no necessary correlation between a sign and an object and underscores the difference between the system and the usage of linguistic signs. The consequences of the arbitrary character of linguistic signs piqued philosophers' interest throughout the 20th century. Rorty (1992) described this as a linguistic turn. The social perspective on language has mostly been taken over by linguistics (Nystrand, 1989). This is, for instance, observed in writing research (Hyland, 2016) and in the field of workplace writing (Bremner, 2018).

Concepts of importance in workplace writing are the understanding of "intertextuality" in the sense that writing always relates to other texts and "collaboration" in the writing process. Particular writing is "taking shape within chains of emails or other interactions, incorporating the work of colleagues as part of the collaborative process, or being informed by templates, practices, and traditions that are specific to an organizational setting" (Bremner, 2018, p. 7). The frame of media and language in an organization (Schön, 1983) can then be observed as the intertexts to which certain writing, e.g., an instructive text, is related. Bremner (2018) wrote about the templates produced for a particular need as generic intertextuality. Moreover, intertextuality is linked to the collaboration of colleagues in a workplace: "[I]nput and influences will come from the work of colleagues—workplace writing is essentially intertextual in that writers are collaborating, building on and revising each other's work in the process of knowledge making" (p. 43, with reference to Reither, 1993; Prior, 2004). The authority of authorship concerning workplace texts in companies shifted when companies started to notice the value of corporate identities. Authorship then slipped away from individuals to organizations. In technical communication in organizations, this has long been the case (Debs, 1991). Collaborative writing, Bremner (2018) writes, "is an almost integral element of any organization" (p. 55).

Instructive texts are commonplace in the case we have studied in the manufacturing industry. Delin (2000) wrote about instructions as an everyday type of text, where texts exist in relation to products and actions in contexts. The relationship with activities carried out in the context of the instruction emphasizes that the exactness of language depends on the writer and the receiver sharing or knowing the context in which the text is to be understood. This close relationship between using a product and the instructions, in association with the relevance of time issues, tends to make the language "telegraphic" (p. 68). Delin (2000) also wrote about how authority affects the choice of the directive. If no authority exists in the relationship between the speaker and the hearer of an instruction, it affects the form. This is also the case with texts in the field of technical communication, as mentioned earlier, bordering on information design, linguistics, and engineering. Kirkman (2005) addressed this in *Good Style: Writing for Science and Technology* as early as the 1990s (see also Pettersson, 2002). More recently, in this domain, Schneider (2002) has problematized "clarity" in language and claimed that the closeness of the technical communicator and the user through interaction in the same context is a key element. A workplace context is also not static but "constituted, moment to moment" (p. 212). Schneider (2002, with reference to Hayman, 1994) suggests "strategic talks" to create clarity in communication. Plain language is not always the answer to clear communication; jargon can communicate more effectively among people in a specific context. Blakeslee and Savage (2013) also wrote that, as a designer of technical communication, one should ask oneself what it means to write well in the industry, the field, and the company. The context is thus decisive.

In applied linguistics concerned with professional practices, texts used in an activity are sometimes called "inscribed objects" (Prior, 2020). In an ethnomethodological study, Due (2020), e.g., wrote about how information sheets in optician settings, sometimes called "charts," "leaflets," or "guides," are understood as inscribed objects consisting of many different signs and used cooperatively in social interaction, *in situ* in a work process as a resource for decision-making. The sheets are used in relation to pointing gestures and stares in communication about buying and selling glasses or contact lenses in an optician shop. Due's (2020) case shows that the informative sheets are used to establish shared "attention and common ground through verbal, spatial, and embodied orderly actions. Pointing practices, embodied orientation in space, gaze, and the use of the sheet are deeply embedded in ways that exploit the specificities of the situated action" (p. 140). The instructive sheets then have a central position for the activity taking place. Sticky notes can also be considered such inscribed objects.

Sticky notes are understood as both material objects, easily attached to various surfaces, and as inscribed objects that bear information (Landgrebe and Rye Marstrand, 2020, with reference to Caglio et al., 2014; Weilenmann and Lymer, 2014). An interesting parallel can be drawn between our study and the investigation conducted by Landgrebe and Rye Marstrand (2020), as both studies examined organizations that have adopted "lean management" principles. This philosophy involves engaging the workforce in continuous improvement, which is also the context of our case. This is interesting because informative sticky notes also play a role in our setting of continuous improvements.

We are drawing on the theories and previous research mentioned above to support an understanding of what text design could be. Informative text design can be found in settings with high social interaction with and about work activities. In this sense, text design takes place in smaller groups, where there is a close relationship between both doings and descriptions of doings. The team effort is rather a realm of collaboration than a linear process of writing and using. Instead of concrete genres, like "instructions," as we shall notice in our case, there are inscribed objects of information and ongoing writing of texts. The case article will give relevance, through the thinking above, to a type of text production and use that is rarely studied in the sense of informative text design. In the next section, we will turn to the choice of setting, the materials, and the methods used.

2. Setting, materials, and methods

2.1. Setting

The production line studied in this case is in a factory that belongs to a multinational company. The company has a lean philosophy inspired by Japanese manufacturing thinking (Liker, 2004). In the corporation's way of working, each individual and team is important for the quality of the production. Quality is achieved through internal efficiency, which means that there is a standardization of routines and tools throughout the company. It is an ongoing work that also acknowledges the bottom-up perspective, involving operators on the workshop floor to eliminate all sorts of waste, i.e., matters that do not contribute to efficiency and quality. Lean production and standardization work have been implemented in the factory since 2011, and the operators work in groups, at workstations, and on the production line[1,2]. In standardization work, the operators also participate in writing the existing standard of an operation in manual assembly, that is, the method used so that another operator can read it[3], e.g., new personnel or workers from another part of the line. According to Liker (2004) and Liker and Meier (2006), standardization work is not the fixing of a final method but a starting point from which one continuously improves. The reason for choosing the company for our study is that they have undergone a transformation into lean production, meaning that they do acknowledge a bottom-up perspective concerning improvements. At the same time, there is continuous calibration concerning processes taking place. This would, we believe, affect the language used in instructive texts for manual assembly.

2.2. Materials and methods

We conducted a small case study (Merriam and Tisdell, 2016), collecting data through interviews and on-site observations on the workshop floor, with one interview in a small meeting room with the production technician (PT)[4], and conducting text analysis. The nature of the data is ethnographical and in line with the study of a research-into-design process. Observations took place at one assembly line in the manufacturing company in Sweden, in interaction with operators (O1, O2),[5,6] the team leader (TL)[7], and the assistant team leader (ATL), explaining the site, the situations in which the texts were used, and their functions (see text footnote 2). Using a semi-structured guide, we conducted in-depth interviews on-site. Both authors participated in collecting the empirical data. The interviews were recorded, transcribed, and used for the text analysis. The text analysis, done by the lead author, was conducted to look for how language appeared in instructive texts and to recognize the social interaction in the setting where the text was both produced and used. The form of the language was linked to the setting, the situations and functions, and the participants in the communication. Only partly is this analysis in line with the social understanding of language in Halliday (1978, p. 11); the analysis is *not* a proper systemic, functional linguistic analysis. The functions we acknowledged are indeed the institutional setting (field) of the text and the relationship between the contributors of meaning (tenor), as well as the media through which communication takes place (mode) (Halliday, 1978, with reference to Doughty et al., 1971, 185–6). However, the analysis does not aim at establishing what selections of meaning the grammar implies to readers and writers but only *that* language in the texts studied takes its form, similar to transcripts of talk, yet functions in its social setting were compared with the information design ideal and rules mentioned earlier.

An initial examination, locating different types of documents, was first performed, and three types of texts were found linked to the manual assembly on the production line. Eleven documents, called Element sheets (confidential and cannot be disclosed), a form

[1] ATL, assistant team leader, was interviewed on 25 November 2013.

[2] Observation (2013). Observation at location done on 25 November 2013.

[3] XPS, The Company's Production System (2013).

[4] PT, production technician, was interviewed on 25 November 2013.

[5] O1, operator 1, was interviewed on 25 November 2013.

[6] O2, operator 2, was interviewed on 25 November 2013.

[7] TL, team leader, was interviewed on 25 November 2013.

of the information sheet, were thereafter singled out to be analyzed further concerning the language used. The initial examination of the documents showed that, in practice, the Element sheets functioned as instructions for how to perform the manual assembly. The sheets had a central position at the intersection between standardized work procedures, humans' social interactions, and capturing oral language in documents sharing features with transcripts. In this analysis, no visuals were considered because of the focus on language in text design.

3. Results

The findings will be presented below according to how, when, and why the documents were used. Thereafter, the process of document creation is elucidated and explored. Finally, the language of the Element sheets is discussed in relation to their ability to facilitate effective use within the setting.

3.1. Empirical findings

3.1.1. The documents' usage: how, when, and why?

Three types of documents linked to the manual assembly were found in a binder located at the balancing board of each assembly station on the production line (see text footnote 2): (1) The *Work instructions* were the oldest type of document, also used before standardization, consisting of an abstract drawing of components with arrows and different product parts' article numbers, showing the design of the product (see text footnotes 3, 4). We initially assumed that we should study these documents when looking for instructive texts; (2) *Standard operation sheets* (SOSs) [Swe. *Standard Operations Blad* (SOB)] and the Element sheets were introduced in the standardization in 2011. The SOS included the order of operations, times, a layout of the operator's movements, and variants of products in assembly. It had short sentences on "what" to do, pictograms on safety, critical moments, quality, and ergonomics, and it was hung at the workstation (see text footnotes 2, 3); and (3) *Element sheets* contained the best-known agreed-upon practice, the standard of "what," "how," and "why" in manual assembly. These sheets followed the same template form in the whole factory, as did the SOS, and showed one activity linked to a certain time (the tempo of the operation), the shortest standard time in the production cycle, and the minima. They contained the information required to perform work safely with the right quality at the right time (see text footnote 3). Photographs were used "to facilitate understanding" (see text footnote 3), and they sometimes had arrows marking movements and circles showing focus points. They also included pictograms, as in the SOS (see text footnote 3). If there were options, there was a sheet for each variant. Moreover, product parts' article numbers were not allowed in the *Element sheets* (see text footnote 1). It can be noted that all product parts used in the assembly were to be found at the assembly stations, ready for use and rewriting.

In the first examination, we turned to the use of the documents, which also led us to single out the *Element sheets* for further analysis of the language. It is important to note that the *ideal* on the workshop floor would be to *not* use documents during assembly. This is significant because of the effect spoken interaction has on instructive texts. Together with the company's bottom-up perspective regarding efficiency and quality work, the operators' discussion regarding their work is crucial for determining the best practice in a continuous negotiation among the operators. The operators thus first ask other, more experienced co-workers if a question arises. The oral mode was the priority, and there was intricate knowledge about whom to ask about what on the workshop floor (see text footnotes 1,2, 4). When the text mode in the documents *was* used, this occurred in four situations: (1) if a need for information arose during assembly, (2) in daily work observations, (3) in education, and (4) in seeking/solving a problem in production.

If (1) *a need for information arose* during assembly, the operators first turned to a colleague, as mentioned above, and then to documents (see text footnotes 1, 6). However, the assistant team leader noted the importance of checking the standards in the Element sheets (see text footnote 1). The main function of the Element sheet here was to provide correct instructions on how to perform a work operation. The SOS was the text quickly looked at for time, order of events, and considerations needing special attention, such as safety and quality (see text footnotes 2, 5). The Work instruction was not used at all if a need for information arose. In daily (2) *work observations*, the Element sheets were used to see if the standard was preserved or needed improvement (see text footnote 3). The main function of the sheet in work observations was to be a description of the best-known standard at the time. The SOS was updated if affected by an update of the sheet. The Work instruction document was not used in these daily work observations. In (3) *education*, the three documents were used when an operator was new or if there was a new model on the production line. The operator read the binder with all documents and worked under supervision for up to 2 weeks (see text footnote 1). The main function of the Work instruction in education was to teach about the design of the product to be assembled, which was relevant for memorizing actions linked with product parts used in assembly. The main function of the SOS during training was to teach the sequence, time of operations, and movements of the operator (see text footnote 3). Another function of the SOS was to teach when safety, quality, critical moments, and ergonomics were highlighted. The main function of the Element sheet here was to explain the standard and the best practice for the smallest operation in assembly. Especially relevant for learning was the reason given for a method, the "why" in the Element sheet (see text footnote 6). The texts were also used to (4) *seek/solve a problem* in production at large. If the standard was followed, there should not be a problem with a particular operation (see text footnote 1).

To summarize, the Element sheets were used if questions were raised during assembly, in work observations, in training, or if problem-seeking/solving was needed in the factory. Only the Element sheet had three functions: *documentation, instruction,* and *educational* material in training. It was the knowledge of the operators that enabled the establishment of a standard and the maintenance of these functions. According to all informants, all functions of the Element sheets sustained three functions (see text footnotes 1, 4–7). In the following section, we shall observe that the education function of the Element sheet had the potential to suffer the most because of its transcription-of-talk character, which is because, in learning, there is an assumed decontextualization between the texts and the new operator, hence the need for learning. The most problematic document and most distant from the actual assembly was, surprisingly, the document called *Work Instructions*. During training, the operators had to memorize the product parts' article numbers—the main information—but the Work instruction was thereafter not used and was not continuously updated (see text footnotes 2, 4).

3.1.2. The text design processes

Concerning the initial production of the documents, a "preparer" and production technicians created the Work instructions before a new model was introduced in production (see text footnotes 1, 4). Production technicians, assistant team leaders, and team leaders wrote the SOS (see text footnote 4). Operators could participate in this initial phase through the technician when they had suggestions. This could be discussed in their teams or at daily meetings. Because of the bottom-up perspective of the company, there were many opportunities to participate in the construction of the SOS. In an earlier chapter of standardization, the operators updated the SOSs themselves without involvement from the production technician, assistant team leaders, or team leaders, which created problems with the accuracy of time-related information, prompting the discontinuation of this practice (see text footnotes 4, 6). An important point to note in this study was that the Element sheets, from the beginning of the standardization work, were written by the production technicians. However, the operators (see text footnote 4) did not use the top-down text design. The production technician and the assistant team leader then insisted on the importance of involving operators in the writing process; it was regarded as quality work (see text footnotes 1, 4). Element sheets, functioning as the work instructions, were then really "owned by production", that is, assistant team leaders and team leaders created them when a new model was introduced or the pace was changed *together* with the operators. Initially, this was a heavy job, but they were used (see text footnote 1).

In the process of capturing the best practice, the ongoing and daily rewriting of the Element sheets took place during daily work observations and when solving problems in production. The team leaders or assistant team leaders took the binder and followed a chosen standard daily while observing an operation. Small discrepancies between text and activity were frequently found (see text footnotes 1, 5). The issue was then discussed and negotiated; it was then determined whether a rewrite had been missed or if the operator needed to be informed of the revision to the standard (see text footnotes 1, 5). The negotiation took place among operators in different work shifts and with team leaders and assistant team leaders, sometimes also with the production technician, regarding both the manner of assembly and the formulation of the instructions. Operators, with their language, then participated in the rewriting of the sheets (see text footnote 1). The writing resulted in multiple authorship. In our case, 11 sheets had seven authors (or combinations of authors). This collaboration also, as we shall notice, affected the use of language.

All informants also underlined that updating required much ongoing work (see text footnotes 1, 4–6). The binder also contained several sticky notes for suggested updates, discussed or not yet discussed, still not written into the standard. Language was frequently discussed during the workday, which all the notes on the binder about changes in progress in the texts revealed (see text footnote 2). As will be discussed, there were also suggestions or questions concerning writing and working in the sheets. The language was intimately entangled with the operators' unfolding knowing and doing at the manual assembly workstation of the production line. The sheets had a central position, as in Due's (2020) study of information sheets in opticians' shops. In the following sections, we shall observe how the way of writing was related to talking during the activity, which was not the case with Due (2020) inscribed objects.

3.1.3. The Element sheets: language and the affordance of use in the setting

Regarding the way of writing in the Element sheets, in which operators participated in the rewriting, we examined the relationship between the manufacturing setting (field), the written language (mode), the functions, and the group to whom the text was addressed (tenor). The setting was manual assembly in a manufacturing industry; the functions were those of documentation, instruction, and educational material, and the readers/writers were the operators, team leaders, assistant team leaders, and new co-workers.

Related to the manufacturing setting, or frame, to use Schön's (1983) concept, the Element sheets had a lot of technical terms, such as "reversing alarm" and "check valve", concepts that were crucial for documentation and instructions and were familiar to the team or concepts to be learned in training. There was also a standardization of terms (see text footnote 4). We found recurring verbs related to the engineering field, such as "affix" and "install", examples of the company's regular terms. The function of documentation of the standard and the instruction on how to work might suffer in language, as will be discussed below, but training could be a function suffering in the distance between the time of writing and reading if certain terms in the texts were not made consistent. The template form and the layout were also the same throughout the factory, which made it easier to understand for a trainee or someone who had worked in another team during the learning phase. The main parts of the Element sheets were related to the uniform layout and consisted of columns for "what," "how," "why" illustrations, and time. There were also possibilities for writing down the history of safety and quality problems.

It should be noted that most texts in the sheets were about "how" something was to be done, supporting all the functions in the usage of the sheets. In the setting, since neither the Work instruction nor the SOS had texts on how to assemble, the Element sheets in practice were the instructions. Interestingly, our discussion below shows that the text named "Work instruction" did not function as such, and the text serving as an instruction was not named "instruction". One element of the sheets that pointed toward a cultural setting of talking or procedures that were not yet captured in language was that the fields and columns were left blank in the sheets. The history of safety, for instance, was often left blank. Sometimes even the "how" to assemble was absent, leaving the "what" to do as the only instruction, documentation, and education. There was also a shortage of text on the "why", affecting the educational function since an incentive for the operation expands on the reasons for the work to be done. When the "why" element was present, this motive could be found in upcoming workstations or in the product's final use. Sometimes, the "how" to assemble was absent. In the column of what to do, it said, "Assemble the X clutch on Z," but "how" was not described (yet the whole operation, with five activities and only two with an explanation of "how", had three authors). It could also be assumed that the activities were self-explanatory and that there was no other way of working. This is also a sign of the contextualization of the texts, common in transcripts, capturing only the necessary information required for assembly at the specific site, reflecting the situational nature of the task.

The technical, standardized, and formal words were mixed with features in language coming from a culture of talking about work, giving them the character of transcriptions amid the activity. This was shown in the use of the Swedish word *skav* [Eng. "scrape"] instead of the proper *skavande*. We also found mash-ups of words; for example, "tighten" was written in Swedish as *dra fast,* but, here, it was inscribed as *drafast.* Moreover, related to both the setting and everyday local talk, there was an expression that was used both locally and technically. "Enter the nut" [Swe. *Äntra*] in Swedish is a specific word for "boarding", as in boarding a ship, but it is not used in "boarding" an airplane or a train. Here, the imperative was used for placing a nut on a screw before fastening it. It was the operators' term, and functioned precisely in the practical situation, sustaining documentation and instruction.

The texts had a vagueness in their appearance, yet they functioned well-enough as descriptions for others' intuitive understandings (Schön, 1983), although education, as we have observed before, might be a problematic function. The texts featured vagueness when referring to practical knowledge on-site in the situation, e.g., "make sure the X is in the right position". Furthermore, both the "what" and the "how" were imprecise in the text: What to do was to "[p]lace X on the assigned place", and the way of doing it was "[a]ssigned place in the pallets". It can be assumed that an operator would know the meaning of a "right" position and an "assigned" place. Imperative sentences were frequent: "Assemble the lower X" and "Dismantle the plug in the X." However, sometimes, a definite form and, other times, an indefinite form were used about the same operation: A/the "upper X tube," "partition wall," or "X console" was used. At the site, in assembly, this sporadic openness to any kind of, e.g., tube, would not have been a problem in this case since there was only one upper tube, partition wall, and X console to use in the task. In other cases, this was not easy because a diversity of objects and metaphors were used to handle the variety in the operations. "The stomach" [Swe. *mage*] of a tube was used for its convex bending in the analyzed sheets. This was a way of aiding documentation, instruction, and education among the group. The assistant team leader stressed that the operators, before standardization, used more metaphors than that. He argued that the groups' own metaphorical language could be good information: "If you have 40 differently marked tubes, you might need a way of remembering. This is not allowed in the transcripts anymore... A 'yellow-pink' tube going up was called 'China,' and a 'blue-pink' tube going down was 'USA.'" This was, however, still used orally for memorization during training (see text footnote 1). We shall return to this topic later.

Another characteristic of the language was the different forms of spelling, presumably due to writing in haste and being influenced by talk and/or the multitude of authors. The Swedish word for "thorough" or "careful" was spelled both as *noggrann* (correct) and *nogran*. We also found a frequent loss of punctuation in single sentences and at the end of paragraphs, as well as grammatical errors. In the mix, there were also some signs of a formal character in the written language. The Swedish old-fashioned *ej* (Eng. "not") was used in writing instead of the everyday spoken word *inte*. The verb *kontrollera* [Eng. "control"] was used instead of the spoken and shorter *kolla* [Eng. "check"]. This could have been because of the sense of formality given to the act of writing down the practice. Additionally, the variations in spelling and other aspects could potentially be attributed to the multiple authors involved, each with their own backgrounds, experiences, and levels of comfort with writing. None of these formal ways of writing seemed to disturb the functions of the sheets. Collaborative writing is intricately related to the continuous discourse surrounding work within all teams. Whoever was working when the negotiation over the standard and the documentation in writing were updated contributed to the form the language took.

The team leader and assistant team leader played a crucial role in upholding the texts within the company's standardization efforts by maintaining continuous communication with the operators, which included activities such as work observations. The evolving nature of the sheets can be perceived not only through the incorporation of numerous sticky notes but also through tangible evidence found in specific notes within the texts themselves. In one place where the history of safety could have been written, there was a note in the form of the abbreviation "Upd." [Swe. *Uppd.* for *uppdatera*; Eng. "update"] to mark a wish to revise the sheet because of a problem in assembly. Another sign of the evolving nature of the texts was the presence of three question marks following a description of the operation, highlighting an element of uncertainty or the need for further clarification.

To sum up, despite the language on the Element sheets consisting of a combination of standardized language and operators' own vernacular, characterized by grammatical, spelling, and punctuation errors, metaphors, and local vocabulary, the Element sheets were still considered well-functioning in the context of their use. The results showed that the texts were not decontextualized in most functions, as transcripts can be in other fields (cf. Gilbert and Heydon, 2021), but one function of the documents was educational: they were used in the training of new co-workers, and this was where standardizations in language became crucial.

Paradoxically, while standardization aimed to enhance efficiency and engage workers across the entire factory, it did not always effectively fulfill its educational function. Conversely, oral metaphors used in conversations proved to be helpful. In the described situation involving numerous tubes, the capturing of operators' dialogue was prohibited, leading to a detriment in both instructional and educational functions. Furthermore, the documents labeled "Work instruction," which included article numbers for parts, were disconnected from their practical application and production processes, resulting in not updated information.

4. Discussion

The aims of our study in the information design field were to highlight the meaning-making in language as being socially situated in its context. Entering the study, we asked ourselves the question, what are the consequences for information design's idea of text design when meaning-making in a language is understood as being socially situated in an activity? In our case, the continuous capture of the best practice, focusing on the smallest units in the operations, required both knowledge and language from the operators, but this affected the texts. This is where the topic of the transcript-like character becomes relevant; the formulations from the operators provided the texts with a spoken language character, dependent on the immediate context. In this discussion, we combined the theories with the findings to answer our research question concerning the Element sheets used and considered to function as per the documentation, instructions, and educational material. We highlighted the consequences for the design of informative text while also acknowledging the process of writing these sheets as text design.

First, sometimes a *mix of designer and user roles* is necessary, i.e., when certain informative text design needs the user's situated knowledge and "feel" for action, *along* with the user's formulating the action in the language (Schön, 1983), as these are relevant for the design to be used. It involves acknowledging and incorporating the operators' tacit and intuitive knowledge, their innate sense, and familiarity with the assembly process through their cognition, physical touch, and muscular strength, all intertwined with language (Schön, 1983). The operators are far from the passive position of being "given a set of instructions" (Delin, 2000, p. 59) by a distant designer. The operators rather take on a "makeshift-user role" (cf. McHardy et al., 2012) and cannot be categorized as the "other" to the designer within a theory of user-subordination.

Second, the language has intertexts from the workplace setting (Bremner, 2018), protected by team leaders and assistant team leaders. We observed that the sheets' learning function could suffer from the writings' closeness to talk. According to this understanding, text design needs continuous *protection of intertexts in the continuous change of spoken language*. Teamwork becomes important here. Intertextuality can ensure the possibility of also communicating with new co-workers or operators from another part of the factory. There is no proper decontextualization of the readers of the sheets, but there have been some efforts to strive for formalization so that others, later in time and/or new to the context in a learning situation, will be able to understand the transcripts. There is a first short distance, one might say, between, on the one hand, an operation and the talk about it negotiated into a text, and, on the other hand, a new reader in time. This small distance shows the struggle between spoken language about the uniqueness of an activity and a formalization of the language referring to it. The case, however, also shows that standardization sometimes sweeps away the most precise and efficient discussions using metaphors. Moreover, on another assembly line, they may never use the word "enter" [Swe. *Äntra*"], as in "enter the nut". A new reader would still recognize the template and the fixed terms, and through supervision, the activity and talk would make the transcripts clearer. In addition, the team leader and the assistant team leader ensured that local terms that become a problem in communication were replaced.

Third, some text design practices could be understood as *collaborative writing based on negotiations*. Despite the influence of discussing assembly and reflection on action (Schön, 1983), which resulted in a telegraphic writing style (Delin, 2000), the presence of "strategic talks" (Schneider, 2002; see also Blakeslee and Savage, 2013) contributed to enhanced clarity in the communication for others. This clarity was achieved through the negotiation process among operators and team leaders. The clarity in the text emerges through a collaborative "struggle" of discussions (McHardy et al., 2012). This collaboration is part of the close struggle for continuous improvements and formulating new methods, even at the smallest level of tasks, known as the minima, among the operators who possess ownership of the *Element sheets*. To understand collaborative writing within this context, it is crucial to recognize the negotiation that takes place regarding both content and language in texts, highlighting the significance of this approach in text design.

Fourth, recognizing the *becoming character of texts*. The work with the Element sheets is an ongoing work, a becoming of design work (Schneider, 2002; Baerenholdt et al., 2012; Olsen and Heaton, 2012; Roth et al., 2017).

In addition to Roth et al. (2017) concept of becoming design, which emphasizes the interaction and correspondence between the designer and material in a specific setting, we want to emphasize the continuous and collective nature of the texts themselves as they evolve and develop through ongoing collaborative writing (Lee, 2008; Bremner, 2018). This practice thus challenges the idea of a finished design product at the end of a design process that can be evaluated or tested. In our case, the evaluation and testing of the texts within the design context itself occurred through daily work observations or when issues arose and discrepancies between the text and actual activities were identified. Rather than relying on separate evaluation stages, the assessment and refinement of the texts were integrated into the ongoing design process within the daily work environment.

The fundamental essence of design activity, which is open to change, is extended to a continuum. If the manufacturing company needs the manual assembly and capturing of best practices, the focus lies on the continuous evolution of the text itself rather than a fixed and final version. This aligns with the objective of the text becoming a dynamic entity. The evolving nature of the texts is also linked to the challenge of unclear genre categorization within their specific contexts of use (compared with principles of how to write, e.g., "instructions", and so on). The inscribed objects and ongoing writings are "intertwined forms of linguistic, prosodic, bodily" (Due, 2020) information.

Fifth, the design process is *not linear but a realm,* which we have already touched upon. The practice we have studied closes much of the distance between the designer's abstract space and the user's concrete space (cf. Lee, 2008). Based on the case study, we deduced an image of a situated activity realm that features the ability to capture the minima in an ongoing collaborative text-design work with texts of somewhat unclear genres, having different authors, functions, and intertexts from the organizational frame, oral conversations, and the "feel" for the activity at hand. The text results cannot meet the information-design generic claim of what "good" design is. However, the results are complex text designs that satisfy many of their functions. By adding a living description of the operations in the Element sheets to be used by the team, it can easily be understood how the text can have meaning in practice while not always having a coherent style.

Finally, in the information design field, it would be interesting to draw on more cases with knowledge from both the applied linguistic fields and the design discipline at large. Moreover, it would be important to continue the discussion of the boundaries between informative text design and information design. Concerning our case, it should be noted that there was much work with the Element sheets. The continuous conversations and updates to the language in the sheets constituted a demanding job. We wrote earlier that acknowledgment of the minima in the functional texts appears to be a need in our case while following the company's standardization of work. However, because of the challenging workload and the developments in digitalization in Industry 4.0 (European Union, 2015), the instructional texts risk being produced in, to use Lefebvre's (2003) words, an abstract space before ending up on a digital screen at the assembly. For future research, it would be interesting to investigate how Industry 5.0, focusing on a human-centric industry (European Commission, 2023), considers the operators' best practices and their formulations in language about methods during digitalization processes.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical review and approval was not required for the study involving human participants in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants in accordance with the national legislation and the institutional requirements.

Author contributions

A-LC has collected and analyzed the data, created the theoretical framework, and written the article. NS contributed by collecting the data. All authors contributed to the article and approved the submitted version.

Funding

This work was funded by The School of Innovation, Design, and Engineering at Mälardalen University in Eskilstuna, Sweden (A-LC) and the Knowledge Foundation (NS).

Acknowledgments

The authors wish to thank the School of Innovation, Design, and Engineering at Mälardalen University in Västerås and Eskilstuna, Sweden, for the opportunity for collaboration between an information design researcher and a Ph.D. student in engineering at the time of the data collection. We also wish to thank the three reviewers from *Frontiers* for their valuable comments on an earlier version of the article. Their insights and efforts were highly appreciated.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Baerenholdt, J. O., Buscher, M., Damm Scheuer, J., and Simonsen, J. (2012). "Perspectives on design research," in *Design Research: Synergies from Interdisciplinary Perspectives* eds J. Simonsen, J. O. Baerenholdt, M. Büscher, and J. D. Scheuer (Abingdon, VA: Routledge).

Bezemer, J., and Kress, G. (2016). *Multimodality, Learning and Communication: A Social Semiotic Frame*. London; New York, NY: Routledge.

Blakeslee, A. M., and Savage, G. J. (2013). "What do technical communicators need to know about writing?" in *Solving Problems in Technical Communications*, eds J. J. Johndan-Eilola, and S. A. Selber (Chicago, IL: University of Chicago Press), 362–385.

Bremner, S. (2018). *Workplace Writing: Beyond the Text.* Abingdon, VA: Routledge.

Coates, K., and Ellison, A. (2014). *An Introduction to Information Design*. London: Laurence King.

De Saussure, F. (2015). *Cours de linguistique générale*. Lund: Arkiv förlag.

Debs, M. B. (1991). Recent research on collaborative writing in industry. *Tech. Commun.* 38, 476–484.

Delin, J. (2000). *The Language of Everyday Life: An Introduction*. London: Sage.

Due, B. L. (2020). Respecifying the information sheet: an interactional resource for decision-making in optician shops. *J. Appl. Linguist. Prof. Pract.* 14, 127–148. doi: 10.1558/jalpp.33663

European Commission (2023). *What is Industry 5.0?* Avaailable online at: https://research-and-innovation.ec.europa.eu/research-area/industrial-research-and-innovation/industry-50_en (accessed August 07, 2023).

European Union (2015). *Industry 4.0. Digitalization for Productivity and Growth. A Briefing September 2015*. Available online at: https://www.europarl.europa.eu/RegData/etudes/BRIE/2015/568337/EPRS_BRI(2015)568337_EN.pdf~2023-05-19 (accessed May 19, 2023).

Frascara, J. (2015). "What is information design?," in *Information Design as Principled Action*, ed J. Frascara (Champaign, IL: Common Ground).

Gilbert, D., and Heydon, G. (2021). Translated transcripts from covert recordings used for evidence in court: issues of reliability. *Front. Commun.* 6, 779227. doi: 10.3389/fcomm.2021.779227

Halliday, M. A. K. (1978). *Language as Social Semiotic: The Social Interpretation of Language and Meaning*. London: Edward Arnold.

Hyland, K. (2016). *Teaching and Researching Writing, 3rd Edn. Applied Linguistics in Action Series*. New York, NY: Routledge.

Jacobson, R. (2000). *Information Design*. Cambridge, MA: MIT Press.

Kirkman, J. (2005). *Good Style: Writing for Science and Technology, 2nd Edn*. New York, NY: Routledge.

Kress, G., and van Leeuwen, T. (2006). *Reading Images: The Grammar of Visual Design. 2nd Edn*. London; New York, NY: Routledge.

Landgrebe, J., and Rye Marstrand, R. (2020). Sticking to the rules: post-its and members' competence in a Lean meeting activity. *J. Appl. Linguist. Prof. Pract.* 14, 172–199. doi: 10.1558/jalpp.33806

Lee, Y. (2008). Designer participation tactics: the challenges and new roles for designers in the co-designing process. *CoDesign* 3, 31–50. doi: 10.1080/15710880701875613

Lefebvre, H. (2003). *The Urban Revolution*. Minneapolis, MN: University of Minnesota Press.

Liker, J. K. (2004). *The Toyota Way*. New York, NY: McGraw-Hill.

Liker, J. K., and Meier, D. (2006). *The Toyota Way Fieldbook: A Practical Guide for Implementing Toyota's 4Ps*. New York, NY: McGraw-Hill.

McHardy, J., Wolf Olsen, J., Southern, J., and Shove, E. (2012). "Makeshift users," in *Design Research: Synergies from Interdisciplinary Perspectives*. Abingdon, VA: Routledge 95–108.

Merriam, S. B. and Tisdell, E. J. (2016). *Qualitative Research. A Guide to Design and Implementation, 4th Edn*. San Francisco, CA: Jossey-Bass.

Nystrand, M. (1989). A social-interactive model of writing. *Written Commun.* 6, 66–85. doi: 10.1177/0741088389006001005

Olsen, P. B., and Heaton, L. (2012). "Knowing through design," in Design Research: Synergies from Interdisciplinary Perspectives, eds J. Simonsen, J. O. Baerenholdt, M. Büscher, and J. D. Scheuer (Abingdon, VA: Routledge), 79–94.

Pettersson, R. (2002). *Information Design*. Amsterdam: John Benjamins.

Prior, L. (2020). Inscribed objects and professional practices: a postscript. *J. Appl. Linguist. Prof. Pract.* 14, 304–308. doi: 10.1558/jalpp.40428

Rorty, R. (1992). *The Linguistic Turn: Essays in Philosophical Method*. Chicago, IL: University of Chicago.

Roth, W. -M., Socha, D., and Tenenberg, J. (2017). Becoming-design in corresponding: Re/theorising the co- in codesigning. *CoDesign*. 13, 1–15. doi: 10.1080/15710882.2015.1127387

Schneider, B. (2002). Clarity in context: rethinking misunderstanding. *Tech Commun.* 49, 210–218.

Schön, D. A. (1983). *The Reflective Practitioner: How Professionals Think in Action*. New York, NY: Basic Books.

Simonsen, J., and Hertzum, M. (2012). "Iterative participatory design," in *Design Research: Synergies From Interdisciplinary Perspectives*. Abingdon, VA: Routledge, 16–32.

Simonsen, J., Svabo, C., Strandvad, S. M., Samson, K., Hertzum, M., and Hansen, O. E. (eds.). (2014). *Situated Design Methods*. Cambridge, MA: MIT Press.

frontiers | Frontiers in Communication

TYPE Brief Research Report
PUBLISHED 27 October 2023
DOI 10.3389/fcomm.2023.1178516

Check for updates

OPEN ACCESS

EDITED BY
Mila Vulchanova,
Norwegian University of Science and Technology, Norway

REVIEWED BY
Dian Dia-an Muniroh,
Universitas Pendidikan Indonesia, Indonesia
Luna Filipovic,
University of California Davis, United States

*CORRESPONDENCE
Kate Haworth
✉ k.haworth@aston.ac.uk

RECEIVED 02 March 2023
ACCEPTED 09 October 2023
PUBLISHED 27 October 2023

CITATION
Haworth K, Tompkinson J, Richardson E, Deamer F and Hamann M (2023) "For the Record": applying linguistics to improve evidential consistency in police investigative interview records. *Front. Commun.* 8:1178516. doi: 10.3389/fcomm.2023.1178516

"For the Record": applying linguistics to improve evidential consistency in police investigative interview records

Kate Haworth[1]*, James Tompkinson[1], Emma Richardson[2], Felicity Deamer[1] and Magnus Hamann[2]

[1]Aston Institute for Forensic Linguistics, College of Business and Social Sciences, Aston University, Birmingham, United Kingdom, [2]School of Social Sciences and Humanities, Loughborough University, Loughborough, United Kingdom

The "For the Record" project (FTR) is a collaboration between a team of linguistic researchers and police in the England & Wales jurisdiction (E&W). The aim of the project is to apply insights from linguistics to improve evidential consistency in police interview transcripts, which are routinely produced by transcribers employed by the police. The research described in this short report is intended as a pilot study, before extension nationally. For this part of the project, we analysed several types of data, including interview audio and transcripts provided by one force. This identified key areas where current transcription practise could be improved and enhanced, and a series of recommendations were made to that force. This pilot study indicates that there are three core components of quality transcription production in this context: Consistency, Accuracy, and Neutrality. We propose that the most effective way to address the issues identified is through developing new training and guidance for police interview transcribers.

KEYWORDS

transcript, interview record, police interview, investigative interview, language as evidence, forensic linguistics, applied linguistics

1. Introduction

The FTR project applies linguistic findings to the process of producing written transcripts of police investigative interviews with suspects. The current standard procedure is that these interviews are audio recorded, then for any case which will proceed to court,[1] a transcript is produced by administrative staff employed by the relevant police force. This process is of particular importance given that these are evidential documents, presented in court as part of the prosecution case, yet we know from linguistics that original spoken data are necessarily substantially altered through the process of being converted into written format (see below). Yet once a transcript or ROTI (Record of Taped Interview) has been produced, it is generally heavily relied upon rather than the audio recording, making its accuracy all the more important.

The overall objective of this research is to substantially increase the accuracy and consistency of investigative interview evidence, especially in terms of the representation of spoken language features. Our aim is to enable transcribers to produce interview records

[1] For our pilot force, this now only applies to cases which will be heard in the Crown Court, but there appears to be variation in practice across forces.

which encapsulate more of the meaning conveyed by the original spoken interaction, and to enable consistency of interpretation of features such as punctuation and pauses for the reader (i.e., investigating officers, Crown Prosecution Service, courts, juries), thus removing a major source of subjective and potentially inaccurate interpretation of criminal evidence. We emphasise that the intended outcome is not the production of a "perfect" transcript, since this is an impossibility. Instead, the intention is to reduce the "contamination" or distortion which transcription can introduce, and to raise awareness in legal contexts of the fundamental limitations of transcripts.

2. Rationale

In E&W, before the full national implementation of the Police & Criminal Evidence Act 1984 (PACE), written records of interviews with suspects were created by the interviewer after the event, based on any contemporaneous notes which had been made during interview, and on their own memory. A series of infamous miscarriages of justice (e.g., Bridgewater Four, Derek Bentley; see e.g., Coulthard, 2002) shone a harsh light on this practise, proving that these records could not only be highly inaccurate, but even completely fabricated. PACE therefore introduced the mandatory audio recording of all interviews with suspects (with only a handful of exceptions, e.g., in terrorism cases). This was of course a substantial improvement to policing practise, and one in which E&W has led the way internationally. Audio-recorded interviews have subsequently been treated as the solution to the problem of inaccurate or unreliable interview evidence; however, that is not entirely the case. In fact, it gives rise to another potential source of contamination or distortion, through the production of the written record of the interview. Although the audio (or video) recording is always available, in practise the written transcript is heavily relied upon once it has been produced (see Haworth, 2018). The written record becomes a central piece of criminal evidence, passed on to the Crown Prosecution Service (CPS) as part of the case file, then presented as part of the prosecution case in court, thereby being routinely presented to the jury as part of the package of evidence on which they must reach their verdict. Juries are of course free to make whatever judgments they wish of these materials; we do not seek to interfere with this. Our concern here is simply that any evidence presented to the court should be as accurate and unaltered as possible.

However, we know from decades of research in linguistics that it is not possible to convert spoken language into a written text without changing it. Linguistic research has indicated that spoken and written modes are essentially different "languages;" they are non-equivalent (e.g., Biber, 1988; Halliday, 1989). Conversion from one to the other is therefore almost like a process of translation and interpretation; this means it is necessarily subjective and inexact. The challenges of transcribing spoken data have in fact long been addressed as a methodological challenge by linguists (see e.g., Ochs, 1979; Edwards and Lampert, 1993; Leech et al., 1995; Bucholtz, 2007, 2009), since we ourselves often need to create written records of the spoken data we record for our research. This has been a particular methodological concern in Conversation Analysis (e.g., Jefferson, 2004; Hepburn and Bolden, 2012). This work shows that transcription is actually a very complex and challenging task, if it is to be done accurately and fairly. A particular problem identified is that it is impossible for any transcriber not to bring in their own perspectives and unconscious biases; in fact Bucholtz (2007) describes transcription as "an inherently and unavoidably sociopolitical act" (p. 802).

Yet transcription of speech routinely occurs in various legal contexts, several of which have been studied by linguists. All such studies have found serious problems with the official transcripts produced. This includes studies of transcripts of courtroom proceedings (e.g., Walker, 1986, 1990; Eades, 1996; Tiersma, 1999, p. 175–99), covert recordings (e.g., Shuy, 1993, 1998; Fraser, 2014, 2018, 2022), and interpreted interviews (e.g., Filipović, 2022); see also our own prior work which informs this project (Haworth, 2018; Richardson et al., 2022).

All of the above research background indicates a strong likelihood that official transcripts of police investigative interviews may not be as accurate and balanced as is generally taken for granted. This is even more the case when we consider that most Records of Taped Interview (ROTI) involve a good deal of editing and summarising, rather than being an attempt to provide a "full," "verbatim" transcript. Editing, or summarising, is a highly selective and subjective process, with the summariser having to make choices as to what to include and what to omit. This process has not been the subject of sustained prior research (although see Haworth, 2018; Filipović, 2022).

Despite these clear warning signs from the linguistic research, none of this has yet made its way into professional practise within the legal system. In fact, not only are the potential problems not recognised, it has actually been built into practise through case law[2] and legislation[3] that tapes, transcripts, and summaries should be treated as interchangeable, and in essence identical. Our starting point for this project, then, is that potentially serious contamination of interview evidence is currently routinely overlooked and unrecognised; but also that linguistic research and analysis can readily be applied in order to redress this.

3. Method

Given that interview records are produced within force, and the process varies from force to force,[4] we chose to work with one force first as a pilot project. This enabled us to conduct detailed analysis across all aspects of the process, from multiple angles and methodological approaches; in other words to prioritise depth over breadth. It enabled us to take into account specific local practises, and also ensures that our findings are as relevant as possible to our partner force. We collected two types of data from our partner force: (1) interview recordings and their corresponding official transcripts; and (2) practitioner input through focus groups and an online questionnaire. Our research questions for these data were:

[2] *R v Rampling* [1987] Crim LR 823.

[3] s.133 & 134(1) Criminal Justice Act 2003.

[4] As revealed through FOI enquiries made by us, and the lack of any national guidance.

- How are written records of interview currently produced and used in this force?
- Is there an unrecognised problem regarding evidential consistency in those records?

Alongside this, we conducted experiments to test our hypotheses around the changes in format of the data (i.e., changing from spoken to written, and transcription choices) having an effect on its interpretation. This was to ensure that there was a sound evidence base for any recommendations we made.

The project thus involved three strands, each with its own methodological approach and data, but which were interrelated with each informing the other as the project progressed. Findings from all three strands were then combined into one unified analysis, through which key themes were identified. As an overall objective, we sought to investigate what insights from linguistics can offer in terms of improving the process.

3.1. Experiments

Our experiments were designed to do two key things: (1) test the assumption that people treat audio and written information similarly, and (2) examine how changes in the representation of different linguistic features could influence the way people think about the information contained within transcripts.

In an initial experiment (see Deamer et al., 2022), a 3-min clip of a publicly-released police interview with a suspect in a UK murder enquiry, sourced from You Tube, was used to elicit views about the interviewee from participants, recruited using convenience sampling (data provided by our partner force were not used due to data protection and confidentiality). A total of 30 adult participants heard the original audio recording; 30 saw a written transcript of the same extract (groups were matched for gender and age). The transcript was produced by the research team with the aim of including as much detail as possible, while also maintaining legibility for a lay audience. Participants were then presented with a series of questions (quantitative and qualitative) to determine their interpretation of the interview, and the interviewee. We wanted to assess whether there would be any differences in the judgements of those who heard the audio compared with those who read the transcript. Responses to questions about what, in the language, had led participants to give their answers, enabled us to identify specific features which may have influenced participants' perceptions.

We then ran a second experiment which further explored these issues (see Tompkinson et al., 2023). Using the same interview data, but additionally manipulating one variable which both prior research (e.g., Nakane, 2007, 2011; Heydon, 2011) and the qualitative findings of the first experiment indicated to be of interest, we created versions of the transcript which represented pauses/silent hesitations in different ways. This experiment was much larger, eliciting responses from 250 participants, recruited via Prolific.[5] Again, we tested whether changing the mode of representation (audio vs. transcript) would affect participants' perceptions, and we also wanted to assess whether the different representations of pauses would impact the judgements that people were prepared to make about the interviewee.

3.2. Linguistic analysis of interview data

A total of 25 recent audio-recorded suspect interviews and 4 video-recorded witness interviews,[6] ranging from 6 to 92 min, and their accompanying transcripts, were provided for analysis by the force under a Data Processing Agreement, and with ethical approval from Aston University. The original data were redacted, anonymised and pseudonymised on police premises. A comparative analysis was undertaken of the interactional activities captured by the audio recording, and what was represented in the written records (see Richardson et al., 2023). This involved close qualitative linguistic analysis informed predominantly by Conversation Analysis. This enables us to identify the social actions that are performed by speakers as they interact, and to evidence the substantial changes that can occur in the process of transforming the spoken interaction into a written representation. In particular, this makes features of the talk which go beyond the words spoken accessible and analysable, including through documenting them through detailed technical transcripts (following Jefferson, 2004).

3.3. Questionnaires and focus groups

An online, anonymous questionnaire was completed by the full cohort of force transcribers at date of completion ($n = 9$), covering basic aspects of their job and their approach to transcribing, along with a very short transcription task. Focus groups with transcribers ($n = 6$) and police interviewers ($n = 13$), recruited as volunteers via our internal force contact, were subsequently conducted on police premises across 3 sites, to minimise participant inconvenience. These were held separately, thereby amounting to 6 focus groups and over 11 h of audio-recorded data. This was anonymised and transcribed, and a thematic analysis undertaken using NVivo. Once the main research was concluded the research team returned to the force for two further focus groups, at which we presented our main findings and proposed recommendations, inviting feedback and discussion. These return focus groups combined both transcribers and interviewers from the original focus groups, enabling direct discussion between these cohorts.

4. Results

The FTR project has produced a large volume of research findings. More detailed findings of the individual project strands are available in Deamer et al. (2022), Richardson et al. (2023), and Tompkinson et al. (2023), with more to follow. Detailed combined findings and outcomes from the FTR project as a whole will also

[5] Available online at: https://www.prolific.com/.

[6] The use of witness interviews in the legal process is very different to suspect interviews, especially in terms of their presentation as evidence in court. However, we included these in this strand of the project as part of our analysis of current transcription practices, since they are produced by the same transcribers in the same conditions.

be published in due course. The key combined findings can be summarised as follows:

- Transcribers are highly aware of the stakes and the potential consequences of their work, and they take this very seriously, aiming to produce balanced and fair records. However, numerous aspects of current transcription practise undermine this aim.
- The transcribers receive no training in transcription. Instead, they report relying on their peers for *ad hoc* support; practise has thus developed within-group, without official input or oversight. They also receive very little, if any, feedback on the transcripts they produce. Bad or inappropriate practise can therefore easily become embedded at a local level, and there is no mechanism for ensuring consistency. There is also no established checking procedure, and therefore no system in place to catch errors and mistakes.
- For the parts of the interview rendered "verbatim"/"in full," we did not find systematic or widespread problems with the basic accuracy of recording the bare words spoken. However, some errors were found, including simple "typos" but also instances where content was apparently misheard, leading to incorrect transcription. Such errors may not be common, but they can be of real significance: we identified at least two instances where meaning was affected regarding important evidential points. For example, one transcript included "he met someone knew." This confuses two very different propositions, with opposite meanings: "he met someone *he knew*," or "he met someone *new*." It is not possible to work out which was meant from this transcript alone.
- There was variation in use of the standard layout on the interview transcript pro forma, which in places could give rise to unintended interpretations. As well as consistency, it gives rise to questions of neutrality, given that these involve subjective decisions on the part of the transcriber. For example, the most common practise was to use a new text box for a new speaker's turn. But we also found examples of turns being split into more than one box, which has the effect of visually highlighting a particular part of that turn, creating a risk that that part is taken out of context and thereby misinterpreted. For example, an apparently incriminating admission was "highlighted" in this way, but had been separated from the very important conditional it followed on from: the interviewee stated that they didn't know what had happened and had no memory of doing the act they were accused of, but then said "if there's enough evidence to say I've done it I'll put my hands up and say || yeah I've done it." These final words were presented on a new line in a new text box with the timing also given alongside, all of which gave them arguably undue prominence.
- Consistency was found to be a key issue. There was a lack of consistency in the way that different transcribers represented different aspects of speech in the transcripts, giving rise to potential confusion as to what was meant. There were also instances of inconsistency within the same transcript. For example, several different methods were observed to be used to represent inaudible parts of the recording, such as "\\\ unintelligible"; "inaudible"; "……". As an added complication, the same resource was found to be used to represent different features. For example, a series of dots ("…..") was used to indicate four different phenomena: transition from one mode of transcribing to another (e.g., summary to "verbatim"); silence; cut-off talk; and overlapping speech. Unsurprisingly, interviewers reported a range of interpretations of this feature when they encounter it in their interview transcripts, demonstrating that meaning is being lost due to this practise. One interviewer described having to go back to the transcriber for clarification of the meaning of "… " in one case, demonstrating how transcription inconsistency is giving rise to inefficiency.
- Another key identified area of inconsistency was in the representation of pauses/silence. This is of importance given that these can be highly significant interactionally (e.g., Nakane, 2007), and thus create meaning for listeners, as borne out in our experimental findings. Our finding that pauses were either omitted, or transcribed inconsistently, in our dataset is therefore a cause for concern.
- Emotion is not represented in the transcripts in our dataset. We use the term "emotion" here to cover a broad range of audible non-verbal aspects of a person's talk, such as laughter or crying. The display of emotion is a crucial part of human social interaction, conveying a great deal of additional meaning beyond the bare words spoken. This was borne out in our experimental findings, with numerous participants commenting on displays of the interviewee's emotion either as heard in the audio or represented in the transcript. The omission of emotion from transcripts can therefore have serious consequences, especially where the emotional state of the interviewee becomes relevant evidentially. This is a phenomenon with which interviewers are very familiar, as reflected in several case examples discussed in the focus groups, including interviewee displays of anger and loss of emotional control. Interestingly, it is also well recognised by the transcribers, which begs the question as to why they do not include such details. The main answer that arose from the focus groups was that it is often mistakenly viewed as being subjective, when they are aiming to be as objective as possible. However, what is currently not recognised is that omission is a subjective choice in itself, affecting the meaning conveyed. The transcribers may have the right intentions, but current practise is arguably achieving the opposite outcome to that desired.
- However, our experimental work indicates that determining the most appropriate way to represent such features in a transcript is not as straightforward as first envisaged, and further work is therefore required before firm recommendations can be made as to best practise and standardisation of interview transcription.
- The process of summarising, rather than writing everything said "verbatim," has a substantial impact on the official record. Transcribers are not provided with specific guidance or training about how to summarise information, or about what to include. Instead, they are left to attempt to identify the most evidentially relevant details themselves, without any legal training or experience. There was extensive use of summaries

across the transcripts analysed, and we found a wide variety of practise, with once again an overall lack of consistency. In addition, the requirement for the use of a reporting verb when producing such summaries ("Smith *said/claimed/insisted...*") introduces a further avoidable element of subjectivity and transcriber interpretation. Further, the fact that the questioning sequence is often not preserved in the transcript is a source of frustration for interviewers, who may well have had specific tactical and evidential reasons for including certain aspects whose significance is (understandably) not recognised by the transcriber and therefore omitted from the record.

- Interviewers reported viewing transcripts as an inadequate reflection of the actual interview interaction, and therefore tend not to use them as an investigative tool. Instead, they may rely on their notes and memory of the interview. This is a risky practise and of some concern.
- Overall, the strong message from the focus groups with both transcript producers and users is that official transcripts currently do not capture interviews effectively. Practitioners are aware of some inaccuracies in what was said, but mainly recognise a failure to capture how it is said. There was strong support for standardisation and training being introduced.

Overall, we conclude that the current process for producing interview records in this force does result in problems with evidential consistency. In other words, this type of evidence undergoes alteration as it is processed; something which would likely not be considered acceptable for physical evidence, for example. However, do these types of changes actually matter in practise? Our experimental work sought to address this.

- Our experimental findings demonstrate how the format in which police interview evidence is presented can significantly affect how it is interpreted, supporting our basic point that converting interview evidence into written format can significantly alter how that evidence is perceived. This demonstrates the importance of the factors identified above, and the potentially serious implications, particularly for the use of interview transcripts as evidence in court.
- Our initial experiment (Deamer et al., 2022) found a range of significant differences between judgements of the interviewee depending on whether participants were presented with an audio recording or transcript of the interview. Those who read the transcript perceived the interviewee as *more anxious, less relaxed, more agitated, more nervous, more defensive, less calm, less cooperative* and, perhaps most importantly, less likely to be telling the truth [$\chi^2_{(1)} = 4.022$, $p = 0.045$]. Participants identified a range of language and speech features which influenced these perceptions of the interviewee.
- Our expanded second experiment (Tompkinson et al., 2023) replicated these findings, again showing significant differences across judgements of the interviewee between the Audio and Transcript conditions. In this study, the interviewee was judged as being significantly less *credible*, *plausible*, *sincere*, *cooperative*, *calm*, *friendly* and *relaxed* by participants who read the transcript, as well as significantly more *agitated*, *nervous*, *surprised* and *panicked*. The interviewee was also significantly more likely to be judged as not telling the truth if the person making the judgement read a transcript as opposed to listening to the audio recording [$\chi^2_{(2)} = 23.82$, $p < 0.001$], with a similar number of participants using the "don't know" option in both conditions. Overall, these findings show the clear potential for instability in perception between audio recordings and transcripts of the same interview data.

In order to address the issues identified through our research, we have created a set of criteria which encapsulate our findings, using terms which are readily understandable and applicable by a non-linguistic, non-technical user group: consistency, accuracy and neutrality (CAN). We propose these three areas as the foundational features that should underpin any police interview transcript. Our key recommendation is the introduction of training and guidance to embed the CAN model into police transcription practises; however further research is required to assess its applicability beyond our pilot force.

5. Discussion

Overall, this project has demonstrated that transcription practises certainly do matter in this context. The way in which police interview evidence is presented can have a substantial effect on how it is perceived and interpreted, to the point of altering whether receivers believe an interviewee is telling the truth or not. Such differences should not occur in the presentation of criminal evidence. Likewise, accuracy and consistency should be expected as minimum requirements for official interview transcripts, so that they can be correctly evaluated by readers, especially those tasked with using interview records as part of the evidence on which to base vital decisions about the interviewee's future (e.g., CPS, judge, jury). Yet we have also shown that transcripts are currently less accurate and consistent than we might wish, especially when it comes to the practise of summarising parts of the interview. Leaving such an important evidential task to clerical staff with no legal training—as appears to be standard across the sector—seems especially troubling and risky.

Some aspects of this can readily be addressed, and series of recommendations were produced for our partner force. These are a combination of known good practise, points which emerged from our research, and solutions suggested by police practitioners themselves. This comes with the recognition that many factors extend well beyond the remit of individual police forces and will require national uptake and implementation, which in turn requires extending the scope of the FTR project beyond one force. Some aspects may even require changes to criminal procedure, which we acknowledge is a steep hill to climb. However, we continue to work towards these objectives, through engaging more police forces, national organisations and policy initiatives, and through conducting further research.

Our experimental findings indicate that solutions around introducing transcription standardisation are not as straightforward as we had initially hoped, so our original intention of producing a set of implementable standards cannot yet be realised. However, these findings demonstrate the importance of not making simplistic recommendations based on assumptions, but of instead conducting targeted research in order to provide a sound evidence base for best practise. It should be emphasised that the research presented here is a pilot project, and we hope that it has successfully demonstrated that this is an issue worthy of continuing, fuller study.

Data availability statement

The datasets presented in this article are not readily available because they include highly sensitive and confidential material including police interviews with suspects, and discussions of such material with police practitioners. As such, even in anonymised form it is considered too sensitive for public access. Requests to access the datasets should be directed to k.haworth@aston.ac.uk.

Ethics statement

The studies involving human participants and police data were reviewed and approved by the Aston Institute for Forensic Linguistics Research Ethics Committee, Aston University. Written informed consent for participation in the research was provided by the participants of the experimental study, focus groups and questionnaire. The police interviews were provided for analysis by the force under a Data Processing Agreement and were fully redacted, anonymised, and pseudonymised.

Author contributions

KH devised and led the project overall, conducted the questionnaire and focus groups, and produced the synthesised analysis of the combined project findings. KH is the main author of this report. JT took over the experimental strand from FD in January 2022, conducted the second and subsequent experiments, and contributed substantial additional analysis across the project. ER led the interview data analysis strand and contributed additional analysis across the project. FD devised the experimental strand and conducted the first experiment. MH contributed to the design of the interview data analysis strand and made a substantial contribution to the analysis of the interview data. In drafting this text, a full 60-page report was initially written by KH, which includes analysis and writing from all team members. From this document JT wrote a 2-page overview summary, which KH then expanded into this short report. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

The research was funded as part of an award to the Aston Institute for Forensic Linguistics by Research England's Expanding Excellence in England (E3) fund.

Acknowledgments

We gratefully acknowledge the support and cooperation of the police force with which we collaborated for this research project, especially all participants and data providers. Without their dedication, commitment, and enthusiasm, this project would not have happened. We are respecting their wish to remain anonymous.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511621024

Bucholtz, M. (2007). Variation in transcription. *Disc. Stud.* 9, 784–808. doi: 10.1177/1461445607082580

Bucholtz, M. (2009). Captured on tape: Professional hearing and competing entextualizations in the criminal justice system. *Text Talk* 29, 503–523. doi: 10.1515/TEXT.2009.027

Coulthard, M. (2002). "Whose voice is it? Invented and concealed dialogue in written records of verbal evidence produced by the police," in *Language in the Legal Process,* ed. J. Cotterill (Basingstoke: Palgrave Macmillan), 19–34. doi: 10.1057/9780230522770_2

Deamer, F., Richardson, E., Basu, N., and Haworth, K. (2022). For the Record: Exploring variability in interpretations of police investigative interviews. *Lang. Law/Linguagem Direito.* 9, 23. doi: 10.21747/21833745/lanlaw/9_1a2

Eades, D. (1996). "Verbatim courtroom transcripts and discourse analysis," in *Recent Developments in Forensic Linguistics, eds.* H. Kniffka, S. Blackwell, and M. Coulthard (Frankfurt am Main: Peter Lang GmbH), 241–254.

Edwards, J. A., and Lampert, M. D. (1993). *Talking Data: Transcription and Coding in Discourse Research*. Hillsdale, NJ: Lawrence Erlbaum.

Filipović, L. (2022). The tale of two countries: Police interpreting in the UK vs. in the US. *Interpreting* 24, 254–278. doi: 10.1075/intp.00080.fil

Fraser, H. (2014). Transcription of indistinct forensic recordings: problems and solutions from the perspective of phonetic science. *Lang. Law/Linguagem Direito.* 1, 5–24. Available online at: https://ojs.letras.up.pt/index.php/LLLD/article/view/2429

Fraser, H. (2018). "Assisting" listeners to hear words that aren't there: dangers in using police transcripts of indistinct covert recordings. *Austral. J. Forensic Sci.* 50, 129–139. doi: 10.1080/00450618.2017.1340522

Fraser, H. (2022). A framework for deciding how to create and evaluate transcripts for forensic and other purposes. *Front. Commun.* 7, 898410. doi: 10.3389/fcomm.2022.898410

Halliday, M. A. K. (1989). *Spoken and Written Language*. Oxford: Oxford University Press.

Haworth, K. (2018). Tapes, transcripts and trials: The routine contamination of police interview evidence. *Int. J. Evid. Proof* 22, 428–450. doi: 10.1177/1365712718798656

Hepburn, A., and Bolden, G. B. (2012). "The conversation analytic approach to transcription," in *The handbook of Conversation Analysis,* eds. J. Sidnell, and T. Stivers (Oxford: Wiley-Blackwell), 57–76. doi: 10.1002/9781118325001.ch4

Heydon, G. (2011). Silence: Civil right or social privilege? A discourse analytic response to a legal problem. *J. Pragm.* 43, 2308–2316. doi: 10.1016/j.pragma.2011.01.003

Jefferson, G. (2004). "Glossary of transcript symbols with an introduction," in *Conversation Analysis: Studies from the First Generation,* ed. G. Lerner (Amsterdam: John Benjamins). doi: 10.1075/pbns.125.02jef

Leech, G., Myers, G., and Thomas, J. (1995). *Spoken English on Computer: Transcription, Mark-up and Application.* Harlow: Longman.

Nakane, I. (2007). *Silence in Intercultural Communication: Perceptions and Performance*. Amsterdam: John Benjamins Publishing. doi: 10.1075/pbns.166

Nakane, I. (2011). The role of silence in interpreted police interviews. *J. Pragm.* 43, 2317–2330. doi: 10.1016/j.pragma.2010.11.013

Ochs, E. (1979). "Transcription as theory," in *Developmental Pragmatics,* eds. E. Ochs, and B. B. Schiefflen (New York: Academic Press), 43–72.

Richardson, E., Hamann, M., Tompkinson, J., Haworth, K., and Deamer, F. (2023). Understanding the role of transcription in evidential consistency of police interview records in England and Wales. *Lang. Soc.* 7, 1–32. doi: 10.1017/S0047404523 00060X

Richardson, E., Haworth, K., and Deamer, F. (2022). For the Record: questioning transcription processes in legal contexts. *Appl. Ling.* 43, 677–697. doi: 10.1093/applin/amac005

Shuy, R. (1998). *The Language of Confession, Interrogation and Deception*. Thousand Oaks, CA: Sage. doi: 10.4135/9781452229133

Shuy, R. W. (1993). *Language Crimes: The Use and Abuse of Language Evidence in the Courtroom.* Blackwell.

Tiersma, P. M. (1999). *Legal Language*. Chicago: University of Chicago Press.

Tompkinson, J., Haworth, K., Deamer, F., and Richardson, E. (2023). Perceptual instability in police interview records. *Int. J. Speech, Lang. Law* 30, 22-—51. doi: 10.1558/ijsll.24565

Walker, A. G. (1986). Context, transcripts and appellate readers. *Just. Quart.* 3, 409–427. doi: 10.1080/074188286000 89041

Walker, A. G. (1990). "Language at work in the law: The customs, conventions, and appellate consequences of court reporting," in *Language in the Judicial Process,* eds. J. N. Levi, and A. G. Walker (New York: Plenum Press), 203–244. doi: 10.1007/978-1-4899-3719-3_7

frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 14 February 2024
DOI 10.3389/fcomm.2024.1281407

Check for updates

OPEN ACCESS

EDITED BY
Dominic Watt,
University of York, United Kingdom

REVIEWED BY
Chiara Barattieri Di San Pietro,
University Institute of Higher Studies in Pavia, Italy
Vincent Hughes,
University of York, United Kingdom
Cristina Aggazzotti,
Johns Hopkins University, United States

*CORRESPONDENCE
Debbie Loakes
✉ dloakes@unimelb.edu.au

RECEIVED 22 August 2023
ACCEPTED 09 January 2024
PUBLISHED 14 February 2024

CITATION
Loakes D (2024) Automatic speech recognition and the transcription of indistinct forensic audio: how do the new generation of systems fare?
Front. Commun. 9:1281407.
doi: 10.3389/fcomm.2024.1281407

Automatic speech recognition and the transcription of indistinct forensic audio: how do the new generation of systems fare?

Debbie Loakes*

Research Hub for Language in Forensic Evidence, School of Languages and Linguistics, The University of Melbourne, Parkville, VIC, Australia

This study provides an update on an earlier study in the "Capturing Talk" research topic, which aimed to demonstrate how automatic speech recognition (ASR) systems work with indistinct forensic-like audio, in comparison with good-quality audio. Since that time, there has been rapid technological advancement, with newer systems having access to extremely large language models and having their performance proclaimed as being human-like in accuracy. This study compares various ASR systems, including OpenAI's Whisper, to continue to test how well automatic speaker recognition works with forensic-like audio. The results show that the transcription of a good-quality audio file is at ceiling for some systems, with no errors. For the poor-quality (forensic-like) audio, Whisper was the best performing system but had only 50% of the entire speech material correct. The results for the poor-quality audio were also generally variable across the systems, with differences depending on whether a .wav or .mp3 file was used and differences between earlier and later versions of the same system. Additionally, and against expectations, Whisper showed a drop in performance over a 2-month period. While more material was transcribed in the later attempt, more was also incorrect. This study concludes that forensic-like audio is not suitable for automatic analysis.

KEYWORDS

forensic linguistics, transcription, automatic speech recognition (ASR), phonetics, artificial intelligence

1 Introduction

This study provides an update on Loakes (2022), which aimed to demonstrate how automatic speech recognition (ASR) systems work with indistinct forensic-like (poor-quality) audio, in comparison with good-quality audio. The original study was motivated by misunderstanding, particularly within the law, around the problem of what is said in indistinct forensic audio being solved automatically. As discussed in that study, this is a question that needs to be explored experimentally, and the current study is intended as confirmation that the unsuitability of ASR for forensic transcription remains, despite recent improvements.

Forensic audio is audio that is generally captured in high stakes and often criminal contexts. This type of audio is defined by Fraser (2022): 8, as

> *…speech that has been captured, typically in a covert (secret) recording obtained as part of a criminal investigation, and is later used as evidence in a trial. Such recordings provide powerful evidence, allowing the court to hear speakers making admissions they would not make openly. One problem, however, is that the audio is often extremely indistinct, to the extent of being unintelligible without the assistance of a transcript.*

The original idea behind the research in Loakes (2022) was to address the fact that computational methods are sometimes seen as a solution to solve the issue of what was said in indistinct recordings. This is part of a wider belief system, dubbed as *technosolutionism* (Morozov, 2013) in which technology is seen as the solution to any problem. Loakes (2022) looked at a poor-quality recording, which was livestreamed via an iPhone, and contained multiple voices with overlapping noise and variable distances from the microphone. That study analysed a good-quality recording by comparison, also recorded via an iPhone, but containing only one speaker who was specifically focussed on being understood. Based on the results of Loakes (2022), it was concluded that AI systems work well when applied to tasks they are designed for—with non-overlapping speech in a language variety the system is familiar with—but poorly when there is background noise, speakers who are not stationary, and when the signal is indistinct, which is all characteristic of forensic audio.

In the short time since that study was published, the availability of more advanced AI systems, especially Open AI's ChatGPT, has changed the artificial intelligence[1] landscape. ChatGPT in particular has received swathes of attention in academic and popular literature. The availability of ASR systems has also risen rapidly, again in particular Open AI's *Whisper*. While there are some critical analyses of artificial intelligence and its role in society (Bender, 2022; Preston, 2022; Bridle, 2023; Perrigo, 2023), there is also still much, less critical, attention on how well these ASR systems work and how much time they save. For example, there are popularly available articles citing Whisper as being 'an ASR model that shows human levels of accuracy and robustness' (Rodriguez, 2022), yet this itself assumes human accuracy is infallible, and anyway the accuracy and robustness appear true only in some limited circumstances.

This study aimed to critically assess the use of Whisper, and some other ASR systems using large language models (e.g., Kallens et al., 2023), to determine how accurate they are in transcribing a section of poor-quality forensic-like audio. Specifically, the aim was to provide new data to compare how the current generation of ASR systems performs when tasked with the transcription of indistinct forensic-like audio (e.g., Loakes, 2022).

2 Background

Automatic speech recognition is not designed for forensic transcription, yet it is often seen by legal professionals as a possible option for the solution of what is being said in indistinct recordings (see, e.g., the discussion in Loakes, 2022). This belief assumes that automatic methods are somehow free from bias and should be more objective than human transcription. However, these automatic systems are of course designed by humans and have in-built biases in their training data (e.g., Koenecke et al., 2020; Wassink et al., 2022). In fact, the more advanced these systems are becoming, the more these inherent biases are also coming to the fore. Talking about ChatGPT's predecessor, GPT-3, Perrigo (2023) notes that its outputs originally involved inappropriate and offensive content, which was then later screened to improve usability by very low-paid workers so its 'huge training dataset was the reason for GPT-3's impressive linguistic capabilities, but was also perhaps its biggest curse'.

In automatic speech recognition, biases are in the direction of better recognition of 'standard' accents (Markl, 2022; Wassink et al., 2022; Harrington, 2023), one or other of male or female voices depending on the system (Markl and McNulty, 2022) as well as non-pathological voices (Benzeghiba et al., 2007; Markl and McNulty, 2022). Additionally, as noted by Benzeghiba et al. (2007), children's voices and elderly voices are also generally not modelled well and cause performance issues with ASR.

It is, nevertheless, important to continue to investigate the issue experimentally to determine limits in ASR performance, as the current study aims to do. In Loakes (2022), the good-quality recording was transcribed well, while the poor-quality recording was not. For example, one of the commercially available systems, Descript, had approximately 96% correct recognition of the good-quality recording and 1.7% correct recognition of speech in the poor-quality recording. That study also demonstrated that using a transcript and trying to align it with speech events using a forced-aligner is replete with problems—the system forces boundaries onto speech events that are not present and may look correct to non-linguists even when it is clearly not. For example, in that study, drumming noise and laughter were aligned with speech events (Loakes, 2022): 9.

Since Loakes (2022) addressed the issue of how ASR copes with indistinct forensic-like recordings, some new work in this space has been conducted with the newer generation of ASR systems which further demonstrates some of the issues discussed above; however, this new research has not made use of *Whisper*. Similar to Harrington et al. (2022), Loakes (2022) carried out a comparison of various ASR systems with recordings known to be difficult for human transcribers, as reported by Love and Wright (2021). They used 18 British English utterances of which they could be certain of the content and used 12 commercially available ASR systems to compare how well the systems transcribed forensic-like audio. They found extreme variability in system responses, ranging from a 70% match across the *Microsoft Transcribe* and the ground truth transcripts, compared to 13.9% for Sonix [which also had low performance in Loakes (2022) for the data analysed in this study]. Harrington et al. (2022) note that errors relate to a degree of phonetic similarity between the error and the actual word spoken, as well as predictability errors from training data. Examples are the word *worrying* mistaken to be *varying* and *chicken tikka masala* (likely low frequency) mistaken to be *she can take*. The authors conclude that managing and interpreting the output of such systems is more effortful than having a human transcribe the data in the first place.

Harrington (2023) considered the use of ASR for police-suspect interviews, with a view to making the process more efficient and

[1] In Loakes (2022) *artificial intelligence* was defined as "intelligence demonstrated machines instead of humans" (c.f. McCarthy, 2007). Other researchers have also noted that artificial intelligence nevertheless has origins in "human contrivance and ingenuity" (Fetzer, 1990).

potentially using human post-editing. She compared three commercial ASR systems (*Rev AI, Amazon Transcribe*, and *Google Cloud Speech to Text*) to assess how they performed across accents and recording qualities. She looked at audio from the DyViS database (Nolan et al., 2009), with Standard Southern British English and West Yorkshire English speakers, using both studio quality files and files with speech-shaped noise added to degrade the signal. Harrington (2023) observed three main kinds of errors with the systems, which involved insertion of material (extra words in the output compared with the transcript), deletion (missing words), and substitution (a mismatch between the reference transcript and the output). She also describes varying levels of success across the systems, noting that errors were higher with West Yorkshire English speakers, whose accents were likely represented less in the training materials, and she also noted different kinds of errors across the accents. Unsurprisingly, Harrington (2023) found that the audio quality affected the performance of the systems. She found that *Amazon Transcribe* had the lowest error rates regardless of whether it was focussed on the studio condition or the speech-shaped noise condition, while *Rev AI* was the most variable. Similar to findings from Loakes (2022), she found that even the best performing system did not accurately transcribe all of the material. Harrington (2023) showed a 13.9% word error rate (WER) for *Amazon Transcribe* with Standard Southern British English in the studio condition and 15.4% WER in the speech-shaped noise condition. The worst performance was for *Rev AI*, which had an error rate of 42.5% with the degraded speech for the West Yorkshire accent.

Another recent study by Harrington and Hughes (2023) looked at the variability of the ASR system. Using *Amazon Transcribe*, which was the best performing system in the study by Harrington (2023), the aim was to look at variability in performance with a homogenous group of speakers, and whether the errors observed correlated with particular phonetic properties. Using the DyViS database (Nolan et al., 2009) and focussing on 'homogenous' speakers with the same accent, Harrington and Hughes (2023) observed that for 99 speakers, WERs ranged between 11.2 and 33%, with a mean of 20% errors across the entire sample. They analysed various phonetic properties which included F0, formants, articulation rate, and voice quality to determine which features predicted performance and found that only articulation rate predicted WER. Taking all results together, Harrington and Hughes (2023) discuss how phonetic reasons for performance issues (even in clear speech) are not clearly predictable, and identifying causes of variability is also problematic. Harrington and Hughes (2023, 3134) note that the number of errors they observe in their homogenous sample of clear speech recordings (with 11.2% WER being the best performance) is 'worrying given the favourable [speech and recording] conditions… and raises issues about the general utility of ASR for many applications'.

The findings of Loakes (2022) and Harrington (2023) in particular are entirely consistent with known issues in automatic speaker recognition when degraded audio is used. For example, in a review paper about trends and developments in ASR, O'Shaughnessy (2023, 2) describes how sponsored challenges address the matter of 'noisy, far-field multi-speaker conversations' being difficult for systems, having up to 50% word error rates for automatic systems; other studies have shown approximately 15% word error rate for automatic systems in which humans can understand speech well. However, as has been shown in this section, even clear speech recordings (Harrington and Hughes, 2023) can have relatively high error rates without a predictable cause.

Analysis of the performance of ASR also brings into question how well human transcribers perform in forensic-like transcription tasks, and while this is not the focus of the study, it is important to address how humans perform in comparison. As mentioned earlier, Harrington (2023) analysed the output of 12 ASR systems, and this same audio was transcribed by professionally trained human transcribers in the study by Love and Wright (2021). While neither the humans nor the systems were able to provide accurate transcriptions of the entire recording, Harrington (2023) concluded that 'at present, it is more effective for humans to transcribe indistinct audio 'from scratch' as opposed attempting to manage and interpret the output of such systems'.

There is a similar finding to this in a recent experiment (Fraser et al., 2023), in which our team focussed on transcription performance from human transcribers who were presented with a section of audio from the same recording as the one used in this study (as well as in Loakes, 2022). Fraser et al. (2023) focussed on how well transcribers performed and saw that overall accuracy was relatively low, but still the top 11 transcribers (of a total of 40) were able to accurately transcribe between 50 and 62% of the material.

The new generation of automatic speech recognition systems needs to be tested because they are iterative and predictive and have access to masses of data compared to systems available only a year ago (e.g., Kallens et al., 2023). Any discussion of how automatic speech recognition performs with poor-quality forensic-like audio, therefore, needs to include these updated systems, because they have the potential to perform better than the older systems which do not draw on large language models, but their performance nevertheless needs to be analysed critically.

3 Aim

The aim of this study was to continue to update knowledge of ASR, and how it performs when applied to indistinct forensic-like audio. This research report is a direct update on a previous article (Loakes, 2022) which looked at forced alignment and smaller language model ASR systems and how they transcribed good-quality audio compared with poor-quality audio. It is also an update of some work by other teams which has looked at the matter of how naturalistic forensic-like audio is handled in modern ASR systems (e.g., Harrington et al. 2022). Given the rate of rapid technological advancement, it is imperative to test the new generation of automatic speech recognition systems, which have to date not been included in work on forensic transcription. In total, eight different systems using deep-learning and large language models are tested in this study—and taking into account updated versions and different file types there are 14 different ASR attempts on both the good- and poor-quality audio files.

The scope of this study is purposefully limited in experimentally providing an initial focus on how the newer generation of ASR systems performs on a sample of poor-quality audio, compared to a sample of good-quality audio. This means that only broad conclusions about the efficacy of automatic speech recognition in forensic contexts can be drawn, but this study nevertheless aims to contribute to the ongoing conversation about this rapidly advancing technology and how it is used and understood in forensics. The ensuing conclusions of this approach may indeed be obvious to linguists, but the goal of this study was to inform a broader audience about the issues.

4 Methods

To give more detail about the audio files used in this study (also used in Loakes, 2022), these are as follows:

4.1 Poor-quality audio

This is a 44.2-s stretch of audio from a recorded rehearsal by a singer and some musicians. This audio includes speech and instrument noise and is forensic-like in that there are varying background noises, there are multiple speakers who are at a distance from the microphone, and there is overlapping speech. This audio was recorded by one of the speakers via an iPhone and streamed to Facebook live, where it was retrieved with permission. The reference transcript has been verified with one of the speakers who organised and streamed this event, and the researcher's access to the accompanying video meant that the sample was clearer than the audio-only version (Loakes, 2022). The recording used has one female voice and three male voices, and all speakers are using Australian English. The speakers knew they were being recorded but were focussed on the task at hand and not attempting to be clear to the audience.

A transcript of this audio is provided below.

4.1.1 Poor-quality audio transcript

> *Yeah so just slowly building energy and nnnn and then I yeah*
> *What about what about another big drum fill will you let us know when you*
> *Yeah*
> *Alright*
> *Nah nah*
> *You gonna give us a hand signal or tell us what you do*
> *I I can't [laughter] ok*
> *From the from the top are we fine to go there*
> *Mel you don't need to do it so you know*
> *I mean this song I think is OK no it's relatively OK I I mean from the top of the set just marking it out what do you think yea nay care*
> *Sorry my brain just*
> *What song are we practising?*
> *Run through*
> *From the top*
> *yeah*

The good-quality audio file was also recorded on an iPhone, by the author. This file is shorter than the poor-quality audio, at 8.4-s duration. The speaker is an Irish English speaker, who knew she was being recorded and was specifically speaking into the microphone with the aim of being understood. The quality of the file is stable, with no background noise or overlap. The context of this recording is a short greeting, where the speaker introduces herself and also refers to a speech programme called *MAUS*, which was used in Loakes (2022) but is not used in the current study. The aim of this research was to deliberately stretch the systems, to determine whether ASR performance is better using systems with large language models.

4.1.2 Good-quality audio transcript

> *Hello*
> *my name's Chloé*
> *I live in Melbourne*
> *I'm from Ireland*
> *I moved from Galway*
> *two and a half years ago*
> *and I love MAUS*

There are eight commercially available systems used in this report. Unless otherwise stated, the files inputted are .wav files. The systems are as follows:

Descript[2]—This is the system used in Loakes (2022), and the results from previous research are also reported here. In November 2022, Descript upgraded and began using large language models (see, e.g., Plumb, 2022), so a new Descript attempt is also made here with both .wav and .mp3 files.

Sonix[3]—This is an automated transcription service that is described on its website as 'fast, accurate, and affordable'. This was a system used by Harrington et al. (2022).

Google Cloud[4]—This is a suite of services using Google infrastructure, also including speech-to-text based on generative AI. This was another system used by Harrington et al. (2022).

Assembly AI[5]—This is a service for speech-to-text, described on the company's site as a system that 'makes up to 43% fewer errors on noisy data'. It is also described as being trained on over 1.1 million hours of data.

Deepgram[6]—This service is considered on the company website to be a 'world-class speech and domain-specific language model'.

Amazon Transcribe[7]—This is a speech-to-text transcription platform within Amazon Web Services. It is described as a platform for developers who want to add speech-to-text to their applications. It is often used for call centre and medical transcription. Amazon Transcribe is the best performing system in Harrington (2023), when compared with Microsoft Azure and Rev.

Microsoft Azure[8]—This is speech-to-text software that now operates within Microsoft Word 365. It uses a 'Universal Language Model' and also allows customisation. This method was also used by Harrington (2023).

Whisper[9]—This is run by the company Open AI. It is described as a system that 'approaches human level robustness and accuracy on English speech recognition' and 'has been trained on 680,000 hours of multilingual and multitask supervised data collected from the web'.

[2] https://www.descript.com/
[3] https://sonix.ai/
[4] https://cloud.google.com/speech-to-text
[5] https://www.assemblyai.com/
[6] https://deepgram.com/
[7] https://aws.amazon.com/transcribe/
[8] https://learn.microsoft.com/en-us/azure/ai-services/speech-service/index-speech-to-text
[9] https://openai.com/research/whisper

There are multiple versions of Whisper available, and this research used those with large language models. Aside from the 'Whisper AI March 2023' attempt, Whisper was run through a third-party app. Whisper was used with different audio files, so there is a 'Whisper June 2023 .wav' and 'Whisper June 2023 .mp3' version as well.

5 Results

Turning attention first to the good-quality audio file, Table 1 shows the system used, the number and % of words transcribed, the number of these words *correctly* recognised, the proportion of the entire attempt which was correct, and the WER (the % of errors compared to the total words spoken). This breakdown shows performance and where there are trouble spots in the outputs of the systems.

With the good-quality audio, the worst performances were from the older version of Descript reported in Loakes (2022) and from Google Cloud. Assembly AI and Descript (the August attempts) performed best with the good-quality recording. These systems correctly identified all of the material in the audio—the low predictability *MAUS* was transcribed *mouse*, which is not technically incorrect because the phonemes are exactly the same for both. Amazon Transcribe and Microsoft Azure also recognised *mouse*, while Sonix produced *my house* and Whisper (in all three attempts) produced *mouths*.

The other systems had some other minor errors, such as *two and a half* transcribed as *to ½* (Descript and Google Cloud), and one larger error in Deepgram's output with the place name *Norway* used instead of *Galway*. Additionally, some systems (the later Whisper attempts, and Microsoft Azure) also used *name's* instead of *names* as uttered by the speaker, which may be an attempt at producing a more readable transcript, but technically introduces an error. Google Cloud had the greatest number of errors overall—also transcribing *mouse* as *maths* and missing the word *I'm* entirely.

To sum up how the systems responded to the good-quality audio, we can see that this audio is largely recognised by these systems, retaining the sense of what the speaker was saying in almost every case. While there is not full accuracy in recognition for most systems, despite the clear quality of the file, these transcriptions can still be classified as being useable overall, and in some cases error-free, or almost error-free.

Turning now to the poor-quality audio, the results in Table 2 show a clear reduction in the number of words transcribed by the systems, as well as a reduction in their accuracy.

Comparing Tables 1, 2, it is clear that less of the material is attempted, and less is correct, for the poor-quality audio. Better performance of a system is indicated by the results in the final two columns—both the proportion of the attempt correct and the WER.

Where we see '100% accuracy' for two of the systems in the second last column, it is important to remember that this is showing *% of attempts correct.* For example, Descript, before the large language model upgrade, only recognised three words in total, and these words happened to be correct, but this is by no means a good performance as can be seen by the error rate of 97.4%. Arguably though, this poor performance could be seen as useful forensically, because the lack of content eradicates the issue of whether the material is correct or not (also see Harrington et al., 2022).

Later attempts using both a .wav and .mp3 file had marked improvement as should be expected, with 57 and 59 of the total 116 words transcribed. While around half of these attempts were correct (52.5% with the .wav file, and 49.1% with the .mp3 file), the total word error rates were still very high, at 73.3 and 75.9%, respectively.

Whisper (March 2023), on the other hand, recognised 21 words, but this is only 18.1% of the total number of words used in the audio (meaning around 82% of the audio is not transcribed). This version of Whisper may be considered to perform relatively well in the sense that of the 21 words recognised, there are no errors, as shown below:

> *Yeah, so just slowly building energy. And then I…Yeah?*
> *It's relatively okay.*
> *I'm just marking it out.*
> *What do you think?*
> *Okay?*

However, this is far from ideal because a closer look at the performance of the system shows that the material comes from different parts of the audio and only from the female speaker, with

TABLE 1 Results for good-quality audio.

System	No. (and %) of words recognised (n = 25)	No. of words correct	% of attempts correct	% errors (WER)
Descript (Loakes, 2022)	24 (96%)	19	76%	14%
Descript (August 2023) *.wav* and *.mp3*	25 (100%)	25	100%	0%
Sonix	26 (104%)[a]	25/26	96%	4%
Amazon Transcribe	26 (104%)	25/26	96%	4%
Microsoft Azure	25 (100%)	24	96%	4%
Google Cloud	24 (96%)	18	75%	18%
Assembly AI	25 (100%)	25	100%	0%
Deepgram	25 (100%)	22	88%	12%
Whisper AI (March 2023)	25 (100%)	24	96%	4%
Whisper AI (June 2023) *.wav* and *.mp3*	25 (100%)	24	96%	4%
Whisper AI (Aug 2023) *.wav* and *.mp3*	26 (104%)	24	92%	8%

[a] In some cases such as this, an additional word *name is* instead of *name's* was recognised, so 26 words (of the original 25) have been counted.

TABLE 2 **Results for poor-quality audio.**

System	No. (and %) of words recognised ($n = 116$)	No. of words correct	% of attempts correct	% errors (WER)
Descript (Loakes, 2022)	3 (2.5%)	3	100%	97.4%
Descript (August 2023) *.wav*	59 (50.8%)	31	52.5%	73.3%
Descript (August 2023) *.mp3*	57 (49.1%)	28	49.1%	75.9%
Sonix	53 (45.7%)	20	37.7%	82.8%
Amazon Transcribe	29 (25%)	11	37.9%	90.6%
Microsoft Azure	33 (28%)	17	51.5%	85.4%
Google Cloud	0 (no attempt)	0	(no attempt)	(no attempt)
Assembly AI	32 (27.6%)	21	65.60%	81.9%
Deepgram	29 (25%)	14	48.30%	88%
Whisper AI (March 2023)	21 (18.1%)	21	100%	81.9%
Whisper AI (June 2023) *.wav*	80 (68.9%)	58	72.5%	50%
Whisper AI (June 2023) *.mp3*	82 (70.69%)	49	59.7%	57.6%
Whisper AI (Aug 2023) *.wav*	96 (82.7%)	55	57.3%	52.6%
Whisper AI (Aug 2023) *.mp3*	97 (83.6%)	57	58.7%	50.9%

large amounts of material from her speech (and all of the speech from male speakers) ignored. The WER for this Whisper attempt was 82%.

A number of the other systems do not perform well at all with this audio, for example, Google Cloud made no attempt, with an error message stating 'we could not process your audio with this model', which was presumably because of the audio quality given that the good-quality audio worked with this system. Of the attempts made by the systems, Sonix recognised the most words (53/116) but also made the most errors (37.7% accuracy). This system also performed poorly in the study by Harrington et al. (2022). Looking more closely at the output from Sonix in this study, some totally incorrect phrases are used in the output. For example, in the poor-quality audio, this section of speech:

> *You gonna give us a hand signal or tell us what you do*
> *I I can't [laughter] ok*
> *From the from the top are we fine to go there*

Is transcribed as:

> *How are you gonna go through this with the High Court, huh?*
> *OK*
> *from the from the top are we fine to go that.*

Here, there are some sections that are relatively accurate and some that have no resemblance to the original. It is worth noting that when processing this file using Sonix, the system came back with an error message warning about the 'low accuracy potential' due to the nature of the audio, so the poor performance is not unexpected.

Amazon Transcribe, which performed well in Harrington (2023) and was then used for a more in-depth analysis in Harrington and Hughes (2023), performed poorly for this data. The values reported above are actually from the United States-English model because the Australian English language model transcribed only *Wait. What* of the entire 116 words. For Amazon Transcribe, neither using the American English model, nor the Australian English model, have given a good outcome. Microsoft Azure also performed relatively poorly with this audio. To give some further examples of the performance of these systems, this is the entire output for Amazon Transcribe (using United States-English):

> *Yes, I just got all your building in. Yeah.*
> *Signal or, uh, OK.*
> *To talk we find because there's nothing.*
> *It's relatively unpayable said just marking it up, okay?*

Some words and phrases are recognisable from the ground truth transcript, but even phrases that are almost correct are still wrong in some way. For example *it's relatively ok no?* is transcribed as *it's relatively unpayable* and *just marking it out* is transcribed as *just marking it up.*

The best performance of all systems tested was Whisper (the June 2023 .wav file attempt) in which almost 69% of the 116 words were transcribed—of that attempt 72.5% was correct. However, 'good performance' is relative; the overall WER is still 50%. This attempt also correctly recognised some speech produced by the male speakers, unlike the March 2023 version. Interestingly, the .mp3 file of the exact same audio had a similar rate of words transcribed in the June 2023 attempt, but this version had more errors, with an error rate of 57.6%. The later August 2023 Whisper attempts, with both the .mp3 and .wav file, had the most words transcribed of all systems used but had slightly higher error rates than the June attempts.

It is also worth noting that the sections of transcripts correctly transcribed were different across all of the Whisper attempts, sometimes completely different, and sometimes just slightly different. For example, the (arguably) low predictability phrase *What about what about another big drum fill will* was correctly transcribed, minus the repetition, in the August 2023 .wav attempt as *What about another big drum fill?* The later August 2023 .mp3 attempt transcribed this as *What about not being comfortable with my weight?* The June 2023 attempts both produced *What about now being comfortable?* As shown above, the March 2023 attempt did not recognise any of the content from this phrase. Another example is the phrase *This song I think is okay, no? it's relatively okay* was transcribed correctly in the August 2023 .wav attempt, and almost the same transcription was produced using the .mp3 file except the word *no* was transcribed as *now.*

When there are incorrect transcriptions, there are also some similarities across how systems dealt with this; for example, the phrase *yea nay care* (which is asked with questioning intonation for each word) is transcribed by Descript in both August attempts and the June Whisper .wav attempt as *Gay, no? Gay?*, by the August 2023 .mp3 Whisper attempt as *Gay enough? Gay?*, and by the August 2023 .wav Whisper attempt as *Okay? Okay.*

6 Discussion

The aim of this research was to determine how well automatic speech recognition works on indistinct forensic-like audio with the new generation of systems that have large language models. Here, we have seen that the good-quality audio file had 24 or 25 (of 25) words recognised (and in some cases one extra word) with error rates between 0 and 18%. As demonstrated, the new generation of ASR systems largely perform well with that audio, despite some errors. This is unsurprising, as the older generation of ASR systems used in Loakes (2022) also had very good performance for this particular audio file, and as mentioned in that study each system is responding to a task it is designed to do.

For the poor-quality audio, the results were much more variable. The best performance was with Whisper (the June 2023 .wav file attempt) in which almost 69% of the 116 words were transcribed, and of that attempt 72.5% was correct. While this is a better performance than seen in the other systems and in Loakes (2022), this still leaves one quarter of the attempt either wrongly transcribed or missed by the system which is problematic for forensic contexts—equally problematic is the total word error rate of 50%. However, the better performance of this system compared to what was observed in Loakes (2022), and compared to other studies such as Harrington et al. (2022) and Harrington and Hughes (2023), needs to be acknowledged—this speaks to the fact that the audio used in the training of large language models is so diverse and so the systems can indeed respond better to new types of data (e.g., Kallens et al., 2023). Comparing back to the literature discussed earlier, the WER of 50% obtained for Whisper is exactly the same error rate mentioned earlier for sponsored competitions for ASR on multi-party speech in noise (e.g., O'Shaughnessy, 2023), so at this point, Whisper appears to be performing as well as any other system currently reported for this kind of audio recording.

The least accurate performance in this research is technically Descript (reported in Loakes, 2022) in which only three words were recognised by the system. While those words were correctly transcribed, there was no usable transcript. Later Descript attempts using large language models had a superior performance in comparison but still had error rates of approximately 75% for both .wav and .mp3.

Another result that should be noted is the earlier Whisper attempt from March 2023, where only 21 of 116 words were recognised, and these were all correct. While that appears to be a cautious response in the sense that if words could not be recognised no attempt was made to transcribe them, 18% accuracy is not a usable output.

The Google Cloud system had poor performance overall for these data overall, not recognising any of the poor-quality audio and having only 75% accuracy for the good-quality file. As seen in Table 2, Amazon Transcribe, Sonix, and Deepgram also had relatively low levels of recognition for the poor-quality file. Assembly AI, touted as performing well on noisy data, performed as well as a number of other systems using this data.

While Whisper in particular, using a .wav file, worked well *compared to the other systems tested*, in terms of correct transcriptions for low predictability items, its performance was not accurate enough to use for forensic transcription. Additionally, problems such as a correct transcription of a phrase in the June attempts being wrong in later attempts are a cause for concern in situations where the need for accuracy is so important. Finally, comparing the .wav files, the error rate increased across the June and August attempts of Whisper but *decreased* slightly when a .mp3 file was used.

Before concluding, this difference in transcription output when an .mp3 file is used compared to a .wav file is worthy of note. In the poor-quality condition, this study showed differences in the transcriptions depending on which file type was used, but there were no differences in performance for the good-quality audio files. With both Descript and the June 2023 Whisper attempts, the .wav files resulted in more accuracy (lower WER)—the difference was small for Descript, but for the June 2023 Whisper attempts there was a 7.6% difference in WER. However, in the August 2023 Whisper attempts, the .mp3 file had a slightly better performance than the .wav file. This variability is likely due to the fact that mp3 audio is compressed; one study looking at the effect of this compression on automatic speech recognition has shown .mp3 files can reduce transcription errors in some types of noise and induce transcription errors in other types—and that the effects are not consistent (Andronic et al., 2020). This result is simply one to be mindful of when working with ASR systems, and this highlights a topic worthy of further study so that better predictions can be made about how automatic speech recognition systems respond to the various types of audio that users may feed in.

7 Conclusion

This study compared the performance of a number of ASR systems, looking at how well they transcribed spoken language from a good-quality recording and a poor-quality recording. Taking into account the study as a whole, Open AI's Whisper performed far better than the other systems, having the lowest error rates overall. The study also showed that different versions of the same system (used at differing time points) do not always have equivalent outputs, and the later Whisper attempts are not necessarily the best attempts.

While Whisper performed best amongst the systems tested, it also needs to be remembered that forensic transcription is a task that is necessarily done without any ground truth to compare against. The potential for such a large error rate (50% WER at best) is not appropriate for forensic contexts; a transcription in which only 50% is correct is not useable. While the results of this study, for Whisper in particular, are a marked improvement in performance compared to the systems trialled on the same audio in Loakes (2022), this study advocates for the use of human transcription done in a measured and systematic manner (e.g., Fraser, 2022, also Loakes, 2022; Harrington, 2023) and for keeping ASR methods limited to tasks they are designed for. This aligns with the findings from Harrington (2023) discussed earlier, who observed that it is more efficient to do a transcription from scratch than to try and use the output of ASR systems which contain relatively high error rates.

Another important finding of this study was that .mp3 and .wav files can induce different outputs from ASR systems. With a good-quality recording, the ASR outputs were the same, while for the poor-quality recording, the results were variable. While the differences may not be large between them, it is nevertheless an important consideration when using ASR systems with noisy data. More generally, it is not apparent from the outset whether there are key similarities or differences across the ASR systems in terms of how they function and exactly which differences might predict variable performance. However, parameters can be adjusted in some systems (including Whisper), and the amount of material the systems are accessing is constantly changing, so at the very least we can predict variable performance, and be mindful of the inevitable variability in resulting outputs, even if it is not clear exactly what the variability will be linked to. Given this, it is likely that the variable performance demonstrated by the different versions of Whisper will happen almost every time one of these systems is used, even with the same audio. The lack of information and full transparency about the exact architecture of the systems, and the resulting lack of certainty about what causes differing levels of performance, is another reason that ASR systems are currently not useful or suitable for the forensic domain.

Finally, the fact that the data used in this study are forensic-like, and not from a real forensic case, does not *per se* limit its implications for forensics. The issues about recognition of particularly infrequent lexical items, background noise, speakers being at variable distances from the microphone, overlapping speech, and background noise still remain and (as noted by O'Shaughnessy, 2023) have hindering effects on speech recognition. In this study, however, these variables are conflated, and future work should focus on specifically controlling variables such as the degree of background noise. Arguably, it could be expected that Whisper, with its particularly large language model (not entirely trained on studio quality audio) and iterative processing, should be one of the best performing systems on the market, and we have seen that is indeed the case in this study.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the author upon request.

Ethics statement

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

DL: Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft.

Funding

The author declares financial support was received for the research, authorship, and/or publication of this article. The author received funding support from the School of Languages and Linguistics and the Faculty of Arts, at the University of Melbourne.

Acknowledgments

The author thanks Helen Fraser for assistance with ideas in this manuscript and discussion of issues surrounding the main themes within and Yuko Kinoshita for running the March 2023 Whisper attempt.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Andronic, I., Kürzinger, L., Chavez Rosas, E.R., Rigoll, G., and Seeber, B.U. (2020). "MP3 compression to diminish adversarial noise in end-to-end speech recognition" in *Speech and Computer: 22nd International Conference, SPECOM 2020*. St. Petersburg, Russia, October 7–9, 2020, Proceedings 22, 22–34; Springer International Publishing.

Bender, E. (2022). Resisting dehumanisation in the age of AI. Cognitive Science Society YouTube. Available at: https://www.youtube.com/watch?v=wuU-5rGPbyg (Accessed July 11, 2023).

Benzeghiba, M., Mori, R. D., Deroo, O., Dupon, S., Erbes, T., Jouvet, D., et al. (2007). Automatic speech recognition and speech variability: a review. *Speech Comm.* 49, 763–786. doi: 10.1016/j.specom.2007.02.006

Bridle, J. (2023). The stupidity of AI. The Guardian. 16 March. Available at: https://www.theguardian.com/technology/2023/mar/16/the-stupidity-of-ai-artificial-intelligence-dall-e-chatgpt (Accessed July 11, 2023).

Fetzer, J. H. (1990). "What is artificial intelligence?" in *Artificial Intelligence: Its Scope and Limits. Studies in Cognitive Systems, Vol. 4* (Dordrecht: Springer), 3–27.

Fraser, H. (2022). A framework for deciding how to create and evaluate transcripts for forensic and other purposes. *Front. Commun.* 7:898410. doi: 10.3389/fcomm.2022.898410

Fraser, H., Loakes, D., Knoch, U., and Harrington, L. (2023). "Towards accountable evidence-based methods for producing reliable transcripts of indistinct forensic audio" in *Presentation at the 31st IAFPA Conference*. Zurich, Switzerland, p. 21.

Harrington, L. (2023). Incorporating automatic speech recognition methods into the transcription of police-suspect interviews: factors affecting automatic performance. *Front. Commun.* 8:1165233. doi: 10.3389/fcomm.2023.1165233

Harrington, L., and Hughes, V. (2023). "Automatic speech recognition: system variability within a sociolinguistically homogeneous group of speakers" in *Proceedings of the 20th International Congress of Phonetic Sciences Guarant International*. (eds.) R Skarnitzl & J Volín, Paper ID: 593; 3131–3135.

Harrington, L., Love, R., and Wright, D. (2022). "Analysing the performance of automated transcription tools for indistinct audio recordings" in *Poster presented at the 2022 Conference of the International Association for Forensic Phonetics and Acoustics*, Prague, Czech Republic. Available at: https://robbielovelinguist.files.wordpress.com/2022/07/harrington-et-al._iafpa.pdf

Kallens, P., Kristensen-McLachlan, R., and Christiansen, M. (2023). Large language models demonstrate the potential of statistical learning in language. *Cogn. Sci.* 47:e13256. doi: 10.1111/cogs.13256

Koenecke, A., Nam, A., and Lake, E. (2020). Racial disparities in automated speech recognition. *PNAS.* 17, 7684–7689. doi: 10.1073/pnas.1915768117

Loakes, D. (2022). Does automatic speech recognition (ASR) have a role in the transcription of indistinct covert recordings for forensic purposes? *Front. Commun.* 7:803452. doi: 10.3389/fcomm.2022.803452

Love, R., and Wright, D. (2021). Specifying challenges in transcribing covert recordings: implications for forensic transcription. *Front. Commun.* 6:797448. doi: 10.3389/fcomm.2021.797448

Markl, N. (2022). "Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition" in *Proceedings of 2022 5th ACM Conference on Fairness, Accountability, and Transparency (FAccT 2022)*. ACM Association for Computing Machinery Seoul. 521–534.

Markl, N., and McNulty, S.J., (2022). Language technology practitioners as language managers: arbitrating data bias and predictive bias in ASR. arXiv [Preprint]. doi: 10.48550/arXiv.2202.12603

McCarthy, J. (2007). What is Artificial Intelligence? Available online at: http://www-formal.stanford.edu/jmc/whatisai/whatisai.html (Accessed January 23, 2024).

Morozov, E. (2013). *The Folly of Technological Solutionism*. New York: Public Affairs.

Nolan, F., McDougall, K., de Jong, G., and Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *Int. J. Speech Lang. Law.* 16, 31–57. doi: 10.1558/ijsll.v16i1.31

O'Shaughnessy, D. (2023). Trends and developments in automatic speech recognition research. *Comput. Speech Lang.* 83, 101538–101522. doi: 10.1016/j.csl.2023.101538

Perrigo, B. (2023). OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic time magazine. Available at: https://time.com/6247678/openai-chatgpt-kenya-workers/ (Accessed July 25, 2023).

Plumb, T. (2022). How descript's generative AI makes video editing as easy as updating text Venturebeat.com. Available at: https://venturebeat.com/ai/how-descripts-generative-ai-makes-video-editing-as-easy-as-updating-text/ (Accessed August 2, 2023).

Preston, L. (2022). Becoming a chatbot: my life as a real estate AI's human backup. The Guardian. Available at: https://www.theguardian.com/technology/2022/dec/13/becoming-a-chatbot-my-life-as-a-real-estate-ais-human-backup (Accessed July 11, 2023).

Rodriguez, J. (2022). OpenAI's new super model: Whisper achieves human level performance in speech recognition medium. September 27. Available at: https://medium.com/@jrodthoughts/openais-new-super-model-whisper-achieves-human-level-performance-in-speech-recognition-78289d48f9a1 (Accessed July 11, 2023).

Wassink, A. B., Gansen, C., and Bartholomew, I. (2022). Uneven success: automatic speech recognition and ethnicity-related dialects. *Speech Comm.* 140, 50–70. doi: 10.1016/j.specom.2022.03.009

Frontiers in Communication

Investigates the power of communication across culture and society

A cross-disciplinary journal that advances our understanding of the global communication revolution and its relevance across social, economic and cultural spheres.

Discover the latest Research Topics

See more →

frontiers

Frontiers in Communication

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact