Informatics in Education, 2025, Vol. 0, No. 0, 0–0
© 2025 *Vilnius University*
DOI: https://doi.org/10.15388/infedu.2025.09

Transcription in Computing Education Research: A Review and Recommendations

Lutz TERFLOTH[1], Vivien LOHMER[2], Friederike KERN[2], Carsten SCHULTE[1]

[1] *Paderborn University, Faculty of Computer Science, Germany*
[2] *Bielefeld University, Faculty of Linguistics and Literature, Germany*
e-mail: {lutz.terfloth, carsten.schulte}@uni-paderborn.de, {vivien.lohmer, friederike.kern}@uni-bielefeld.de

Received: February 2025

Abstract. Transcripts play a crucial role in qualitative research in computing education, with significant implications for the credibility and reproducibility of findings. However, unreflective and inconsistent transcription standards may unintentionally introduce biases, potentially undermining the validity of research outcomes and the collective progress of the field. In this article, we introduce transcription as a theoretically guided process rather than a mere preparatory step, illustrating its role using a case example. Additionally, through a systematic review of 107 qualitative research articles in computing education, we identify widespread shortcomings in the reporting and implementation of transcription practices, revealing a need for greater intentionality and transparency. To address these challenges, we propose a three-step framework for selecting, applying, and documenting transcription standards that align with the specific context and goals of a study. Rather than advocating for overly complex, one-size-fits-all transcription strategies, we emphasize the importance of a context-appropriate approach that is clearly communicated to foster trust and reproducibility. By advancing a more robust transcription culture, this work aims to support computing education researchers in adopting standards that enhance the quality and reliability of qualitative research in the field.

Key words: literature review, transcription systems, qualitative research, conversation analysis.

1. Introduction

Qualitative research is a crucial methodology regularly used by researchers in computing education (CER). Although not always the case, the empirical data are transcribed to be analyzed in subsequent steps. However, *transcripts are not neutral representations of spoken language* because transcribing is neither simple nor objective (Davidson, 2009); they can even lead to misinterpretation of the data (O'Connell and Kowal, 1995b). The phrase "I never said she stole my money", for example, has seven different interpretations, depending on which word is stressed (Rudzicz, 2016). In qualitative social sciences, transcribing oral data is not just a preparatory step but also already part of the analysis. For that, numerous transcription systems for different analytic purposes were developed (cf. overview of systems by O'Connell and Kowal (2009), a system for transcribing talk-in-interaction –

GAT2, by Selting *et al.* (2011)). The underlying motivation for the development of each of the respective systems differs, yet all strive towards the same overarching goal: improving the reliability and quality of the data analysis through the definition of transcription systems or standards. In contrast to such rigor, other fields consider transcription merely as the transformation of spoken into written data. Often, the choice of a transcription system is not critically evaluated in relation to the research questions, and detailed information about the chosen system is regularly missing from the methods sections of papers (Point and Baruch, 2023).

In CER, qualitative research is used, for example, to assess beliefs of computer science (CS) teachers (Bender *et al.*, 2016) or to analyze dialogues during pair-debugging (Murphy *et al.*, 2010). These transcripts are typically created before the analysis begins. In this instance, verbatim transcription remains the most prevalent approach, yet it is rarely specified with sufficient precision. Its typical implementation focuses mainly on semantic content and may overlook critical elements in computing education contexts. This shortcoming becomes especially pronounced in contexts where technology-mediated interactions (e.g., cursor movements in coding environments) or multimodal learning behaviors (e.g., non-verbal communication, such as gestures during pair programming) should be analyzed.

Ethnomethodological Conversation Analysis — an approach to the study of social interaction employed both in (interactional and anthropological) linguistics and sociology — has a long tradition of putting conversational interactions in writing. EMCA also has a long-standing tradition of discussing different transcription systems and adapting them to varying scientific research interests. As a research field, EMCA is primarily interested in the *how* of social interactions (Goodwin and Heritage, 1990), analyzing everyday interactions between people (Sidnell, 2012). For example, EMCA might explore how two participants initiate an explanatory interaction or *how* and when they use minimal feedback such as audible *'mhm'*s, hand gestures, head movements, or eye gaze for signaling or displaying understandings and misunderstandings to one another. For this kind of analysis, established transcription standards within the community are important because it is critical that transcriptions can be reproduced and that transcript-based interpretations can be understood by everyone in the community.

Drawing on ethnomethodological conversation analysis (EMCA) theory, we emphasize that transcription constitutes an inherently interpretive process that actively constructs analytical possibilities. Within CER, the choice of transcription system needs to be reflected and discussed. We experienced firsthand how habitually choosing a verbatim transcription system without reflection involves the risk of negatively impacting results. We argue that a shared understanding of transcription theory is beneficial for the community. Moreover, the development of standardized transcription strategies to improve qualitative research in our field is desirable.

This paper addresses that gap. Its main contributions are (1) a theoretical introduction to EMCA transcription theory; (2) a case example from a recent study to outline how the change of a transcription system improved the accuracy of the analysis; (3) a review of current transcription practices in our community; and (4) a practical three-step guideline for choosing, using, and communicating the transcription system in one's own research. We include a literature review of qualitative research articles from our community to empirically support our argument that CER needs better transcription standards. Readers may choose, depending on their interest, to read the theoretical foundations or to skip to the literature review, depending on their interests. This paper follows a recent trend of publications that review community practices (Sanders *et al.*, 2023; Oleson *et al.*, 2022). Importantly, we do not advocate for always choosing the most complex transcription system. Instead, we recommend a context-sensitive, reflective selection process and transparent communication of transcription choices within CER articles. The proposed three-step framework offers a structured, practical guide to support researchers in our field.

2. Background: A Theory of Transcription

In qualitative research, transcripts and transcribing are crucial, as it allows researchers to analyze and interpret the content of spoken interactions efficiently. However, oral conversations contain richer information, and much information is usually lost when a conversation is transferred into a written form. Generally, transcripts reduce the available data (such as video footage) in terms of complexity and are thus not to be understood as neutral representations of the original material. Instead, "the process of transcription generates the data upon which the analysis is built." (Ayaß, 2015, p.510) Transcribing is not just transforming (video-)recorded data into written text. In verbatim transcripts, for example, information about pauses, rhythm, intonation, dialect, or about gaze, gestures, or posture (for video data) is omitted. Therefore, as such information might be useful later during the analysis and interpretation of the data, the choice of whether to include it should be (a) carefully considered and (b) transparently documented.

To be able to preserve such information in written form, great effort has been put into the development of different transcription systems, especially since Ochs (1979) introduced the idea of *transcription as theory*. Some systems define a notation system for gestures, eye movements, and prosodic information (Davidson, 2009). More complex systems also include video frames. All efforts share the same goal: to systematically include paraverbal (e.g., pitch, volume, intonation) and nonverbal (e.g., gestures, gaze) elements in a standardized manner. In the 1990s, interest in these systems began to arise among the EMCA research community (O'Connell and Kowal, 2009, p. 241). Today, various standardized transcription systems are in use (Romero *et al.*, 2002, p. 620).

In the CER community, qualitative research methodologies are enjoying growing popularity (Fitzgerald *et al.*, 2011), referencing and reflecting on the choice of transcription system has yet to become standard practice (cf. Section 4). In the next section, we will therefore first elaborate on the theoretical background of the transcription of oral speech. To discuss why creating transcripts is always a *first step of data analysis*, the difference between spoken and written language is discussed. Subsequently, different forms of representing oral speech in written transcripts (sentences, utterances, intonation phrases) and

Word	Definition
GAT2	An abbreviation for 'Gesprächsanalytisches Transkriptionssystem' (translated: discourse and conversation-analytic transcription system); a system for transcribing speech in conversational interaction developed by (Selting *et al.*, 2011).
Intonation	The variation in pitch that conveys information beyond the literal meaning of words, including aspects such as sentence type (e.g., statement, question) and emotional expression.
Intonation Phrase	A segment of speech characterized by a single intonation contour, typically corresponding to a clause or phrase.
Notation	The symbols usable in a transcript to represent aspects of speech, for example, timing, pitch, and intonation.
Prosodic Information	An umbrella term used to describe information that is contained in pitch changes, intonation changes, pauses, or rhythm.
Pitch	The perceived frequency of a sound, often associated with the highness or lowness of a tone. In speech, pitch variations convey prosodic information such as stress, intonation, and emotion.
Transcription System	A set of conventions and notation signs used to represent spoken language in written form, which includes conventions for how, for example, pauses or simultaneous speech are represented in the transcript. Typically, transcription systems also include notations for phonetic, orthographic, and prosodic elements.
Paraverbal Elements	Vocal characteristics separate from the actual words, including tone of voice, volume, speed, and rhythm. These elements can convey additional meaning, emotion, or intent besides the literal content of the words (Mandal, 2014).
Nonverbal Elements	Communication cues (excluding words), such as facial expressions, body language, gestures, eye contact, and physical distance. These visual and tactile elements provide context and can significantly alter the interpretation of verbal communication (Frank *et al.*, 2015).

Table 1
An overview of commonly used terms in linguistics and EMCA. These are used throughout the paper and provided here for a quick reference.

their implications for data analysis are explained. Finally, we provide a practical overview of three different transcription systems that may be useful for various research interests in the CER field. Table 1 provides an overview of important terms used throughout the section to support readers unfamiliar with the vocabulary of EMCA.

2.1. *Spoken and Written Language: A Perspective from Linguistics*

Even disciplines that deal with language in its various forms, such as (interactional) linguistics or linguistic anthropology, have long neglected the differences between spoken and written language, even though they have important implications for transcription theory and researchers using transcripts. The differences become more obvious when comparing spontaneously produced speech (e.g., during a conversation) versus speech produced by reading from a teleprompter. Imagine people speaking spontaneously. They rarely produce what would be considered a syntactically complete sentence in written language. In contrast, a person reading from a teleprompter will produce syntactically complex sentences, sometimes even with many subordinate clauses, when they are in the written template. But individuals reading out loud — at least when they are not professionally trained — sound rather artificial as prosodic features (e.g., rhythm, stress, pitch) differ considerably from speech produced spontaneously. In some instances, these differences may even lead to understanding problems when trying to follow the argumentation or content (Carlson, 2009).

The reason for this artificiality lies in the differences in how *spoken* and *written* language are produced and perceived. Spoken language is volatile, dialogic, and prototypically embedded in an environment of face-to-face situations. In such situations, all participants share time and (physical) space, which enables them to perceive each other through various sensory modalities (Stein, 2018; Roberts and Street, 2017). The dialogic nature of spoken language blurs categories of speaker and listener because they constantly change during an exchange; they become even more blurred when participants speak simultaneously, interrupt each other, or provide continuous verbal and non-verbal feedback. Additionally, spoken language features more variation, for example, across regions, social classes, or speech situations. Written language, on the other hand, is codified and rather standardized. It is also persistent and able to be archived, monologic, and relatively context-free. In written language, the modalities that are missing (intonation, stress, gesture) are compensated for by higher verbal explicitness. Differences between spoken and written language can also be found in the area of grammar and syntax, with a tendency for spoken language to feature less complex and often incomplete clause structures. There is plenty of empirical evidence that the grammatical structures of spoken language have their origin in its volatile and dialogic nature (Couper-Kuhlen and Selting, 2018; Auer, 1992).

In summary, creating transcripts requires more than simply converting spoken language into text. Instead, transcription systems are tools for preserving the volatile information within conversations and for capturing their transient nature. They provide guidelines for the transfer of audible (intonation, stress, rhythm, speech rate) and visible (gestures, facial expression, body posture) linguistic features into a written transcript. Depending on the research questions, preserving certain volatile information can be beneficial. For example, when analyzing classroom discourse or learners' retrospection during think-aloud techniques, often not only the content but especially the process of comprehension and the formation of ideas and beliefs are of interest. To support coders in being accurate, objective, and reliable in their interpretations, a suitable and well-considered choice of transcription system—one that preserves crucial information within the recording—is used.

In the next section, we will reflect on the segmentation of spoken language for the purpose of offering alternatives to the linguistic category *sentence*, which is basically a unit of written language. There are other units that are more useful when transcribing dialogic spoken language.

2.2. *Segmentation of Spoken Language in Transcripts*

As spoken language rarely contains grammatically and semantically complete sentences, transforming spoken language into sentence-based written language is a challenge that is seldom reflected upon outside of linguistics or EMCA. However, during transcription, the decision of when a sentence starts or stops is challenging, as the syntactic structures are systematically adapted to the needs and specifics of dialogic interactions (Auer, 1992). Indeed, research has compellingly shown that people segment speech differently from written language. Therefore, different approaches for segmenting spoken language in transcripts have been developed. Instead of using sentences, speakers naturally chunk their speech into so-called *intonation phrases* by grouping interpretable units together using mainly prosodic means (such as stressed syllables, intonation contour, and pitch) (Speer and Ito, 2009; Selting *et al.*, 2011). Du Bois *et al.* (1993) define an intonation phrase as "[r]oughly speaking, [...] a stretch of speech uttered under a single coherent intonation contour. It tends to be marked by cues such as a pause and a shift upward in overall pitch level at its beginning, and a lengthening of its final syllable." Speakers use intonation phrases to express meaning, highlight and chunk information, and implement and signal discourse structure (Selting, 2000; Selting *et al.*, 2011). Additionally, intonation phrases often align with syntactic units within discourse, which are generally shorter than sentences. In doing so, they convey linguistic information such as focus, signals of completion or continuation, in addition to semantic information (Chafe, 1994; Selting *et al.*, 2011; Bergmann and Mertzlufft, 2009).

Understanding intonation phrases is therefore of significant importance in EMCA-grounded linguistics, given their pivotal role in conveying subtle layers of meaning, outlining speech structures, and enriching the overall communicative effectiveness of spoken language. For computing education studies, using intonation phrases for the segmentation of speech has certain implications and offers various potentials. Empirical evidence indicates that intonation phrases reflect cognitive processes to a certain degree. According to Chafe (1994), the amount of information within an intonation phrase is limited not only due to physical constraints (such as breathing) but also cognitive limitations. In a thorough study of intonation phrases and their relation to cognition, Chafe (1994) argued that human physiology and cognition are interrelated. Park (2002, p. 639) provided an in-depth overview of arguments for why intonation phrases may also serve as cognitive units. In her dissertation, Simpson (2016) tested this hypothesis and found empirical evidence supporting it. She also concluded that intonation phrases are ways to "break up the continuous speech stream into processable portions" and are therefore important not only for the *production*, but also the *comprehension and processing* of language by recipients (Simpson, 2016).

In summary, there is overwhelming empirical evidence that intonation phrases reflect the focus of attention of the speaker and are also important for the processing of information by listeners. These findings can be beneficial for computing education researchers, particularly if their research is interested in, for example, how an understanding of a digital artifact develops throughout an explanation. When researchers ask questions that go beyond *what* was said (content) and are instead interested in the *cognitive and interactive processing* during, for example, explanations, transcripts that segment speech into intonation phrases can improve the analysis. These units provide analysts with semantic chunks that align with cognitive processes and serve interactive functions, which can increase the accuracy of the analysis. In Section 6, we describe a method to easily identify prosodic units. Besides considering segmentation in transcripts, choosing the appropriate transcription system also requires attention.

2.3. *Transcription Systems*

Section 2 elaborates on how different transcription systems support different research interests. The systems vary mostly in their complexity, depending on how much of the additional information carried by spoken language they want to capture or emphasize (O'Connell and Kowal, 2009, cf.). The most important aspect for selecting a system is the level and types of details required for analysis. Clearly, there is no all-in-one solution when it comes to choosing a transcription system; more detail is not always better. However, to provide an overview, we introduce three transcription systems, ordered by increasing level of detail: (1) standard orthography (verbatim) or content-based semantic transcription using standard orthography, (2) GAT2 or Jefferson, and (3) multimodal transcription.

```
001  EX   For example, small, dark,
          hollow and round.
002  EE   Mhm.
003  EX   Exist only once. And then I
          start, for example, and give
          you any of these pieces. And
          you have to decide, where
          on the board you want to put
          it.
004  EE   Yes.
005  EX   And it always goes back and
          forth like this.
```

Fig. 1
In this transcript, Explainee (EE) and Explainer (EX) are engaged in an explanatory interaction. This transcript is a typical example of the verbatim transcription system that uses sentence structures for the segmentation of information.

Verbatim transcripts use standard orthography and follow a rather simple set of rules (see Figure 1 for an example); Kuckartz and Rädiker (2019, p. 42) provide a comprehensive overview of the most common rules. The goal of verbatim transcripts is to primarily preserve the content or semantic meaning of the spoken dialogue. In their simplest form, verbatim transcripts smooth dialects, remove laughs and affirmations, and include neither pauses nor parts of simultaneous speech. Punctuation marks are used to indicate the end of an idea or aspect. Each speaker's contribution is placed in a separate paragraph, preceded by an abbreviation indicating who spoke (e.g., I: for interviewer). Syntactical errors, discontinuations, or interruptions are smoothed over, resulting in a certain—and often unreflected—bias based on written language segmentation ('clauses' or 'sentences') and norms ('syntactic completeness'). The aim of this transcription system is easy readability and preservation of content-semantic aspects of the recorded interaction. It is the simplest form, easy to learn and easy to apply. The choice of verbatim transcripts is justified if the research is interested in the content or semantics of the empirical data. Therefore, verbatim transcripts are a suitable choice for methods of qualitative content analysis as described by, for example, Kuckartz (Kuckartz, 2014). However, "[i]n any case in which a speaker deviates from standard pronunciation, the transcription will clearly have a loss of information if that deviation cannot be represented in standard orthography" (O'Connell and Kowal, 2009), potentially hindering accurate interpretation of data.

```
001  EX   for exAMPLE,
002       SMALL dark (.) hollow and
          round;
003  EE   hm_hm
004  EX   exist only ONCE.
005       and THEN;
006       I start for example,
007       and give you ONE,
008       anyONE of these pieces,
009       and YOU have to decide,
010       WHERE on the board;
011       you [(.)      ] WANT to
          place it.
012  EE       [ye_es,      ]
013  EX   and it always GOES back and
          forth like this;
```

Fig. 2
A GAT2 transcript of the same segment as before. The inclusion of pauses, simultaneous speaker contributions and intonations creates a different impression of the content.

Therefore, if research interests go beyond *what* was said at the surface, and the goals of interpretation require more detail, different systems may be necessary (see Figure 2 for an example). Transcription systems such as **GAT2** (Selting *et al.*, 2011) or **Jefferson** (Jefferson, 2004) offer ways to include prosodic features such as intonation or stress[1]. They work based on the principle that a transcript can be extended by various levels of detail. Necessary categories of a transcript are the sequential structure of the interaction, pauses, breathing (when communicatively relevant), elision of words, interruptions, and prosodic features like pitch and stress. Instead of using sentences for chunking information, intonation phrases are used.

Each line in the transcript represents one intonation phrase (see Section 2.2). Everything in this transcription system is written in lowercase by default. Exceptions are the prosodically marked or stressed syllables, which are represented by capital letters. Punctuation is used to indicate pitch movement at the end of each intonation phrase (rising *(,)*, strong rising *(?)*, falling *(;)*, strong falling *(.)*, or level *(-)*). Pauses are marked according to their length (*(.)* for micro-pauses up to a length of .2 seconds, *(..)* for pauses up to a length of .3 seconds; pauses longer than .3 seconds are described in numbers).

The third, and most detailed, transcription system is a **multimodal transcription system** (see Figure 3 for an example). The transcripts include all notations mentioned for GAT2 but additionally include multiple layers of multimodal behavior, such as gestures, facial expressions, or body movements. The amount of additional information about multimodal behavior included depends on the research interest. Transcripts can be a combination of GAT2 or Jefferson's transcription rules and notations of multimodal behavior following the conventions introduced by Mondada (2018, 2019, 2011). Recently, it has become more common to insert timestamped video frames to avoid long verbal descriptions of multimodal behavior. Such systems are suitable for research focused on the fine details of multimodal behavior in interactions. However, the transcripts can be rather difficult to read because they contain a lot of information. Indeed, finding a compromise

[1] We will not distinguish between the two here. The Jefferson transcription system includes prosodic features (pitch, stress, intonation), though it segments speech differently. The system is widely used in the international EMCA community. Within the German community of interactional linguistics and EMCA, the GAT2 transcription system is more commonly used. GAT2 is especially interesting because it includes the segmentation of spoken language into intonation phrases. Our example uses GAT2 (see Table 2).

between readability and level of detail is an ongoing challenge when transcribing a video sequence.

In conclusion, regardless of the theoretical complexities, it is always sensible to reflect the value of including para- and nonverbal elements in transcripts. Certain paralinguistic elements are helpful for interpreting emotions, while others are rather ambiguous. For example, rapidly flowing speech is a marker of positive feelings, whereas stuttering is a sign of negative feelings. Laughed words, for example, are rather ambiguous: they may hint at a person feeling shame but can also simply be a sign of joy (for more interpretations, see Table 1 and Table 2 by Bloch (1996)).

Concerning prosody, "[p]rosodic elements such as intensity, pitch accent (i.e., the pattern of low and high tones used in a stressed word), and intonation, have been suggested to aid in conveying emotional affect (e.g., happiness) in acted speech." (Olsen, 2019). Pause duration and pause occurrence have been found to "consistently mark narrative section boundaries, thus suggesting that pause is a very important structuring device in oral narratives" (Oliveira, 2002). Therefore, these elements of information are often important for *truly* understanding the data beyond its content. It can already be helpful if STRESSED syllables are capitalized in verbatim transcripts to improve the quality of the analysis. Ultimately, transcripts are tools that support us as researchers and can thus be adapted to specific needs. To connect all these theoretical aspects with practice, we will elaborate on the practical experiences and issues we have faced in the following section.

```
001  EX      for exAMPLE,
002          SMALL dark (.) hollow and round;
003  EE      hm_hm
004  EX      exist only ONCE#.
005          and THEN;
006          I start for example,#
007          and give you #ONE,
008          anyONE of these pieces,
009          and YOU have to decide,
010          WHERE on the board;
011          you [(.)        ] WANT to pla#ce it.
012  EE          [ye_es,     ]
013  EX      |and that always GOES
     EX-ges  |„„„„„„„„„„„„„„„„„„„„„„|
             |prep-D                |
014  EX      |back| and |#forth;| like this |
                         #0:01:15.036
     EX-ges  |@EX |- - -|@EE    |„„„„„„„„„„„|
             |stroke-D          |retr-D      |
```

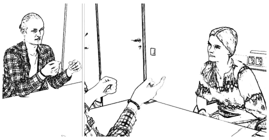


Fig. 3
A multimodal transcript of the same segment as before. Besides the content, stress and intonation, it also has some instances in which gestures are annotated. All points in the transcript that contain a # are moments at which a video frame was extracted to be used in the analysis. Here, only one video frame is provided to serve as an example of how this would look.

3. Case Example: Researching Naturally Occurring Explanations

Our example and experiences originate from a research project that focuses on analyzing how people engage in a naturally occurring explanation of a technical artifact (Terfloth *et al.*, 2023). The overarching research interest was to find explanation patterns and strategies useful for computing education, with the aim of being able to construct understandable synthetic explanations in eXplainable Artificial Intelligence (XAI). The experiments that elicited close-to-natural explanations in a dialogic setting (explainer and explainee) formed the foundation for analyzing how understanding is monitored and scaffolded by the explainer to ensure the explainee's understanding. The aim was to trace how human explainers dynamically and interactively address human learning when explaining a technological artifact.

Twenty explanatory interactions involving a technical artifact-the board game *Quarto!*[2]-were analyzed. Choosing a simple artifact was sensible to gain initial insights that provide a useful basis for further research into how more complex digital artifacts are explained in everyday settings. The explainers-already familiar with the game-were instructed to explain the game such that the other person would have a realistic chance to win if they played. The final video dataset includes recordings from 20 laboratory studies, in which 20 EX explained the game to 20 EE (19 male, 18 female, and 1 non-binary). Participants' ages ranged from 18 to 39 years (M = 24.92, SD = 4.42). Most (36) had an academic background, with 35 identifying as students across various disciplines such as engineering, education, economics, law, computer science, media studies, and linguistics, while two were full-time employees. Among the EX, seven had prior experience explaining the game. Gameplay experience among EX varied between 0 and 18 rounds (M = 5.46, SD = 5.18), and ten reported having general experience in providing explanations (e.g., from tutoring). Each study session was designed to last between two and three hours, including all pre- and post-assessments. The explanations lasted between 02:23 and 16:17 (mm:ss; M = 07:24, SD = 03:22). As is typical in CER, we transcribed the data using standard orthography in a smoothed, verbatim format, followed by coding the transcripts using content analysis. The transcripts included only semantic information and omitted stress or paralinguistic components (such as pauses and laughter).

The coding manual was based on a deductive code system derived from the dual nature theory of artifacts, according to which artifacts can be described either by their architecture and/or their relevance (Kroes, 1998; Schulte, 2008). In explanations, either aspect can be addressed, as artifacts are designed to be *means* to certain *ends*. Therefore, when coding, we identified at which points in the explanation each of the two sides was addressed by the participants (explainer and explainee). This allowed us to assess whether one side of the dual nature was addressed more frequently, and which of the two sides was addressed first. According to dual nature theory, for a holistic understanding of the artifact, both sides need to be addressed and understood.

[2] What follows is a rather compact description of the study and theoretical foundation. For more details regarding the theory, as well as the complete study, see Terfloth *et al.* (2023)

3.0.1. *Issues with Verbatim Transcripts*

Initially, we did not reflect on whether verbatim transcripts were the right choice for our research questions. However, our intended analysis was affected in such a major way that (a) a reliable analysis was not possible, (b) we therefore had to change the transcription system, and (c) we were subsequently motivated to assess our community's practices and reflect on the implications.

Throughout our studies, we tested different iterations of the coding manual in pilots. In these pilots, two independent coders (a student assistant and the first author of this paper) coded a set of 10 verbatim transcripts from the pilot studies. See the example in Figure 1 in section 2.3 for the verbatim transcript excerpt. However, the intercoder reliability was too low (k_{pre}=0.22, SD_{pre}=.17). Following typical conventions, we tried improving the coding manual to address the issue by including better examples, elaborating on corner cases more clearly, and using clearer wording. However, all measures taken were unsuccessful, as the intercoder reliability remained unsatisfactory.

To gain a more profound understanding of the origin of these issues, both coders compared all coded segments in an intercoder session and identified some root causes. The main cause leading to the highest number of disagreements was that coders defined the boundaries of a coded segment (i.e., architecture or relevance) differently. In the coding manual, the minimal coding unit was specified as "at least one word." A typical example of disagreement was an instance in which one coder decided to include a word or phrase in a coded segment that the other coder chose not to include. During the intercoder sessions, many of these disagreements could not be resolved, as different boundaries for a code were due to different plausible interpretations of specific parts of the transcript. Ultimately, even though large parts of these coded segments contained nearly identical semantic information, they could not be counted towards agreement due to a lack of overlap percentage. In summary, in most instances, consensus was not reached.[3]

The switch from verbatim to the GAT2 transcription system resolved this issue. See the example in Figure 2 in section 2.3 for the GAT2 transcript excerpt. It ultimately allowed us to follow the study's initial research interest instead of prematurely rejecting it. As mentioned above, GAT2 transcripts segment speech into intonation phrases, which are smaller units than sentences in writing. The coding manual was altered so that the minimal coding unit was defined as *at least one intonation phrase* (i.e., one line in the transcript). The same two independent coders coded the final set of 20 GAT2 transcripts using the refined manual. This improved the intercoder reliability significantly and especially reduced the standard deviations (from k_{pre}=0.22, SD_{pre}=.17 to k_{post}=0.76, SD_{post}=.03). After finishing coding, both coders reported that, due to clearer rules for identifying the boundaries of a code—enabled by intonation phrase segmentation—, boundary identification was now more precise and less ambiguous than before.

Premature rejection of our research interest due to an incorrect choice of, and lack of reflection on, the transcription system would have resulted in a type II error (beta error): we would have failed to detect an effect or a relationship when there actually was one.

[3] We coded using MaxQDA. The default setting in MaxQDA counts for agreement if 90% of the segment overlaps.

3.0.2. *Theory Guided Interpretation*

From a theoretical perspective, the improved interpretation based on GAT2 transcripts can be attributed to three main points: (1) inclusion of stress information, (2) segmentation into smaller, prosodic units, and (3) the link between intonation phrases and cognitive and interactive processes.

First, GAT2 includes information about stressed (capitalized) syllables within each intonation phrase, which helps to differentiate certain cases better. For example, the phrase "I never said she stole my money" has seven different interpretations, depending on word stress (Rudzicz, 2016). In GAT2 transcripts, stress can aid in pointing out which side of the dual nature is being addressed in certain moments of the explanations. For example, "and then you decide where to put the piece" would address the process and rules of the game (architecture). "And then you decide WHERE to put" would address that there are different positions that one can use to place the piece strategically (relevance).

Second, GAT2 segments information into intonation phrases based on the prosodic characteristics of speech, which improved annotation accuracy for us. Coding the verbatim transcripts reliably was problematic, as there were multiple options for coding single sentences. In some instances, using one, two, or even three codes was sensible. In the case of compound sentences or other complex utterance structures, the boundaries of ideas that receive one code were difficult to assess. This changed with GAT2 due to chunking information into intonation phrases. Because these are based on the prosodic information in the utterances, the boundaries of the semantic chunks were predetermined. In the coding manual, we defined the rule that one intonation phrase always receives one code. Coders thus only had to focus on coding the content, as determining the boundaries of a coded segment was no longer necessary. In comparison with the verbatim transcripts, GAT2's segmentation provided clearer boundaries for coding, making content coding more straightforward and less ambiguous.

Third, empirical evidence suggests that intonation phrases reflect cognitive processes (Chafe, 1994; Park, 2002; Simpson, 2016). During language acquisition, people learn to segment speech into these units by using stress, intonation, and pitch to convey meaning. As our cognitive capacity is limited both for the speaker and the listener; however, there are boundaries to how much one can think ahead while speaking, and there are also boundaries as to how much information a listener can process in one go. For example, organizing — on the fly — which aspect an explainer addresses during an explanation of a game is a challenging task. Thus, if the explainers had reformulated what was said or paused, they might have spotted an inaccuracy or ambiguity in parts of their explanation. Especially these instances often signaled a shift in focus regarding which side of the dual nature was addressed and therefore contained important information for us.

In conclusion, GAT2 transcripts drastically enhanced intercoder reliability in our coding process. The improvements can be attributed to (1) the inclusion of stress information helpful for interpretation, (2) the segmentation based on prosody, and (3) the potential link to cognition. These features facilitated more explicit interpretations, reduced ambiguity, and improved coding reliability, offering a promising approach for analyzing the shifts in explanations. Potentially, due to the intonation-based segmentation, we were able to retrace cognitive processes more accurately when coding GAT2 transcripts, especially in moments when speakers shifted from addressing one side of the dual nature to the other. Regarding other (computing) education research endeavors, this aspect of cognitive chunking could underpin future interpretations in interesting ways.

To assess the existing risk of other researchers running into similar issues, a systematic literature review of articles from our community assesses the community's transcription practices in the next section. Afterward, we discuss whether current practices may lead to similar risks and elaborate on how much detail transcripts need.

4. A Bigger Problem? A Systematic Literature Review

What we experienced may be symptomatic of our community. Phrases such as *sessions were recorded and transcribed*, *both researchers read the transcribed, audio-taped interviews*, or *the retrospective interviews were recorded with the participants' consent and transcribed in full* are epitomes of typical sentences found in the methods sections of papers in our field[4]. To solidify our argument empirically — that CER needs better transcription standards — this section contains a systematic literature review of our community's transcription standards. To this end, we conducted a review of 107 articles from our field. The literature review is guided by the question: to what extent transcription standards are reported in research articles in the CER field (RQ1)?

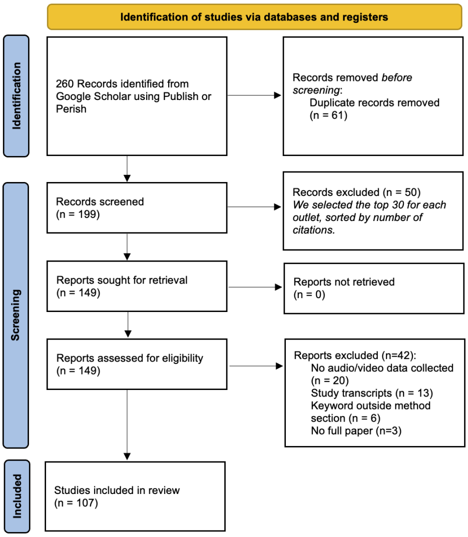

Identification of studies via databases and registers
Identification
Screening
Included
260 Records identified from Google Scholar using Publish or Perish
Records removed *before screening*: Duplicate records removed (n = 61)
Records screened (n = 199)
Records excluded (n = 50) *We selected the top 30 for each outlet, sorted by number of citations.*
Reports sought for retrieval (n = 149)
Reports not retrieved (n = 0)
Reports assessed for eligibility (n = 149)
Reports excluded (n=42): No audio/video data collected (n = 20) Study transcripts (n = 13) Keyword outside method section (n = 6) No full paper (n=3)
Studies included in review (n = 107)

Fig. 4. PRISMA 2020 flow diagram for systematic reviews provides insights into the selection process.

We acquired literature from five popular CER outlets: ICER, ITiCSE, TOCE, KOLI, and SIGCSE, using the keywords: transcript, transcribe, and transcription[5]. Using the software Publish or Perish, we combined the results for each of the three keywords and sorted them by the number of citations. From each of the five outlets, we selected — if possible — 30 unique articles. We screened the articles' abstracts to ensure that the articles

[4] These sentences are not direct quotes but paraphrases of popular phrases. We believe pointing fingers is not needed and is by no means helpful towards the more general point this paper tries to make.

[5] See `https://osf.io/8ejcu/?view_only=15343ef09ff24e3e98571994294dd484` for the article database and queries. It contains the final list of the reviewed articles. Additionally, it contains quotes of the parts where details were shared about the transcripts.

Exclusion Reasons	Total	Outlet				
		ICER	KOLI	SIGCSE	TOCE	ITiCSE
No data transcribed	20	4 (20%)	5 (25%)	3 (15%)	2 (10%)	6 (30%)
Mentioned "study transcripts", which are a different types of transcripts and irrelevant for this article	13	2 (15.4%)	2 (15.4%)	1 (7.7%)	4 (30.8%)	4 (30.8%)
The three keywords (transcribe, transcript, transcription) were *only* mentioned outside the method section of the papers, for example in the background sections. These papers were not working with empirical data in the form of transcripts.	6	0	0	2 (33.3%)	0	4 (66.7%)
Not a full paper (e.g., a poster or work-in-progress)	3	0	2 (66.7%)	0	1 (33.3%)	0
Total	**42**	**6 (14.3%)**	**9 (21.4%)**	**6 (14.3%)**	**7 (16.7%)**	**14 (33.3%)**

Table 2
Summary of exclusion reasons by source and their relative percentages.

included in the review were qualitative studies that used transcripts. See Section 4.1 for more details on the criteria. This resulted in a total of 107 articles from five popular CER outlets (see Figure 4).

4.1. *Inclusion & Exclusion Criteria*

We analyzed research papers from 2015 to 2023 and journal articles published in one of the five outlets and only included papers that were qualitative research *and* included at least one of the words *transcript*, *transcribe*, or *transcription* in the full text. Forty-two articles were excluded from the review for different reasons (see Table 2 for details). For example, 20 articles were excluded because they contained no transcribed data. Additionally, 13 papers mentioned "study transcripts," which referred to a different type of transcript irrelevant to this article. Six papers only mentioned the keywords outside the methods section, such as in background sections, and did not work with empirical data in the form of transcripts. Finally, three papers were excluded because they were not full articles, such as posters or works in progress. This systematic exclusion process ensured that only relevant studies meeting the inclusion criteria were analyzed.

4.1.1. *Threats and Limitations*

Regarding researcher bias, the system for categorizing the different articles by the amount of detail shared about transcript development, we would claim that the bias is mitigated. The category system is fairly simple, and the application of the categories was straightforward. Additionally, two researchers *independently* categorized the papers and subsequently discussed the categorization of the different papers to reduce the risk of human oversight. Both researchers closely read the sections of all articles that provided insights into the transcription strategy.

The representativeness of the sample could be questioned, especially concerning whether the sample can reflect the popular practices of the community. To mitigate this risk, we purposely chose the articles that were well received (i.e., articles were sorted by the number of citations). Additionally, we chose articles from popular venues within the ACM Computing Education community. Moreover, we chose articles from 2015 until June 2024, to reflect current rather than past practices, while leaving some room for potentially identifying improvements and changes throughout these nine years.

4.2. *Data Analysis and Categorization*

We reviewed each of the articles to assess the details shared about the transcription process. The review focused on whether the article provided details about how the transcription was conducted, including the methods, procedures, and any potential biases or limitations associated with the transcription process. We did **not** review whether the transcription system, if identified in the articles, was suitable for the research questions, as this was beyond the scope of our article. Each article was assigned to one of three categories based on the level of disclosure provided:

- Full Disclosure: Articles that provided comprehensive details about the transcription process, including the methods used, the procedures followed, and any potential issues or biases addressed. More precisely, the articles reflected on their choice of system, stated which system was used, and argued how that supported their research objectives or shared the details of the system in, for example, the appendix.
- Insufficient Disclosure: Articles that mentioned the transcription process but lacked comprehensive details, leaving some aspects of the transcription process untransparent. For example, they used "verbatim" or "word-per-word" to communicate the types of transcript. However, the choice of system is not reflected, and the details shared are insufficient to reproduce their transcripts. It is unclear what information is included and what information is excluded.
- No Disclosure: Articles that did not provide any information about the transcription process and lack transparency of their methods. For example, the articles stated that transcripts were developed without any further details, or even withheld information about the creation of transcripts, even though, for example, transcripts were coded in later steps.

As stated before, each article was reviewed independently by two researchers (the first author and a student assistant) to ensure reliability and consistency in the categorization process. Discrepancies in categorization were resolved through discussion and consensus. Table 3 lists the absolute and relative numbers of articles within each of the three categories (full disclosure, insufficient disclosure, no disclosure) for each venue and across all venues combined.[6]

Articles	**#**	**Degree of Disclosure**		
		Full Disclosure of Transcription Strategies (comprehensive, transparent, detailed, reflective)	**Insufficient Disclosure of Transcription Strategies (partial, vague, limited, incomplete)**	**No Disclosure of Transcription Strategies (opaque, absent, undisclosed, withheld)**
ICER	29	0 (0%)	13 (44.8%)	16 (55.2%)
KOLI	22	0 (0%)	4 (18.2%)	18 (81.8%)
SIGCSE*	5	0 (0%)	0 (0%)	5 (100%)
TOCE	27	0 (0%)	6 (22.2%)	21 (77.8%)
ITiCSE	24	0 (0%)	8 (33.3%)	16 (66.7%)
Total	**107**	**0 (0%)**	**31 (28.9%)**	**76 (71.0%)**

Table 3
Summary of articles reviewed and categorized according to the degree of disclosure.

** Publish or Perish just listed articles from 2017, even though we searched from 2015 and 2023. We tried to address the issue by changing the outlet name in the queries without success. Addressing the issue would have induced more complexity in our query logic. Thus, instead of further complicating the process, we used these 5 articles from 2017 instead.*

For a more profound understanding of each article's practices within the three categories, the wording used to describe the transcription process was assessed. However, out of 107 articles, no article provided information that, according to the theoretical discussion and our experiences, can be considered **sufficient**. Thus, the category **Full Disclosure** contained no articles.

The majority of articles in the category **No Disclosure** — at minimum — inform the reader about the existence of transcripts, for example, by stating that transcripts were coded. Some articles in this category fail to mention that transcripts were created, even though it was implicitly clear that transcription of recorded material was required. In all 76 articles, no details regarding the standards applied were shared. The articles did not clarify which information was preserved in the transcripts (e.g., stress, content, prosody, paraverbal or nonverbal components such as laughter).

The 31 articles in the category **Insufficient Disclosure** mentioned, for example, that data was *transcribed verbatim*. Some mentioned that a professional or automatic transcription service was used, which provided a certain transparency of the overall process.

[6] The full list of papers is available at https://osf.io/8ejcu/?view_only=15343ef09ff24e3e98571994294dd484

However, they did not provide sufficient details to support readers in understanding which standards were applied and which information was preserved, thereby hindering comprehension of interpretations and results. An example of this category is: "[t]he interviews were transcribed using automatic transcription software and were afterward manually corrected. The transcripts were anonymized and used for subsequent processing" (Aivaloglou and Meulen, 2021).

Out of 107 articles, 87 (81.3%) quoted excerpts from their transcripts. This can generally be considered good practice in qualitative research. However, as the articles did not share sufficient information about the standards to which the transcripts were created, it was unclear whether the quoted excerpts were altered for the publication. Consequently, even though the excerpts were quoted, it cannot be determined whether these excerpts were formatted or cleaned for readability, or if the format matched that which was coded in their analysis. Identification of standards would resolve this issue as well.

4.3. *Results*

In this section, we address the research question RQ1: To which extent are transcription standards reported in qualitative research articles in the CER field? The analysis of articles revealed a significant lack of transparency in the transcription process within qualitative CER articles that base their results on audio or video data (see Table 3 for an overview of the degree of disclosure of the transcription process within the articles).

In summary, out of the 107 articles reviewed:

- A substantial 76 articles (71.0%) disclosed almost no details about their transcription methods.
- 31 articles (28.9%) provided some but insufficient disclosure.
- No article (0%) provided full disclosure of their transcription processes.

These findings indicate a prevalent culture of insufficiently reporting transcription practices in qualitative CER. Consequently, the critical evaluation and reflection of the results within our community is severely impacted. In contrast to practices within EMCA, in our community neither the status of transcriptions in the research process nor their underlying principles of production are reflected, discussed, or made transparent to the reader. While a small percentage of articles share some details about the type of transcript that were used, no articles shared arguments for the choice of transcription system or standards. We therefore claim that the choice is *not actively reflected*, or at least these reflections are not transparently communicated, in articles in our field. Even though we would like to provide some reasons, the review results do not provide any insights into the underlying reasons for this trend.

5. Discussion

In this section, the issues arising from transcription strategies not being communicated in method sections are discussed. By doing so, this paper follows the path of publications in other fields that have also raised the need for better transcription standards (Point and Baruch, 2023). Our analysis highlights the potential to further develop transcription practices within our research community. In particular, fostering greater reflexivity in dealing with transcripts and establishing more standardized ways of reporting transcription strategies could enhance transparency and comparability in research findings. Before diving into specific discussion points, we want to address two issues: The fact that all the articles reviewed went through a rigorous peer-review process yet still do not provide sufficient details regarding transcription in their method sections indicates room for improvement. Second, the absence of clear communication regarding the standards applied in transcript creation makes it difficult to conduct meta-analyses of our community's research practices. For instance, systematically assessing whether the interpretations and results of qualitative studies in our field are built on solid foundations becomes a challenge. With our pragmatic transcription framework, we aim to provide a useful tool that facilitates a more structured evaluation of spoken data and enhances the overall transparency and reliability of research outcomes.

In CER, verbatim transcripts are the predominant choice. As long as the choice is (a) reflected upon and (b) transparently communicated in papers, this is not a problem per se. A suitable way to reflect is to ask: How much information does a transcript *need*? Firstly, simply adding details in transcripts for the sake of having detailed transcripts is not recommendable. "One of the important features of a transcript is that it should not have too much information. A transcript that is too detailed is difficult to follow and assess. A more useful transcript is a more selective one" (Ochs, 1979). The amount of detail a transcript needs, therefore, depends on the study's objective and thus requires researchers to reflect on which features of the transcript should be excluded and included. More precisely, aligning the transcript system with the study's objectives requires researchers to think about how the potential exclusion of information may "blind us to other features of language which are equally important to human communication" (Olson, 1993).

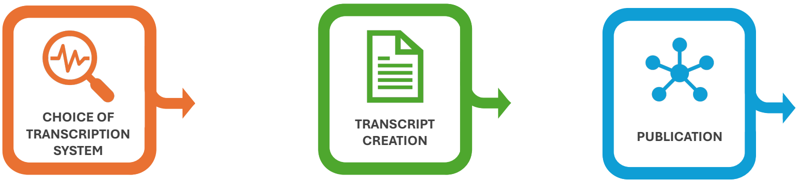
In educational contexts, these other features can be relevant for analysis. Collaborative learning is common, and understanding the dynamics of group interactions is, for some research interests, important. In retrospective techniques, retracing an inner debate of someone may not be possible if only verbatim transcripts that contain information about *what* was said are used. While there is definitely a place for research in our field concerned with *what* was said (e.g., identified themes in Fowler *et al.* (2021)), there is also considerable research in our community that is more interested in or additionally focused on what was *meant* (e.g., Alshahrani *et al.* (2018)). Two examples relevant to CER help illustrate this point: First, in the context of block-based programming, a transcription system that includes gestures such as pointing captures the interplay between verbal interactions and visual coding blocks in pair-programming settings. This is potentially relevant, for example, for understanding how students jointly navigate and interpret programming tasks. Second, incorporating prosodic and multimodal data can provide deeper insights when researching how students struggle during pair programming. Capturing pauses or hesitations in speech might reveal moments of cognitive struggle, which can point researchers to significant instances in the data. In these impromptu examples, smoothed, verbatim transcripts would exclude phenomena relevant for analysis; details relevant for interpretation are missing (cf. the seven interpretations of the phrase "I never said she stole my money" (Rudzicz, 2016)). Sometimes capturing such nuances might be directly related to research interests and can serve as a guide. Thus, if not reflected upon, verbatim transcripts may lead to unintentionally ignoring nuances, resulting in inaccurate or incomplete analyses and interpretations.

Concerning the understandability and transparency of qualitative research results, one important aspect needs reflection: In qualitative research, trustworthiness is increased through transcription. Transcripts provide evidence for the analysis (Duranti, 2006). Therefore, if the process of creating the evidence is not reflected upon and made transparent, analytical claims (Ashmore and Reed, 2005) may be based on a weak foundation. As this would generally require more text in publications, Lapadat (2000, p. 217) states, "[w]hen standardized procedures are used, a few words will suffice, but when researchers contextualize and negotiate a method as a means of interpretive seeing, there is no shortcut to explicit description."

Lastly, we want to discuss an important issue concerning insufficient reflection on the choice of transcription system: Our findings in the case example suggest that unreflected adherence to verbatim transcription could risk prematurely dismissing promising research ideas. This would have led us to overlook valuable insights central to our research proposal in our case example. Therefore, we want to highlight the risks of selecting verbatim transcripts without critical evaluation, particularly the inefficiencies and biases this practice introduces. An unreflected choice of transcription system can potentially cause researchers to dismiss seemingly unreliable coding manuals and doubt their research approach. This risk should motivate us as a community to reflect and adapt current transcription practices.

6. A Pragmatic Framework for Transcribing

In this section, we provide practical recommendations to support researchers in our community in improving their transcription practices. The SIGSOFT standards for qualitative research require authors to "identif[y] data recording methods (audio/visual), field notes, or transcription processes used" (Ralph *et al.*, 2021). For this, a practical yet theoretically sound guideline can be helpful. We condensed various advice found in the literature (e.g., (Point and Baruch, 2023; O'Connell and Kowal, 1995b; Davidson, 2009; O'Connell and Kowal, 1995a)) into a three-step process (see 5). In the first step, after data acquisition, researchers should—keeping their research questions in mind—choose what information needs to be included in their transcripts. During the development of the transcripts in step two, the rules of the selected system should be followed rigorously to create standardized and comparable transcripts. In the final step, when writing the publication, the choice of transcription system needs to be made transparent and discussed in the context of the research question. Only if the article contains sufficient information will readers be in a position to fully reproduce the research and create appropriate transcripts themselves. Furthermore, the argumentation provides insights for readers to evaluate how the research questions are planned to be answered. Although not part of the corpus of articles we reviewed, Tenenberg and Chinn (2019); Kong *et al.* (2022) provide good examples of how to



Prior

- What information is in the data (different for audio and video)?
- What information needs to be included in the transcript to serve the purpose of the research interest?
- Which segmentation (sentences, IPs) is most sensible for the research interest?
- Define and consciously decide which rules of a transcription system you want to use.
- Who codes and interprets the data, and do they know how to interpret the transcription system?

During

- Use consistent formatting and stick to the rules provided by the transcription system.
- Iteratively add information to the transcripts instead of trying to add all information in one go.
- Be aware of how symbols are used in the conventions (e.g., punctuation marks may be used differently in transcripts).
- There may be some kind of training effect. Revise the earliest transcripts to ensure consistency across all transcripts.
- If you adapt a system, carefully document the changes.

After

- In the publication, state what transcription system was used.
  - Why was it used? How does it relate to your research interests?
- Which segmentation was used and why?
- Were thirds parties or AI tools used to create the transcripts?
- How did the researcher get familiar with the data?
- How many pages of transcripts were created?
- Were transcripts verified or was the process overseen?
- If possible, share transcripts as part of the publication

Fig. 5. Flowchart illustrating the steps in the process, with a synthesis of important guiding questions for each step. Depending on the complexity of the qualitative study as well as individual interests, not all aspects need to be strictly adhered to and serve as a guideline supporting the reflection and decision-process.

address transcription practices and can serve as references. In the following subsections, we will discuss further considerations for each of the three steps.

6.1. *The Choice of System — Before Transcribing*

The process of transcribing transfers the volatile nature of spoken language in interaction into the readable and reproducible format of the transcript. To pick the right notation for the transcription, it is important to reflect on the research interest and research question to decide which details need to be preserved (Ochs, 1979). To aid in this decision, a few guiding questions can help select an adequate notation system (Spiegel, 2009, p. 7f):

- What is the transcript for? In which context and for what kind of phenomenon do I transcribe?
- What is the research interest?
- Which information must the transcript preserve so that it can serve this purpose?
- What do the readers of the transcripts (e.g., researchers, student assistants) know about transcripts and transcribing? Will they be able to work with them?

For the analysis, it is essential to remember that research assistants may not be familiar with the transcription system. If multiple people will work with the transcript for interpretation and analysis, they should be familiar with the rules and conventions of the chosen system. Especially with more detailed systems, training may be required.

6.2. *Creating the Transcripts — While Transcribing*

It can be helpful, especially when using more complex transcription systems for the first time, to iteratively improve the level of detail in the transcripts. Rather than aiming for perfect transcripts in one go, it is sensible to start with a simpler initial transcript and refine it through successive iterations. As more complex transcription systems require adherence to a larger set of rules and conventions, keeping all of them in mind simultaneously is often overwhelming and can lead to errors. Instead, one may begin with a simple verbatim transcription. In a subsequent step, pauses and stress are added according to the conventions. Then, intonation phrases may be identified, and the segmentation is changed accordingly, until a refined transcript is ultimately achieved. One final step should always be to revise the transcripts and listen to the source file while reading along with the transcript to spot errors.

Transcribing is a time-consuming task. Nowadays, however, there are helpful ways to increase transcription speed, even when the system is more complex than verbatim. Openly available, locally executable transcription models such as Whisper (Radford *et al.*, 2022) are becoming increasingly popular and effective. Many commercial products for qualitative analysis also provide automatic transcription services (e.g., MaxQDA, Atlas.ti, Transana). In the past, the quality of auto-transcripts was insufficient, and correcting errors often proved to be as time-consuming as transcribing manually. This, however, has changed. Today, models perform well and run locally, so they can be used with sensitive data (if the recorded audio data are of high quality). Even automatic speaker diarization[7] is supported. However, when using such tools, the question of how one familiarizes oneself with the data appropriately becomes increasingly important. Additionally, these tools tend to smooth the language, often more than standard verbatim transcription does. In other words, there is a trade-off between obtaining transcripts quickly, knowing the data well, and having accurate transcripts, all of which are helpful for improving the analysis. While these tools certainly make the creation of a first, rough transcript almost effortless, there are two aspects that we want to emphasize. First, one should use the time saved by auto-transcripts to better familiarize oneself with the data-perhaps even by extending the transcripts beyond their verbatim nature and including certain paralinguistic elements. Second, these tools can be problematic if researchers are unaware of the issues and potential consequences of choosing an inappropriate transcription system (as discussed in this paper). Critically speaking, research communities could increase their distance from the practice of transcribing, thereby missing out on analyzing rich transcripts and yielding fruitful interpretations.

Regarding segmentation in transcripts, it is sensible for each speaker's contribution to be placed in a single paragraph. If one decides on sentence segmentation, each sentence should contain one coherent contribution. However, it is important to remember that this decision process introduces subjectivity. To address this issue, it can be sensible to adhere to rules for segmentation (e.g., a sentence ends after every pause of at least 0.5 seconds)

[7] https://github.com/huggingface/speechbox/tree/main?tab=readme-ov-file#asr-with-speaker-diarization

and to repeatedly listen to the source data closely while reading along to identify instances that may need correction or discussion with colleagues.

If, however, the research interests require a more sophisticated approach to segmentation, such as intonation phrases, sticking to the GAT2 ruleset is viable but can be rather overwhelming for those outside linguistics. To enable broader use of the benefits of segmenting transcripts into prosodic units such as intonation phrases, alternative methods such as Rapid Prosody Transcription (RPT) — though perhaps not as linguistically rigorous — exist (Cole and Shattuck-Hufnagel, 2016). Although these methods are not based on the GAT2 ruleset, they ensure accessibility for a broader audience to make use of different forms of segmentation. In the first step, all words present in the data are transcribed without punctuation as perceived by one person. Afterward, in two passes, other researchers or research assistants identify (1) prominent words and (2) boundaries of information chunks by listening to the audio while annotating the transcripts from step 1. The method stipulates that they cannot pause or stop the playback; however, they should listen twice per pass. No feedback or sample solution of another transcript should be given to the annotators beforehand, as they should rely solely on their intuition without concern for correctness. For these two passes, they are instructed uniformly ("Mark as prominent words those that the speaker has highlighted for the listener, to make them stand out," and "Mark boundaries between words that belong to different chunks that serve to group words in a way that helps listeners interpret the utterance" (Cole and Shattuck-Hufnagel, 2016)). Afterward, the annotations need to be tested for their reliability. In the original method, Cole and Shattuck-Hufnagel (2016) provide a sophisticated approach that we deem useful only if, for example, at least five people annotate the transcripts. As that is often not the case, we would argue that intercoder reliability tests and intercoder sessions are also sensible and quick alternatives.

6.3. *Transparency in Publications — After Transcribing*

When writing the publication, we would strongly recommend reporting sufficient details to increase transparency in such a way that other researchers can produce transcripts of comparable form. It is advisable to share (excerpts of) the transcripts in open-science repositories in anonymized form, not only for transparency but also in the interest of open science.

To quickly share necessary details about the transcription system that was used, it is advisable to simply refer to a pre-existing system by citing it (Kowal & O'Connell provide a good overview in (Flick, 2013, p. 64ff)). If rules of existing systems were combined or adapted, it is helpful to state whether the (1) verbal component (e.g., verbatim, standard orthography), (2) prosodic component (e.g., rhythm, pitch contour), (3) paralinguistic component (e.g., breathing, laughing), or (4) additional information such as gestures or gaze are included, omitted, or adapted.

7. Conclusion

This paper explores the role and standards of transcription in qualitative computing education research (CER). Besides highlighting the need for better standards, it examines how different transcription choices may influence research outcomes. Drawing on Ethnomethodological Conversation Analysis (EMCA), the paper elaborates on transcription theory. It emphasizes that transcription is more than just a preparatory step; it actively shapes analysis and should be critically considered. A systematic review of 107 CER articles was conducted to assess how transcription standards are reported and to identify ways to improve transparency and rigor. Based on these findings, a three-step framework is proposed for selecting, applying, and communicating transcription strategies. A case study illustrates how a structured approach to transcription enabled the pursuit of a research interest that would have been discarded using a different transcription system. Rather than advocating for a one-size-fits-all solution, this paper emphasizes the importance of selecting a transcription method that aligns with research goals and ensuring clear documentation. The aim is to contribute to a more reflective and consistent approach to transcription in CER.

To improve methodological clarity and the qualitative analysis of spoken data in CER: (1) A conscious reflection on the choice of transcription system in light of the research question is required. For example, if one wants to retrace the thought process of someone who was interviewed about a certain topic, verbatim transcripts may already provide too little information to *closely* retrace the thought process. If one decides that a verbatim transcript contains sufficient information, it should be explained why that is the case. (2) All publications should identify how the transcripts were developed and what information they include. It is good practice to use well-defined transcription systems whenever possible (see, e.g., Flick (2013, p. 64ff) for an overview).

As discussed in the previous sections, there is a plethora of empirical and theoretical ideas that underline the importance of rigor in the context of transcription. Ultimately, an unreflective choice or poor strategy may even lead to the discarding of promising qualitative research ideas. We have provided theoretical foundations, shared practical experiences, argued for better transcription practices, and presented a practical three-step scheme (see 5) that researchers can follow. More rigorous standards can improve the findings of our field's research in the long term, especially regarding:

- **Contextual Richness**: Transcribing layers of conversation that preserve details in addition to semantic content can improve the accuracy and reliability of coding.
- **Pedagogical Insights**: Thorough transcription offers additional, valuable resources for researchers in education, potentially aiding in better understanding, for example, teaching methodologies and student interactions.
- **Inclusivity and Equity**: Accurate transcription acknowledges diverse linguistic expressions, ensuring fair representation and potentially avoiding bias in educational research interpretations.
- **Validity of Results**: Detailed transcription may sometimes be needed to pursue certain research interests. Therefore, better transcription practices can enable researchers to

access interpretations that are inaccessible with other transcription systems. This can improve reliability and, ultimately, the validity of results.

Admittedly, opting for the most complex transcription system and strategy is not always practical, as transcription is "[t]ime consuming and can therefore also be a cost factor, [...] it is important to consider what level of accuracy is truly necessary to answer your research questions" (Kuckartz and Rädiker, 2019). Thus, again, our appeal is not an endorsement of always selecting the most complex system. Rather, we call for more rigor and consciousness of the topic, as the identified potentials can only be realized if the community follows clearer standards. By no means do we argue that the method sections of all qualitative research papers should grow significantly in the future. Instead, this paper encourages researchers to articulate, reflect on, and justify their choice of transcription system and to make their guiding rules for transcript creation transparent.

Higher standards can serve as a catalyst for meaningful progress, deter incorrect or erroneous interpretations, and open up fresh avenues in our field. Our community, despite being relatively young compared to others, has established and matured significantly over time, and the inquiries being addressed are becoming increasingly complex, intricate, and often interdisciplinary. To be adequately prepared and to foster openness for interdisciplinary collaborations, maintaining a rigorous standard in the context of transcription is important to remain both relevant and progressive as a field.

Acknowledgments

We thank the anonymous reviewers of a previous version of this article for their valuable feedback and tips that improved the paper's quality. We also thank our colleagues for their feedback on different drafts of the paper and during discussions. We also would like to thank all research assistants for their support during transcription, data acquisition, and data analysis.

Funding

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021 - 438445824.

References

Aivaloglou, E., Meulen, A.v.d. (2021). An Empirical Study of Students' Perceptions on the Setup and Grading of Group Programming Assignments. *ACM Transactions on Computing Education*, 21(3), 17–11722. https://doi.org/10.1145/3440994.

Alshahrani, A., Ross, I., Wood, M.I. (2018). Using Social Cognitive Career Theory to Understand Why Students Choose to Study Computer Science. In: *Proceedings of the 2018 ACM Conference on International Computing Education Research*. ACM, Espoo Finland, pp. 205–214. 978-1-4503-5628-2. https://doi.org/10.1145/3230977.3230994.

Ashmore, M., Reed, D. (2005). Innocence and Nostalgia in Conversation Analysis: The Dynamic Relations of Tape and Transcript.

Auer, P. (1992). The neverending sentence: Rightward expansion in spoken language. In: Kontra, M., Tamás, V. (Eds.), *Studies in Spoken Language: Englisch, German, Finno-Ugric*. Hungarian Academy of Sciences, Linguistics Institute, Budapest.

Ayaß, R. (2015). Doing data: The status of transcripts in Conversation Analysis. *Discourse Studies*, 17(5), 505–528. https://doi.org/10.1177/1461445615590717.

Bender, E., , S. Niclas, , C. Michael E., , M. Melanie, and Hubwieser, P. (2016). Identifying and Formulating Teachers' Beliefs and Motivational Orientations for Computer Science Teacher Education. *Studies in Higher Education*, 41(11), 1958–1973. https://doi.org/10.1080/03075079.2015.1004233.

Bergmann, P., Mertzlufft, C. (2009). Segmentierung spontansprachlicher Daten in Intonationsphrasen – Ein Leitfaden für die Transkription. In: *Die Arbeit mit Transkripten in Fortbildung, Lehre und Forschung* (Gedruckte ausgabe ed.). Verlag für Gesprächsforschung, Mannheim, pp. 83–95. 978-3-936656-34-3.

Bloch, C. (1996). Emotions and Discourse. *Text & Talk*, 16(3), 323–342. https://doi.org/10.1515/text.1.1996.16.3.323.

Carlson, K. (2009). How Prosody Influences Sentence Comprehension. *Language and Linguistics Compass*, 3(5), 1188–1200. https://doi.org/10.1111/j.1749-818X.2009.00150.x.

Chafe, W. (1994). *DISCOURSE, CONSCIOUSNESS, AND TIME*. Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing. University of Chicago Press, Chicago, IL. 978-0-226-10054-8.

Cole, J., Shattuck-Hufnagel, S. (2016). New Methods for Prosodic Transcription: Capturing Variability as a Source of Information. *Laboratory Phonology*, 7(1). https://doi.org/10.5334/labphon.29.

Couper-Kuhlen, E., Selting, M. (2018). *Interactional linguistics: studying language in social interaction*. Cambridge University Press, Cambridge, United Kingdom ; New York, NY. 978-1-107-03280-4 978-1-107-61603-5.

Davidson, C. (2009). Transcription: Imperatives for Qualitative Research. *International Journal of Qualitative Methods*, 8(2), 35–52. https://doi.org/10.1177/160940690900800206.

Du Bois, J.W., Schuetze-Coburn, S., Cumming, S., Paolino, D. (1993). Outline of Discourse Transcription. In: *Talking Data*. Psychology Press, New York. 978-1-315-80792-8.

Duranti, A. (2006). Transcripts, Like Shadows on a Wall. *Mind, Culture, and Activity*, 13(4), 301–310. https://doi.org/10.1207/s15327884mca1304_3.

Fitzgerald, S., McCauley, R., Plano Clark, V.L. (2011). Report on qualitative research methods workshop. In: *Proceedings of the 42nd ACM technical symposium on Computer science education - SIGCSE '11*. ACM Press, Dallas, TX, USA, p. 241. 978-1-4503-0500-6. https://doi.org/10.1145/1953163.1953237.

Flick, U. (2013). *The SAGE Handbook of Qualitative Data Analysis*. SAGE. 978-1-4462-9669-1.

Fowler, M., Chen, B., Zilles, C. (2021). How should we 'Explain in plain English'? Voices from the Community. In: *Proceedings of the 17th ACM Conference on International Computing Education Research*. ICER 2021. Association for Computing Machinery, New York, NY, USA, pp. 69–80. 978-1-4503-8326-4. https://doi.org/10.1145/3446871.3469738.

Frank, M.G., Griffin, D.J., Svetieva, E., Maroulis, A. (2015). Nonverbal Elements of the Voice. In: Kostić, A., Chadee, D. (Eds.), *The Social Psychology of Nonverbal Communication*. Palgrave Macmillan UK, London, pp. 92–113. 978-1-137-34586-8. https://doi.org/10.1057/9781137345868_5.

Goodwin, C., Heritage, J. (1990). Conversation Analysis. *Annual Review of Anthropology*, 19, 283–307. https://doi.org/10.1146/annurev.an.19.100190.001435.

Jefferson, G. (2004). Glossary of transcript symbols with an introduction. *Conversation analysis*, 13–31. https://doi.org/10.1075/pbns.125.02jef.

Kong, M., Mauriello, M.L., Pollock, L. (2022). Exploring K-8 Teachers' Preferences in a Teaching Augmentation System for Block-Based Programming Environments. In: *Proceedings of the 22nd Koli Calling International Conference on Computing Education Research*. Koli Calling '22. Association for Computing Machinery, New York, NY, USA, pp. 1–12. 978-1-4503-9616-5. https://doi.org/10.1145/3564721.3564725.

Kroes, P. (1998). Technological Explanations. *Phil & Tech*, 3(3), 124–134. https://doi.org/10.5840/techne19983325.

Kuckartz, U. (2014). *Qualitative Text Analysis: A Guide to Methods, Practice and Using Software*. SAGE. 978-1-4462-9776-6.

Kuckartz, U., Rädiker, S. (2019). Transcribing Audio and Video Recordings. In: Kuckartz, U., Rädiker, S. (Eds.), *Analyzing Qualitative Data with MAXQDA: Text, Audio, and Video*. Springer International Publishing, Cham, pp. 41–49. 978-3-030-15671-8. https://doi.org/10.1007/978-3-030-15671-8_4.

Lapadat, J.C. (2000). Problematizing transcription: Purpose, paradigm and quality. *International Journal of Social Research Methodology*, 3(3), 203–219. https://doi.org/10.1080/13645570050083698.

Mandal, F. (2014). Nonverbal Communication in Humans. *Journal of Human Behaviour in the Social Environment*, 24, 417–421. https://doi.org/10.1080/10911359.2013.831288.

Mondada, L. (2011). Understanding as an embodied, situated and sequential achievement in interaction. *Journal of Pragmatics*, 43(2), 542–552. https://doi.org/10.1016/j.pragma.2010.08.019.

Mondada, L. (2018). Multiple Temporalities of Language and Body in Interaction: Challenges for Transcribing Multimodality. *Research on Language and Social Interaction*, 51(1), 85–106. https://doi.org/10.1080/08351813.2018.1413878.

Mondada, L. (2019). Transcribing silent actions: a multimodal approach of sequence organization. *Social Interaction. Video-Based Studies of Human Sociality*, 2(1). https://doi.org/10.7146/si.v2i1.113150.

Murphy, L., Fitzgerald, S., Hanks, B., McCauley, R. (2010). Pair Debugging: A Transactive Discourse Analysis. In: *Proceedings of the Sixth International Workshop on Computing Education Research*. ICER '10. Association for Computing Machinery, New York, NY, USA, pp. 51–58. 978-1-4503-0257-9. https://doi.org/10.1145/1839594.1839604.

Ochs, E. (1979). Transcription as Theory. (2nd ed.), New York.

O'Connell, D.C., Kowal, S. (1995a). Basic Principles of Transcription. In: *Rethinking Methods in Psychology*. SAGE Publications Ltd, 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom, p. 93. 978-0-8039-7733-4 978-1-4462-2179-2. https://doi.org/10.4135/9781446221792.

O'Connell, D.C., Kowal, S. (1995b). *Rethinking Methods in Psychology*. SAGE Publications Ltd, London. https://doi.org/10.4135/9781446221792.

Oleson, A., Xie, B., Salac, J., Everson, J., Kivuva, F.M., Ko, A.J. (2022). A Decade of Demographics in Computing Education Research: A Critical Review of Trends in Collection, Reporting, and Use. In: *Proceedings of the 2022 ACM Conference on International Computing Education Research - Volume 1*. ACM, Lugano and Virtual Event Switzerland, pp. 323–343. 978-1-4503-9194-8. https://doi.org/10.1145/3501385.3543967.

Oliveira, M. (2002). The Role of Pause Occurrence and Pause Duration in the Signaling of Narrative Structure. In: Ranchhod, E., Mamede, N.J. (Eds.), *Advances in Natural Language Processing*. Springer, Berlin, Heidelberg, pp. 43–51. 978-3-540-45433-5. https://doi.org/10.1007/3-540-45433-0_7.

Olsen, R.M. (2019). The Acoustics of Feeling: Emotional Prosody in the StoryCorps Corpus. *The Journal of the Acoustical Society of America*, 145(3_Supplement), 1930. https://doi.org/10.1121/1.5102025.

Olson, D.R. (1993). How writing represents speech. *Language & Communication*, 13(1), 1–17. https://doi.org/10.1016/0271-5309(93)90017-H.

O'Connell, D.C., Kowal, S. (2009). Transcription systems for spoken discourse. *John Benjamins Publishing*, 4, 240.

Park, J.S.-Y. (2002). Cognitive and interactional motivations for the intonation unit. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 26(3), 637–680. https://doi.org/10.1075/sl.26.3.07par.

Point, S., Baruch, Y. (2023). (Re)thinking transcription strategies: Current challenges and future research directions. *Scandinavian Journal of Management*, 39(2), 101272. https://doi.org/10.1016/j.scaman.2023.101272.

Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. https://doi.org/10.48550/ARXIV.2212.04356.

Ralph, P., Ali, N.b., Baltes, S., Bianculli, D., Diaz, J., Dittrich, Y., Ernst, N., Felderer, M., Feldt, R., Filieri, A., de França, B.B.N., Furia, C.A., Gay, G., Gold, N., Graziotin, D., He, P., Hoda, R., Juristo, N., Kitchenham, B., Lenarduzzi, V., Martínez, J., Melegati, J., Mendez, D., Menzies, T., Molleri, J., Pfahl, D., Robbes, R., Russo, D., Saarimäki, N., Sarro, F., Taibi, D., Siegmund, J., Spinellis, D., Staron, M., Stol, K., Storey, M.-A., Taibi, D., Tamburri, D., Torchiano, M., Treude, C., Turhan, B., Wang, X., Vegas, S. (2021). Empirical Standards for Software Engineering Research. arXiv. https://doi.org/10.48550/arXiv.2010.03525.

Roberts, C., Street, B. (2017). Spoken and Written Language. In: *The Handbook of Sociolinguistics*. John Wiley & Sons, Ltd, pp. 168–186. 978-1-4051-6625-6. https://doi.org/10.1002/9781405166256.ch10.

Romero, C., O'Connell, D.C., Kowal, S. (2002). Notation Systems for Transcription: An Empirical Investigation. *Journal of Psycholinguistic Research*, 31(6), 619–631. https://doi.org/10.1023/A:1021217105211.

Rudzicz, F. (2016). *Clear Speech: Technologies that Enable the Expression and Reception of Language*. Morgan & Claypool Publishers. 978-1-62705-827-8.

Sanders, K., Vahrenhold, J., McCartney, R. (2023). How Do Computing Education Researchers Talk About Threats and Limitations? In: *Proceedings of the 2023 ACM Conference on International Computing Education Research V.1*. ACM, Chicago IL USA, pp. 381–396. 978-1-4503-9976-0. https://doi.org/10.1145/3568813.3600114.

Schulte, C. (2008). Duality Reconstruction - Teaching digital artifacts from a socio-technical perspective. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5090 LNCS, 110–121. https://doi.org/10.1007/978-3-540-69924-8_10.

Selting, M. (2000). The construction of units in conversational talk. *Language in Society*, 29(4), 477–517. https://doi.org/10.1017/S0047404500004012.

Selting, M., Auer, P., Barth-Weingarten, D. (2011). A system for transcribing talk-in-interaction : GAT 2. *Gesprächsforschung : Online-Zeitschrift zur verbalen Interaktion*, 12, 1–51.

Sidnell, J. (2012). Basic Conversation Analytic Methods. In: Sidnell, J., Stivers, T. (Eds.), *The Handbook of Conversation Analysis* (1st ed.). Wiley, pp. 77–99. 978-1-4443-3208-7 978-1-118-32500-1. https://doi.org/10.1002/9781118325001.ch5.

Simpson, H.E. (2016). *The role of intonation units in memory for spoken English*. Dissertation, University of California, Santa Barbara, Santa Barbara, Calif..

Speer, S.R., Ito, K. (2009). Prosody in First Language Acquisition – Acquiring Intonation as a Tool to Organize Information in Conversation. *Language and Linguistics Compass*, 3(1), 90–110. https://doi.org/10.1111/j.1749-818X.2008.00103.x.

Spiegel, C. (2009). Transkripte als Arbeitsinstrument: Von der Arbeitsgrundlage zur Anschauungshilfe. In: *Die Arbeit mit Transkripten in Fortbildung, Lehre und Forschung* (Gedruckte ausgabe ed.). Verlag für Gesprächsforschung, Mannheim, pp. 7–15. 978-3-936656-34-3.

Stein, S. (2018). 1. Oralität und Literalität. In: *Handbuch Text und Gespräch*. De Gruyter, Berlin, Boston, pp. 3–25. 978-3-11-029605-1. https://doi.org/10.1515/9783110296051-001.

Tenenberg, J., Chinn, D. (2019). Social Genesis in Computing Education. *ACM Transactions on Computing Education*, 19(4), 34–13430. https://doi.org/10.1145/3322211.

Terfloth, L., Schaffer, M., Buhl, H.M., Schulte, C. (2023). Adding Why to What? Analyses of an Everyday Explanation. In: Longo, L. (Ed.), *Explainable Artificial Intelligence*. Communications in Computer and Information Science. Springer Nature Switzerland, Cham, pp. 256–279. 978-3-031-44070-0. https://doi.org/10.1007/978-3-031-44070-0_13.

Lutz Terfloth is a PhD student at the Computing Education Research Group at the University of Paderborn, Germany. His primary research focuses on the empirical study of explanations of technological artifacts. In his dissertation, he develops a novel analytical approach grounded in techno-philosophical theory to systematically investigate how explanations about technology are constructed and understood. By integrating philosophical perspectives with empirical research methods, his work aims to deepen our understanding of the processes and challenges involved in making complex technological systems explainable in educational and practical contexts.

Vivien Lohmer is a PhD student at the Transregional Collaborative Research Centre 318 (Sonderforschungsbereich Transregio 318) at Bielefeld University, where she works as a research associate and doctoral researcher in project A04. Her research focuses on multimodality and gestural behavior in everyday explanations.

Friederike Kern teaches German linguistics and their didactics at Bielefeld University. After studying German literature, linguistics and philosophy in Berlin and London, she was awarded the Dr. phil., from the University of Hamburg on the communicative differences between East and West Germans in Job Interviews. Her research interests are in the area of Ethnomethodological Conversational Analysis, multimodal interaction analysis and ethnography, discourse acquisition and multimodal language development, classroom interaction, and learning in multimodal environments. Her publications include work on rhythm in Turkish German, on the development of children's multimodal storytelling and explanations, and on interactions in learning situations.

Carsten Schulte is a professor for Computing Education Research at Paderborn University, Germany. His work and research interests are the philosophy of computing education, artificial intelligence in education, and empirical research on teaching and learning processes (including eye movement research). Since 2017, he has been working together with Didactics of Mathematics (Paderborn University) on the ProDaBi project, in which data science and artificial intelligence are prepared as teaching topics. He is also a PI in the collaborative research centre TRR318 'Constructing Explainability' on explainable AI.