

VIDEO-MEDIATED TALK-IN-INTERACTION: CHALLENGES FOR THE TRANSCRIPTION OF FACIAL GESTURES

Ulrike Schröder*

Flavia Fidelis de Paula**

- **ABSTRACT:** The pandemic has restricted face-to-face interactions while accelerating the globalization of communication through video calls. This has facilitated the widespread conduct of activities such as collaborative work, scientific conferences, online teaching, medical consultations, religious rites, and family gatherings. These practices have not only promoted the expansion of local boundaries but have also connected the world on a global scale, enhancing intercultural encounters increasingly characterized by multilingual conversations. In this new configuration, “talking heads” assume central importance. This paper presents and compares two methodological models for transcribing facial gestures in video-mediated interactions, as proposed by Brunner and Diemer (2021) and Dix (2022). Using a multimodal conversation analysis approach, we apply these models to a video-mediated interaction (VMI). We will demonstrate the need to develop methodologies that capture the multimodal complexity of facial gestures in these interactions, highlighting how these systems can contribute to both quantitative and qualitative research methods integrate linguistic and gestural dimensions. Additionally, we will discuss the advantages and disadvantages of the two systems and propose enhancements.
- **KEYWORDS:** Multimodal transcription; Video-mediated interaction (VMI); Facial gestures; English as a lingua franca; Intercultural communication.

Introduction

The rise of the digital and virtual world was already underway when the COVID-19 pandemic accelerated the transition to a new normal. This new reality quickly pushed us to learn how to interact through various video-mediated technologies, transcending boundaries from local to global. People began using videoconferencing to chat with friends, collaborate at work, attend meetings, consult healthcare professionals, participate

* Universidade Federal de Minas Gerais (UFMG), Faculdade de Letras, Belo Horizonte, MG, Brasil. Professora Titular. schroederulrike@gmx.com. ORCID: <https://orcid.org/0000-0001-7764-7249>

** Universidade Federal de Minas Gerais (UFMG), Faculdade de Letras, Belo Horizonte, MG, Brasil. Pesquisadora. flaviafid@gmail.com. ORCID: <https://orcid.org/0000-0002-9427-8745>

in international conferences, expand institutional activities, engage in religious rituals, and enhance their leisure experiences through the virtual world. Thus, the restrictions on in-person interaction imposed by the pandemic were accompanied by a rapid opening of various multilingual and intercultural communities as well.

Regarding this digital-virtual universe, prior to the pandemic, research in the fields of linguistics, communication, and interaction was predominantly focused on data from social networks and platforms such as Facebook, WhatsApp, X (formerly Twitter), and YouTube. Attention to virtual interaction in talk-in-interaction is a more recent or an emerging topic, particularly in multimodal conversation analysis (Due; Licoppe, 2021) and distance learning (Salomão; Freire Junior, 2020). In an online and synchronous ‘virtual interaction,’ interlocutors connect through technologies that provide simultaneous access to sound and image in real time (Stivers; Timmermans, 2017; Fidelis De Paula, 2023, p. 12). Recognizing the need to explore this dynamic in a linguistic, interactional, and multimodal manner, in 2022, we initiated the research project *The Multimodal Coordination of Intercultural Video-Mediated Interaction*.¹ This project aims to investigate video-mediated interaction under a multisensory and embodied approach *in situ*, as well as explore the growing intercultural dimension of online interaction. A key question that arises within the new epistemology of virtual interaction is that of transcription. Conversation analysis (CA), interactional linguistics (IL), and gesture studies (GS) have significantly influenced other fields of linguistic studies over the last few decades, contributing to the understanding of an ‘embodied turn’ or ‘multimodal turn’ (Mondada, 2019; Neville, 2015) in language-in-use investigations. However, the transition to virtual interaction raises essential questions, particularly regarding transcription. The body, previously central to interactional studies, undergoes a kind of inverted transposition, given the predominance of “talking heads” (Licoppe; Morel, 2012) in virtual communication. This shift raises questions about participant’s attention, the data provided by the screen and video recording, and therefore, the specific challenges for appropriate transcription.

Thus, the need to reflect on the transcription of video-mediated interactions arises primarily from two aspects: (i) the unstoppable use of digital space for human communication, accelerated since the pandemic, and (ii) a methodological gap that stems from, among other factors, a certain neglect of facial gestures, accompanied by an unsatisfactory integration of these gestures into available transcription systems. In Section 1, *Methodological Framework*, we will address this issue through a discussion of the specific characteristics of video-mediated interaction. We will then discuss the advancements and gaps in multimodal transcription, with a particular focus on facial gestures. We propose a classification of facial gestures, and their functions based on Bavelas and Chovil (Bavelas, 2022), aligned with the theoretical pillars of conversation analysis (CA), interactional linguistics (IL), and gesture studies (GS). In Section 2, we will present two current systems developed to tackle challenges in multimodal

¹ Funded by CAPES and DAAD (2023-2026), see www.lettras.ufmg.br/icmi

transcription: (i) GAT 2+ISWA (DIX, 2022), a system oriented toward the ‘visual turn,’ and (ii) ViMELF (Brunner; Diemer, 2021), designed to handle rich data and corpora of video-mediated English as a Lingua Franca (ELF). We will illustrate the operation of these systems with a short example and conclude with a discussion of their advantages and disadvantages in Section 3.

Methodological framework

Video-mediated talk-in-interaction in the global world

The study of video-mediated talk-in-interaction, in comparison to the abundance of other phenomena in the digital world, is still in its embryonic stage. Over the past two decades, there has been an ‘embodied turn’ in the field of research on interaction, cognition, and language, which marked a dissociation from the neo-Cartesian dualism towards a new claim of intercorporeality (Streeck, 2021).² According to this view, (inter)action, gesture, language, and cognition are interconnected and inherently social phenomena, as individuals naturally inhabit each other’s actions. Therefore, a praxeological and phenomenological approach to *in situ* interaction is based on a holistic understanding of communication, which also conceptualizes gestures as “coupled with the environment” (Goodwin, 2018). In addition to bodily-visual resources, manual gestures, and, especially in the case of online and synchronous interactions, facial gestures (Kendon, 2004; Bavelas; Gerwing; Healing, 2014), prosody assumes a crucial role in embodied interaction. Under the influence of interactional linguistics (IL), prosody has gained new research focus, encompassing “the ‘musical’ attributes of speech—auditory effects such as melody, dynamics, rhythm, timing, and pauses” (Couper-Kuhlen; Selting, 1996, p. 11).

Since its inception in 2010, our Research Center *Intercultural Communication in Multimodal Interactions*³ has focused empirically on real interpersonal and intercultural interactions, meaning that physical presence was conceived as a shared space concept, involving intimacy and immediacy. In contrast, video-mediated interaction involves a series of changes compared to face-to-face encounters in presence. These differences are of high relevance epistemologically, methodologically, and analytically, and should be considered when addressing data from video-mediated interactions. In summary (cf., among others, Due; Licoppe, 2021; Friesen, 2014; Loenhoff, 2012; Licoppe; Morel, 2012; Manstead; Lea; Goh, 2011; Norris; Prini, 2016; Balaman; Doehler, 2022):

² In this article, we adopt a rigorously communicative perspective, that is, *from within*, in accordance with the Conversation Analysis paradigm, grounded in the social phenomenology of Alfred Schütz and in studies from the field of ethnography. Discussions related to other research fields go beyond the scope of this article.

³ www.lettras.ufmg.br/icmi

- (a) Instead of full-body interactions, interlocutors engage in an “ecological eye-to-eye cluster” (talking heads);
- (b) The visual and auditory spaces are fragmented (lighting, set design, body segmentation, vocal characteristics, background noise, etc.);
- (c) Visual interruptions (screen resolution) and auditory interruptions (staccatos, echoes, volume, etc.) occur;
- (d) There are no transitional spaces for informal communication (e.g., elevators, hallways) in institutional interactions;
- (e) The basic mode of experience is 2D instead of 3D (e.g., when objects come into play);
- (f) Interpersonal distance is usually reduced (e.g., in meetings, lectures, etc.);
- (g) Interlocutors must deal with the dilemma of eye contact, as eye movements do not guide attention, allocation, or turn-taking;
- (h) Attention is divided, as interlocutors also control their own transmitted image.

Considering these significant phenomena that have already been revealed as having a crucial impact on online and synchronous interaction, it should be noted that they may become even more relevant in multilingual and intercultural lingua franca communication, situations where the prevention of misunderstandings (‘pre-empting strategies’), for example, becomes highly important. Recently, scholars in conversation analysis have followed a path toward a video-based data collection methodology, aiming to explore, in a very detailed and granular way, how video-mediated interaction is carried out moment by moment. In doing so, it became clear that there are many open questions, both old and new, that arise due to this new context, particularly related to transcription. In the next section, we will draw attention to this issue.

The pitfalls of multimodal transcription

Multimodal transcription is not as recent as it may seem. Its roots lie in ethnographic research, in contrast to conversation analysis (CA), whose interest, for approximately three decades, was almost exclusively related to the verbal level, as reflected in the Jefferson system (2004). The term *multimodal* can be misleading as it suggests an addition of modes, since the idea of ‘multiple modalities’ carries the semantics of separate channels that need to be brought together, even though it refers to an intermodal relationship that contributes to the richness and density of meaning construction. Strictly speaking, the concept of multimodality contrasts with a more holistic, gestural, and synesthetic understanding of our experience (Lyons, 2016; Deppermann, 2013).

With this ‘renaissance’ of the body in the fields of linguistics, communication, and interaction, the process of transcription has gained prominence as an issue of reliability. However, since it is neither possible nor practical to transcribe every aspect of a multimodal interaction, transcribers always face the choice of what to transcribe

and what to omit, which modes to focus on, and what level of detail to include (Lyons, 2016, p. 274). Furthermore, the primary data undergoes an initial reduction of the original during the video recording, which becomes even more relevant in the case of recordings of virtual interactions, as usually only the virtual excerpt is recorded, while the excerpt of the real environment, which includes other parts of the interlocutors' bodies, is neither present in the transcription nor in the video. Thus, the assumption that it is the camera's position that determines which aspects of the given reality are selected for video recording (Cruz *et al.*, 2019) becomes of paramount importance.

The emergence of the multimodal perspective brought significant changes to transcription conventions as well. Transcription shifted from simply converting speech into written form to becoming a multifaceted and multilayered effort, with a preference for tabular transcription formats over sequences that might seem overly focused on verbal content in favor of a holistic approach to multimodal communication. Tabular transcriptions typically segregate specific modes, such as gaze, gestures, vocalization, and interactions with artifacts and objects, into columns. They aim to preserve information about how these modes interact with one another, enabling vertical and horizontal readings to capture the simultaneous occurrence of modes at a specific point in time, as well as the synchronous layout of the timeline (Cowan, 2014; Flewitt *et al.*, 2017).

The issue of how to handle rich data has become increasingly urgent as more and more datasets become available through online sources or multimodal compilation projects. Over the past few years, attempts have emerged to develop new ways to transcribe multimodal interactions based on conventions that represent different levels of resources (Heath *et al.*, 2010; Mondada, 2019). However, to date, no format has reached the desired level of completeness. It remains challenging to provide an adequate representation of the intricate simultaneity of different modes and to uniformly represent visual and bodily resources. Additionally, the excessive complexity of some transcription systems has generated criticism from readers and researchers, who point out the lack of clarity and readability. Bezemer and Jewitt (2010) highlight the dilemma researchers face: too much attention to many different modes can hinder the understanding of the meanings of a specific mode; too much attention to a single mode risks 'fixing things' in just one of the many ways in which people assign meaning. Furthermore, there is no standard way to transcribe rich data. There will always be – and should be – adaptability in transcription to specific data, research questions, and intended use. It is important to emphasize that a key criterion is to maintain a clear alignment and representation of simultaneity and orchestration of different interactive resources, as well as a certain expansibility (onion principle) of transcription in relation to different levels of detail.

As Schröder (2025) points out, a possible way forward could be a two-step procedure: first, create a classic sequential transcription (GAT 2) as the initial step of a larger sequence, and then create a 'zoom-in' as the second step using a tabular transcription for gesture annotation. In gesture studies, for many decades, there has been a primary focus on manual gestures alongside a certain exclusion of a holistic view,

not only of other resources such as prosodic cues but particularly also of other types of gestures, such as facial gestures. In virtual interactions, most of a person's body is 'cut off,' and as we discussed in Section 1.1, at the same time, the distance between interlocutors is reduced. Therefore, facial gestures become more prominent, and the transcriber realizes that there is a rather fragmented repertoire available to deal with this new transcription universe. Bavelas and Chovil (2006), who have studied facial gestures for more than two decades, still stated in 2006 that the study of facial gestures presents "new directions of research in this relatively neglected area" and repeatedly emphasize that "facial gestures in conversation are not emotional expressions" (Bavelas; Chovil, 2018, p. 98), a view that contrasts with what much research suggests. This is why we will focus in the next section on this specific phenomenon to illustrate (i) the particular relevance of facial gestures for the transcription of virtual interaction; (ii) their epistemological conceptualization, and (iii) how to develop, based on this, categories for the transcription of talking heads.

Categories and functions of facial gestures

The analysis of facial gestures is fundamental to understanding human language, as it plays a role in both the perception and production of communication by providing cues related to emphasis, (dis)approval, (dis)engagement, negation, confirmation, lexical repair, and other interactional phenomena among participants. In the context of video-mediated interaction, where participants can see themselves on screen but may struggle to accurately identify who or what they are observing on their devices, facial gestures such as smiles, gaze direction, and head nods significantly alter the systematicity of adjacency pairs and turn allocation in interaction. Moreover, the gestural dynamics of these categories enable participants to communicate availability to initiate an interaction, signal structural (mis)alignment, show social affiliation, and create sequences of (self-)repair (Hjulstad, 2016; Fidelis De Paula, 2023). Given that the lack of visual resources in virtual interactions can challenge the analysis, comprehension, and precise interpretation of conversational elements, paying attention to facial gestures becomes even more critical.

Thus, this section aims to present an overview of facial gestures in the transcription of virtual interactions, alongside their epistemological conceptualization and the development of categories for transcribing 'talking heads.' This framework is based on the referential and pragmatic categories and functions proposed by Bavelas (2022).

Kendon (2004) defines 'gestures' as actions forming a single integrated system, produced within the context of speech or equivalent to a complete utterance. As such, they contribute to the understanding of multimodality in co-verbal expression. In line with this perspective, facial gestures refer to actions involving movements or configurations of the head and/or the muscles and components of the face, encompassing

areas such as the eyebrows, eyes, and mouth. These gestures are often characterized by the anatomical and functional coordination of these movements.

Similarly to manual gestures, facial gestures share several properties. They harmoniously synchronize with the word(s) they complement and enhance, highlighting the interconnection between physical and expressive components within the dynamics of interaction. Consequently, they possess remarkable communicative capacity, providing semantic, syntactic, and pragmatic information. Additionally, facial gestures, along with other types of gestures, are strongly tied to the context of their production and the interactional space (Mondada, 2013), meaning the participants' reciprocal and coordinated bodily arrangements within a given space. This implies that such gestures do not convey meanings in isolation, nor do they correspond to the emotional 'expressions' of individuals (Bavelas; Chovil, 2018; Bavelas, 2022) in oral-auditory modality languages.

This distinction is significant, as the field of psychology dominated the scientific and popular understanding of 'facial expressions' for many decades. The most prominent approach to facial expressions was developed by Ekman and Friesen (1978). Assuming that the face is the primary site for displaying affect, the authors proposed that there is a fixed set of muscular configurations corresponding to innate and universally recognized emotions. This is reflected in their Facial Action Coding System (FACS),⁴ which allows researchers to generate facial expression outputs based on the fusion of nine emotions (including 'anger,' 'contempt,' 'disgust,' 'fear,' 'sentimentality,' 'joy,' 'sadness,' 'surprise,' and 'confusion') and twenty facial expressions, derived from measuring facial muscle activation, to reflect the subject's emotional engagement.

To underscore their interactional importance and variable *in situ* meaning, Bavelas and Chovil prefer the term "facial gesture."

On the other hand, facial gestures exhibit a significant difference in terms of timing and execution. Specifically, they do not require allocation to an exact position within the gestural unit or subdivision into phases, as is typical of manual gestures (Bressemer; Ladewig; Müller, 2013), because their execution tends to be extremely brief. According to Bavelas and Chovil (2018), the distinctiveness of these gestures lies precisely in their ability to occur within fractions of a second and their potential for strong synchronization with speech. However, this does not imply that facial gestures are necessarily tied to a specific point in the oral production.

In a study on the contribution of facial gestures to the perception of pragmatic functions – such as multimodal intonation and focus – in Brazilian Portuguese, Carnaval *et al.* (2023, p. 25) demonstrated that certain occurrences of facial gestures can persist throughout the entire production of a conversational unit without necessarily being synchronized with a specific moment of speech.

The studies conducted by Chovil (1997, 2005), dedicated to analyzing the meaning and function of facial gestures in specific conversational instances, laid the groundwork

⁴ <https://blog.affectiva.com/emotion-ai-101-all-about-emotion-detection-and-affectivas-emotion-metrics>

for the categories and functions outlined by Bavelas (2022). Using an inductive approach, Chovil identified four general categories, each encompassing up to 12 specific functions associated with facial gestures. Bavelas's (2022) subsequent proposal introduced a broader classification, furthering the understanding of facial gestures by categorizing them into

- (a) personal reactions,
- (b) portrayals,
- (c) motor mimicry,
- (d) smiles, head nods, and gaze orientations.

The gestures executed within each of these categories serve specific functions, enriching the understanding of the complexities associated with facial gesture language in its various manifestations in conversational interactions.

Based on a situational and interactional understanding of 'facial gestures' as opposed to 'facial expressions,' these functions not only overlap but also vary depending on the context, which is always in constant flux. The first category proposed by Bavelas (2022), termed *personal reactions*, is frequently associated with referential functions and represents one of the most common occurrences of facial gestures. Actions such as raising eyebrows, widening the eyes, or dropping the jaw reflect speakers' reactions to the ongoing conversation and perform various roles in the interaction. These include emphasizing, providing feedback or confirmation (backchannel), expressing thoughts or recollections, highlighting an interactional element, offering or suggesting, or even performing a facial shrug—a movement equivalent to shrugging the shoulders. It is worth noting that although listeners may also produce facial gestures as personal reactions during interactions, this occurs less frequently compared to speakers.

Portrayals are performed to illustrate a scene, expression, or reaction that previously occurred, either involving the speaker or someone being described during the conversation. These gestures encompass variations in facial expressions and other categories of facial gestures, such as head nods and gaze orientation. They are closely tied to the referential functions of language and can indicate sentence transitions, negation, social focus, turn completion, concession ("but"), sarcasm, or humor.

Facial gestures classified as *motor mimicry* are predominantly executed by listeners in response to the ongoing talk, as though they were placing themselves in the speaker's position during a given moment in the narrative (Bavelas, 2007). These gestures often serve as empathetic reactions, demonstrating that the listener is emotionally connected to the narrated experience. They may show surprise, astonishment, or even mild shock in response to the account. Within this classification, actions such as raising eyebrows, wrinkling the nose, or slightly widening the eyes are recognized for providing an appropriate response to the speaker's expression. Consequently, they can fulfill interactional functions such as signaling topic closure, concluding a conversation,

demonstrating understanding or agreement, explaining, initiating a topic shift, or even signaling a conversational pause.

The *smile*, unlike the other categories, has received significant attention in the study of embodied gestures (cf. Goodwin; Goodwin, 2000; Kendon, 2004; Kaukomaa *et al.*, 2013; Bavelas; Chovil, 2018). Research has identified a variety of functions for smiles in face-to-face interactions. These include the ability to regulate different stages of interaction (Kendon, 2004), serve as a signal for intraturn responses and/or confirmations similarly to head nods (Bavelas, 2022), anticipate or provide cues about conversational content, highlight elements of the ongoing speech, and express humor, irony, or self-deprecation (Bavelas; Chovil, 2018).

Head nods are commonly used to express agreement or approval, as well as negation or disagreement. Additionally, speakers may utilize head angles and movements to represent relationships, emphasize specific points in the conversation, highlight the speaker's role and perspective, or draw attention to a particular direction. Gestures involving head nods frequently combine with gaze orientation to emphasize specific elements in communication, allowing interactants to reinforce the clarity and impact of their non-verbal suggestions.

Gaze orientation, in turn, functions both to regulate turn-taking in conversation (Kendon, 1967; Fidelis De Paula, 2023) and to direct attention or highlight manual gestures, thereby assuming a deictic function (Hjulstad, 2016, 2018). Specifically, the exchange of gazes can aid in the temporal coordination of spoken turns by signaling or requesting confirmation (backchannel).

In summary, the analysis of facial gestures, whose progression is closely tied to the communicative purposes of interlocutors, demonstrates that the significance of these embodied resources extends beyond in-person contexts. It also holds relevance in virtual interactions, directly influencing the accuracy of transcription and the understanding of communicative processes, particularly in a globalized world and during intercultural meetings, as initially discussed.

The following section introduces two recently developed transcription systems: (a) GAT 2 in conjunction with ISWA, proposed by Dix (2022), and (b) ViMELF, by Brunner and Diemer (2021), designed for interaction in English as a lingua franca in video-mediated environments. This methodological approach seeks to provide a more comprehensive and structured understanding of facial gestures and their importance in virtual interactions, aiming to establish concrete tools and a methodologically sound foundation for future research.

Illustration and discussion of two proposed transcription systems for facial gestures in VMI

GAT 2 + ISWA (DIX 2022)

Based on the GAT 2 framework, Carolin Dix (2022) recently developed a proposal to combine transcription conventions for talk-in-interaction with the International SignWriting Alphabet (ISWA) inventory developed by Sutton (2010). The researcher's intent is to provide a tool capable of representing and integrating different levels of bodily resources. As discussed in Section 1.2, Dix critiques the fact that most multimodal transcriptions focus on a single resource, such as manual gestures or gaze. On the other hand, she emphasizes that this integration requires the transcription to be adaptable to the corresponding database, research objective, and intended use.

Broadly speaking, this means that it is possible to integrate all phenomena in a detailed and precise manner while accounting for their orchestration and simultaneity. However, similar to GAT 2, the 'onion' principle applies, where transcriptions are performed across three distinct levels of granularity (Schröder; Nascimento; Silva, 2019). Regarding the notation of bodily aspects within ISWA, this can be further expanded depending on the research focus (Dix, 2022).

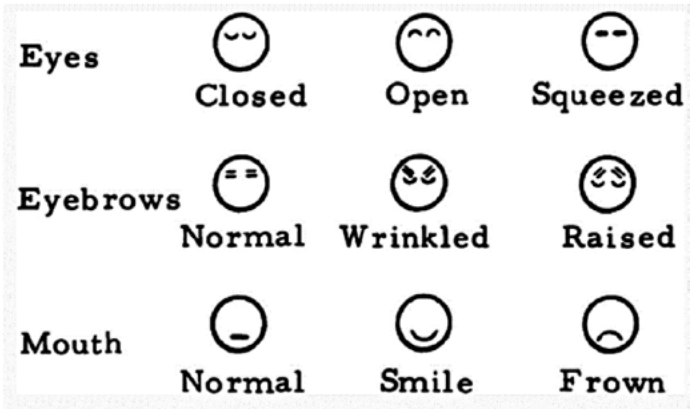
The most relevant feature of ISWA's pictograms is their universal status. Developed in the 1970s by Valerie Sutton (2010), ISWA comprises symbolic icons for notating sign languages. The notations adopt the producer's perspective and cover five categories: (a) hands, (b) movements, (c) dynamics, (d) head and face, and (e) body (<https://www.signbank.org/iswa/>). Dix (2022) demonstrates how a score transcription is created by inserting visual lines below the verbal line (GAT 2) to correspond with the different levels of resources (e.g., body orientation, head positioning, hand position, shoulder movement, facial gestures).

Without delving into details, it can be summarized that the primary challenge lies in the legibility of the symbols and their integration into a transcription system for talk-in-interaction. This challenge arises from the fact that ISWA includes a total of 261 symbols, which are not always easily comprehensible to someone unfamiliar with the system.

In examining the ISWA symbols and the way Dix integrates them into her multimodal transcription proposal, we were particularly struck by the extensive repertoire of symbols for facial resources and their alignment with the forms and functions of facial gestures established by Bavelas (Section 1.3). This category encompasses symbols for the following subcategories: (a) Head movement and face direction; (b) Eyebrow shapes, eye shapes, and gaze direction; (c) Cheeks, nose, ears, and breathing; (d) Mouth and lip shapes; and (e) Tongue, teeth, chin, and neck.

Some symbols, which are easily understandable due to their iconic nature, are illustrated in Figure 1.

Figure 1 – Symbols for Eyes, Eyebrows, and Mouth



Source: Sutton (1982, p. 82 *apud* Dix, 2022, p. 114)

Below, we present the transcription of the introduced sequence using selected symbols from the ISWA inventory. Our methodological framework is still under development; however, after several trials, we have decided—for the purpose of this initial discussion and to ensure maximum readability and adherence to the principle of relevance—to retain a single level of transcription for all bodily resources rather than adopting a score transcription approach.⁵

The first line displays the transcription following GAT 2 conventions (Schröder *et al.*, 2016), along with abbreviations indicating the participants’ origin, in accordance with ICMI standards, as the conversation takes place in a video-mediated interaction in ELF (*Co=Colombian; Br=Brazilian*). The second line illustrates the facial gestures and their timing (*HF = Head; Face*).

⁵ To view the format of a transcription with multiple levels, including the insertion of (potentially) all categories of symbols, and particularly to understand the GAT 2 transcription – though this will not be discussed here – please refer to the Appendix.

Figure 2 – Transcription of Sequence 1 according to GAT 2 + ISWA (DIX, 2022)

Sequence 1: *To face the winter* 2019SkyBrCo01 ((07:13-07:36))⁶

01 Co: (--) so i feel like there's a lot of (.) ↓CULTural 'difference,
HF (KK)

02 Br:
HF (OO) ...

03 Co: (-) between (--) how pEople !↓AC!ting with 'their time?
HF (OO) (KK)

04 that WE do because temporary I WASN'T prepA:red;
HF (UU)

05 Br:
HF (OO)

06 (--) to (--) con!FRONT!.>
HF
.....

07 (--) con'FRONT?>
HF (KK)

08 Co: like (.) to to ↑FACE the wInter?
HF (OO)

09 Br:
HF (OO)

10 Co:
HF (OO)

11 Br: yeah to FACE;
HF (OO) (OO)

12 Co: to 'FADE the wInter because;
HF (OO)

13 Co: (-) like in colOmbia we don't have the STAtions;=
HF (KK)

14 =like the 'SEAsOns.=↓SORry-
HF (KK)

15 (-) we don't have SEAsOns at All (.) sO:;
HF (OO)

16 Br:
HF (OO)

Source: Designed by the author

⁶ The video can be accessed at the following link: <https://youtu.be/t-wjyou1rTE>

Below, we will briefly exemplify the reading and possible understanding of some icons to illustrate how our adaptation of Dix's system (2022) for the multimodal transcription of video-mediated talk-in-interaction:

In lines 01, 03, 07, 13 and 14, it can be seen that the Colombian (Co), in these specific moments, next to the speech, instead of looking at the screen, looks up and to the right, or to the left. This prototypically indicates access to memory rather than addressing the interlocutor, which corresponds to the verbal level in which there are simultaneous metacommunicative insertions. For example, in lines 6 to 7, he inserts a metacomment on his own uncertainty of having chosen the correct term, indicated by the insertion of hesitation markers and pauses. When looking up and to the left, he displays a 'thinking face' (Bavelas; Chovil, 2018), as if searching for the word in another cognitive space. In line 13, when looking to the left, he uses the term *stations*, probably as a transfer from the Spanish *estaciones* to *seasons*, but immediately presents a self-initiated self-repair: =like the 'SEAsons.' ↓SORry- (line 14). Here, the turn constructional unit (TCU) is momentarily suspended (Hayashi; Raymond; Sidnell, 2013, p. 13). Self-repairs are highly relevant techniques in intercultural communication (Liu; Kinginger, 2021; Cogo; House, 2018).

The body symbols that accompany the behavior of the Brazilian interlocutor (Br) also indicate, at first glance, a high level of engagement in the co-construction of meaning and affiliation. In line 02, the blinking of the eye's signals concentration and willingness, while in lines 05 and 16 he nods. What is most interesting is that the symbols in lines 09 and 11 in a way allow access to the cognitive plane and its active co-construction in the search for the correct lexeme. In line 09, the Brazilian stops his rhythmic accompaniment of the conversation to also access his own memory space when his eyes move from the top to the left and back up again, visualized by the icon. Later, in line 11, he nods his head to the left and closes his eyes when he says *yeah to FACE*; . It is precisely in this context that the two smiles with slightly open mouth are shown by the Colombian, perhaps as a sign that he is grateful for receiving or having received support from his interlocutor (lines 08, 10 and 12).

Therefore, the iconic symbols of ISWA, integrated here in the category 'head and face', in an even more simplified way than proposed by Dix (2022), help the researcher and the reader to identify, already at a first moment, to what extent a specific sequence of an interaction becomes relevant and significant for the analysis, by providing clues to the direction that could be explored further.

ViMELF (Brunner; Diemer, 2021)

The development of a notation system for talk-in-interaction in the ViMELF corpus (*Video-mediated English as a Lingua Franca Conversations*, 2018) has highlighted the relevance of transcribing salient non-verbal elements (NVEs), which include gestures, facial expressions, gazes, physical posture, as well as camera changes and background

events, in a brief and structured manner, as emphasized by Brunner and Diemer (2021). The authors' main goal is to provide a robust and concise transcription model designed to efficiently integrate multimodal data into the ViMELF corpus. This corpus comprises informal conversations recorded between participants of various nationalities and unknown to each other, engaged in online and synchronous interactions mediated by English as a lingua franca. Furthermore, the authors hoped that this model will be applicable in a variety of fields; without requiring prior knowledge of gesture research. While acknowledging the existence of a certain subjectivity in defining the criteria of salience and conciseness, the proposed approach offers flexibility to analyze a wide range of multimodal data and allows for both quantitative and qualitative investigations of conversational interactions.

On the other hand, Brunner and Diemer (2021) point out that, despite the considerable history and influence of gesture studies in advancing research with multimodal data, much of the work carried out so far has focused on describing gestures and has therefore neglected the actual transcription of this data. This initial dilemma in describing gestures is mainly attributed to the complexity of notations, which often rely on visual elements to classify gestures, making quantitative analysis difficult. In addition, multi-level and tabular transcription (Goodwin, 2007; Mondada, 2018; Due; Lange, 2021) represents a considerable challenge, both because of its detailed and descriptive nature, and because of the complexity involved in its interpretation. Furthermore, the authors point out that approaches centered on interactional dynamics imply a significant degree of interpretation on the part of the transcribers.

In contrast to CA transcription, whose applicability in quantitative research is limited, the system proposed by Brunner and Diemer (2021) integrates both gestural annotation and other elements such as pauses, hesitation markers, breathing, etc. into the overall text of the transcription, instead of creating a separate layer for non-verbal elements.

The transcription of salient nonverbal elements (NVEs) follows a specific pattern, where actions are placed between curved braces `{ }` with the verb(s) conjugated in the third person singular, always in the present tense, such as `{laughs}`, not `{laughing}`. For consecutive NVEs, the notation is sequential, `{raises head}`, `{rolls eyes}`, while simultaneous occurrences are indicated by the `'&'` symbol `{raises eyebrows & smiles}`. These notations are not related to a specific intonational content, nor are they marked in relation to the speed, intensity or duration of the execution, which is a weak point of this proposal. In addition, it is possible to extend the scope of the transcription to include (i) direct object, for verbs that do not implicitly carry the object (`raises the hand`, `raises the arm`), unlike `{nods}` which includes the 'head' domain; (ii) adverb (three times, repeatedly); (iii) prepositions (up, to the left, with the finger).

In contrast to CA transcription, whose applicability in quantitative research is limited, the system proposed by Brunner and Diemer (2021) integrates both gestural annotation and other elements, such as pauses, hesitation markers, breathing, etc.,

into the main body of the transcription text rather than creating a separate layer for nonverbal elements., (iv) the suffix *-ing* to indicate additional modifications such as movements made by the participants and, finally, if there is a need to include more details, (v) actions can be separated by semicolons {raises hands; palms out}, {raises hand; palm up}, always with the principle of conciseness in mind, so that the transcriptions remain as legible as possible.

Next, we will exemplify the notation and reading of some salient elements in the interaction to illustrate our adaptation of Brunner and Diemer's (2021) system for the multimodal transcription of video-mediated talk-in-interaction:

Sequence 1: *To face the winter* 2019SkyBrCo01 ((07:13-07:36))

01 Co: (--) [so i feel] like there"s a lot of
(.) ↓CULTural 'difference, {looks up to the right }

02 Br: [{blinks, nods}]

03 Co: (-) between (--) how pEople !↓AC!ting with 'their tIme?
{executes manual gesture & looks at the screen}
{leans the head forward & looks up to the right }

04 that WE [do because temporary I WASN"t prepA:red;]
{looks down to the right} {looks at the screen}

05 Br: [{nods}]

06 Co: (--) to (--) con!FRONT!.

07 (--) con'FRONT? {looks up to the left}

08 Co: (--) like (.) to [to ↑FACE] the wInter? {looks at the screen
& smiles}

09 Br: [to;]

10 {rolls eyes upward and to the left, and then back
again}

11 Co: {smiles}

12 Br: yeah to FACE; {leans head forward & closes the eyes}

13 Co: to 'FACE the wInter because; {sorri & olha à esquerda}

14 Co: (-) like in colOmbia we don"t have the STAtions;=
{smiles & looks to the left}

15 =like the `SEASons.=↓SORry- {waves & looks to the left}

16 (-) we don"t have SEASons [at All (.)] sO;;
{looks at the screen}

17 Br: [{nods}]

In lines 01 and 02, we can observe an overlap of the participants' turns, where the Colombian's intonation unit begins so i feel like there's a lot of (.) ↓CULTural´difference (line 01), while the Brazilian makes subtle facial gestures, such as winks and slight nods (line 02). These gestures by the Brazilian occur simultaneously with his gaze directed at the screen and suggest a high degree of involvement in the construction of interactional alignment and the promotion of social affiliation (Stivers, 2008; Fidelis De Paula, 2023). This overlapping of multimodal turns is an adaptation we have developed which differs from the system proposed by Brunner and Diemer (2021), which only advocates the inclusion of ENV, without specifying the temporal or contextual relationship between gestures and verbal content. However, we believe that the association of gestures with specific elements of conversation can indicate different strategies for constructing meaning in interaction. Furthermore, the analysis of these simultaneous occurrences can provide insights into the relationship between facial gestures and narrative and sequential aspects, both in the IMV context and in the participants' use of ELF.

In lines 06 and 07, when looking to the left, the Colombian uses the verb *to confront*, probably as a transfer from the Spanish *confrontar* to *face*, but then makes a self-initiated self-repair: (--) like (.) to [to ↑FACE] the wInter? (line 08), at the same time as he looks straight ahead, possibly towards the camera, and smiles. This suggests that, after accessing the cognitive space to carry out this lexical search, he returns to the interactional environment, looking at the screen and addressing the Brazilian, at the same time as showing a smile. This behavior indicates that the Colombian alternates between the cognitive space, where he performs mental operations, and the interactional space of the screen, where he makes conversational adjustments and signals engagement with the Brazilian. The smile as he looks straight ahead and possibly at the camera seems to mark this transition back to virtual and synchronous interaction (Fidelis De Paula, 2023).

The system proposed by Brunner and Diemer (2021) allows for the concise transcription of nonverbal data (NVE), simplifying the reading of transcripts and providing immediate clues as to the relevance of these interactional elements for analysis. The taxonomy resulting from this model is based on the principles of salience and conciseness, thus constituting an organized, descriptive and comprehensive annotation system. We believe that this approach represents a balanced solution, since it allows researchers to structure complex multimodal data and contribute to the development of rich data corpora and a wide range of applications.

Discussion

A first issue that emerges when dealing with the challenges of transcription in video-mediated interactions (VMI) concerns understanding and delimiting the salience of the gestural-bodily elements that should be transcribed. This task leads us to reflect


particularly on the dynamic and multifaceted nature of facial gestures in VMI in which some gestures, such as movement and addressing the gaze, can be difficult to establish with precision. It is not always possible to determine when a gaze movement becomes a relevant gesture for the interaction, given the extreme brevity of its execution. Furthermore, this speed often makes it difficult to relate facial gestures to specific parts of speech. Next, we discuss how the transcription models proposed by Brunner and Diemer (2021) and Dix (2022) were applied to analyze gaze direction and addressing, and we highlight the analytical aspects related to the notation of facial gestures.


Sequence 1: *To face the winter* 2019SkyBrCo01 ((07:13-07:36))



Transcription output according to Brunner e Diemer (2021)


- 01 Co: (--) [so i feel] like there"s a lot of
(.) ↓CULTural ´difference, {look up to the right}
- 02 Br: [{blinks, nods}]
- 03 Co: (-) between (--) how pEople !↓AC!ting with ´their tIme?
{executes manual gesture & looks at the screen}
{leans head forward & looks up to the right}
- 04 that WE [do because temporary I WASN"´T prepA:red;]
{looks down to the left} {looks at the screen}
- 05 Br: [{nods}]
- 06 Co: (--) to (--) con!FRONT!.
- 07 (--) con´FRONT? {looks up to the left}


Transcription output according to Dix (2022)

01 Co: (--) so i feel like there's a lot of (.) !CULTural 'difference,
 HF 


02 Br:
 HF  ...

03 Co: (-) between (--) how pEople !AC!ting with 'their tIme?
 HF  

04 that WE do because temporary I WASN'T prepA:red;
 HF 

05 Br:
 HF 

06 (--) to (--) con!FRONT!.>
 HF

07 (--) con'FRONT?>
 HF 

Dix's system (2022) reduces movement phrases while seeking to represent a precise and lucid mode for the type of action. We have seen that Dix's system (2022) has an advantage in terms of introducing universality to the proposed icons, which is an aspect of growing relevance to VMI transcription in view of its increasingly multilingual and intercultural nature in a global world, particularly in this interactional field. What's more, as our example illustrates, the transcriber can immediately visually see where intense, multimodal moments of co-construction of meaning, mutual understanding and possible inconsistencies occur in the interaction. Therefore, simplifying transcription by adopting iconic symbols offers an accessible and efficient approach to identifying facial gestures in a reduced and more concise version, aimed at VMI, as proposed here. This simplification not only speeds up the transcription process, but also helps the transcriber develop a clearer understanding of the salience criteria of facial gestures. Gestures are inserted exactly at the point where they begin and their duration is marked with punctuation marks, such as periods (.), indicating the prolongation of the gestural execution in relation to the speech. However, a negative point in Dix's system is not only the enormous effort that a transcription requires, including training to become fluent in transcription, but also the fact that the reader of a transcription is often unfamiliar with the symbols.

Brunner and Diemer's (2021) proposal seeks to offer a standardized transcription that integrates not only clear semantic guidelines, but also considers the integral movement itself more precisely than icons. In this way, all elements can be documented. We have seen that the Colombian's behavior suggests a tendency to look up to the right or left at various times. However, the act of looking straight ahead (lines 03 and 04), noted exclusively in Brunner and Diemer's model (2021), reintroduces the Colombian into the context of virtual interaction and places him back in social alignment and for the organization of turn-taking in this environment (Fidelis De Paula, 2023). This

default case is not provided for in the more static Dix icons. This default case is not provided for in Dix's more static icons. Similarly, the Dix system does not allow the Colombian's full gaze to be annotated in L03. In other words, the object of interest, the hand gesture itself, is outside the focus of the icon because it does not show the relationship between the gaze and the hand gesture. Given that the pointing of the gaze seems to take on a more self-referential than interactive function in this context, the impossibility of recording this dynamic can restrict the researcher's understanding of the construction of meanings in video-mediated interaction. Therefore, a detailed and context-sensitive analysis is fundamental to understanding the dynamics of facial gestures and gaze addressing and their contribution to communicative interaction.

The formalization of nonverbal elements additionally allows those looking to apply quantitative research to then have a solid basis for doing so. Brunner and Diemer offer a rudimentary taxonomy based on simplified syntax so that searchable multimodal corpora can be created. This possibility does not yet exist regarding icons, although there are early projects to integrate such systems into transcription programs such as EXMARaLDA (Schmidt; Wörner, 2009). This more detailed and flexible approach can be particularly useful when it comes to understanding important nuances and complexities of conversational interactions. In addition, the concise and standardized syntax of this system seems to be the best way so far to retrieve metadata and search rich corpus data and is compatible with software that has database creation systems and search tools, such as Exakt (Schmidt; Wörner, 2014). This standardized structure is also favorable for integration with Artificial Intelligence (AI) tools that perform automatic transcription (<https://exmaralda.org/en/>).

However, the Colombian's observation of the direction and pointing of his gaze suggests a self-referential process, where he performs the gesture more to monitor and process his own speech than to explicitly direct it to the interlocutor. This aspect points to the need for a more flexible transcription system, as proposed by Brunner and Diemer (2021), which allows detailed annotation of this relationship between modalities. This could include changes in camera angle and movement; background events (sounds, people and/or animals visible in the interaction, interrupting or interacting with one of the speakers, etc.). This approach may be particularly relevant for analyzing interactions where facial gesture serves to both express and internalize meaning during video-mediated communication.

Concluding remarks

Having started in the last few decades, research into multimodal data still has many limitations when it comes to analyzing facial gestures. This article aims to present methodological and analysis tools for transcribing facial gestures in VMI based on the transcription models proposed by Dix (2022) and Brunner and Diemer (2021). We chose these two proposals for two main reasons: (a) they are part of the current debate

on the complexity and problems related to multimodal transcription in the context of conversation analysis, interactional linguistics and gesture studies; and (b) both deal with issues relevant to video-mediated interactions (VMI), since one approach offers universal iconic tools for conversations in lingua franca, with a detailed inventory for facial gestures, while the other, presenting an almost opposite proposal, was developed based on VMI in ELF, seeking a concise syntax as a basis for quantitative research.

We observed that while Brunner and Diemer's (2021) system allows for a more detailed and comprehensive notation of body, hand and facial gestures, gazes and the situational context of the interaction, Dix's (2022) model, based on the insertion of iconic symbols adapted from ISWA for the head and face, provides a concise transcription, highlighting the most prominent facial gestures in the interactional sequences. Our discussion points out the challenges and limitations in both systems, since Dix (2022) requires training to become familiar with the icons and there is, to date, no possibility of automatically inserting the symbols into transcription software, such as EXMARaLDA, ELAN, and CLAN. Brunner and Diemer's (2021) model does not provide for temporal marking between oral and gestural production. To remedy this shortcoming in Brunner and Diemer (2021), we proposed adaptations to the notation of overlapping participant actions, both for speech and gestures, nods, glances and smiles, making it possible to associate these phenomena with a specific moment of speech.

We understand that with the expansion of multimodal corpora, including those that bring a variety of cultural and linguistic contexts, it is essential to develop hybrid systems, integrating the visual clarity of Dix's icons with the detailed annotation of Brunner and Diemer. The development of these methodological and analytical tools can contribute to the development of future research on investigating the self-referential function of facial gestures in video-mediated interactions, the relationship between gaze direction and pointing with hand gestures, as well as understanding how these gestures are used by interlocutors to monitor and process their own speech.

Acknowledgements and REC number

Ulrike Schröder would like to thank FAPEMIG for funding the FAPEMIG Universal project (2024-2026), CAPES-DAAD for funding PROBRAL (2023-2026) and CNPq for the Productivity Grant (2022-2025). Flavia Fidelis de Paula would like to thank CAPES for the financial support through the CAPES/PROBAL project, which enabled her to carry out her postdoctoral research at the University of Potsdam in Germany from November 2023 to August 2024, as well as FAPEMIG for the BDCTI Level I grant (2024-2025) under the FAPEMIG Universal project. The project "The multimodal coordination of intercultural interaction mediated by video" was approved by the Research Ethics Committee under process number 55218521.1.1001.5149.

- **RESUMO:** *A pandemia restringiu as interações presenciais, enquanto acelerou a globalização das conexões humanas por meio de videochamadas. Isso possibilitou a realização em larga escala de atividades como trabalho colaborativo, conferências, ensino, consultas médicas, ritos religiosos e encontros familiares. Essas práticas não apenas promoveram a expansão das fronteiras locais, mas também conectaram o mundo globalmente, o que fortaleceu encontros interculturais caracterizados cada vez mais por conversas multilíngues. Nesta nova configuração, as “cabeças falantes” (talking heads) assumem uma importância central. O presente trabalho apresenta e compara dois modelos metodológicos para a transcrição de gestos faciais em interações mediadas por vídeo, conforme propostos por Brunner e Diemer (2021) e Dix (2022). Escolhemos estas duas propostas atuais dentro de uma discussão sobre transcrição refinada no campo da Análise da Conversa Multimodal, como uma primeira tentativa de testar a aplicabilidade desses modelos em interações mediadas por vídeo. Nosso estudo ressalta a necessidade de desenvolver metodologias que capturem a complexidade multimodal dos gestos faciais nessas interações, ao mesmo tempo que demonstra como esses sistemas podem contribuir para métodos de pesquisa quantitativos e qualitativos que integram dimensões linguísticas e gestuais. Além disso, discutiremos as vantagens e desvantagens dos dois sistemas e faremos propostas para seu aperfeiçoamento.*
- **PALAVRAS-CHAVE:** *Transcrição multimodal; Interação mediada por vídeo (IMV); Inglês como língua franca, Comunicação intercultural.*

Author Contributions (according to the CRediT taxonomy)

Ulrike Schröder: Conceptualization, Data curation, Data analysis, Funding acquisition, Investigation, Methodology, Project administration, Writing – original draft, Writing – review & editing

Flavia Fidelis de Paula: Conceptualization, Data analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing

Data Availability Statement

All datasets supporting the findings of this study have been published within the article itself.

REFERENCES

- BALAMAN, U.; DOEHLER, S. Navigating the complex social ecology of screen-based activity in video-mediated interaction. **Pragmatics**, v. 32, n. 1, p. 54-79, 2022.
- BAVELAS, J. Face-to-face dialogue as a micro-social context. *In*: DUNCAN, S. D.; CASSELL, J.; LEVY, E. T. (org.). **Gesture and the dynamic dimension of language**. Amsterdam, The Netherlands: Benjamins, 2007. p. 127-146.
- BAVELAS, J. **Face-to-face dialogue**: Theory, research, and applications. Oxford University Press, 2022.
- BAVELAS, J.; CHOVIL, N. **Nonverbal and verbal communication**: Hand gestures and facial displays as part of language use in face-to-face dialogue, 2006.
- BAVELAS, J.; CHOVIL, N. Some pragmatic functions of conversational facial gestures. **Gesture**, v. 17, n. 1, p. 98-127, 2018.
- BAVELAS, J.; GERWING, J.; HEALING, S. Hand and facial gestures in conversational interaction. *In*: HOLTGRAVES, T. M. (org.). **The Oxford handbook of language and social psychology**. Oxford: Oxford University Press, 2014. p. 111-130.
- BEZEMER, J.; JEWITT, C. Multimodal analysis: Key issues. *In*: LITOSSALIS, L. (org.). **Research methods in linguistics**. London: Continuum, 2010. p. 180-197.
- BRESSEM, J.; LADEWIG, S.; MÜLLER, C. A linguistic annotation system for gesture. *In*: MÜLLER, C.; CIENKI, A.; FRICKE, E.; LADEWIG, S. H.; MCNEILL, D.; TEßENDORF, S. (orgs.). **Body – language – communication**. An international handbook on multimodality in human interaction. Volume 1, Berlin, Boston: De Gruyter Mouton, 2013. p. 483-501.
- BRUNNER, M.; DIEMER, S. Multimodal meaning making: The annotation of nonverbal elements in multimodal corpus transcription. **Research in Corpus Linguistics**, v. 9, n. 1, p. 63-88, 2021.
- CARNAVAL, M.; MIRANDA, L. S.; MORAES, J. A.; RILLIARD, A. Funções dos gestos faciais na prosódia audiovisual. *In*: AVELAR, M.; PACHECO, V.; OLIVEIRA, M. (org.). **Linguística e Estudos de Gestos**: Interfaces. Campinas: Pontes Editores, 2023. p. 11-49.

CHOVIL, N. Facing others: A social communicative perspective on facial displays. *In: RUSSELL, J. A.; FERNÁNDEZ-DOLS, J. M. (org.). **The psychology of facial expression**. Cambridge: Cambridge University Press, 1997. p. 321-333.*

CHOVIL, N. Measuring conversational facial displays. *In: MANUSOV, V. L. (org.). **Sourcebook of nonverbal measures: Going beyond words**. Mahwah, NJ: Lawrence Erlbaum Associates, 2005. p. 173-188.*

COGO, A.; HOUSE, J. The pragmatics of ELF. *In: **The Routledge handbook of English as a lingua franca**. In: JENKINS, J.; BAKER, W.; DEWEY, M. (org.). London: Routledge, 2018. p. 210-223.*

COUPER-KUHLEN, E.; SELTING, M. Towards an interactional perspective on prosody and a prosodic perspective on interaction. *In: COUPER-KUHLEN, E.; SELTING, M. (org.). **Prosody in conversation**. Cambridge: Cambridge University Press, 1996. p. 11-56.*

COWAN, K. Multimodal transcription of video: Examining interaction in Early Years classrooms. **Classroom Discourse**, v. 5, n. 1, p. 6-21, 2014.

CRUZ, F. M. D.; OSTERMANN, A. C.; ANDRADE, D. N. P.; FREZZA, M. O trabalho técnico-metodológico e analítico com dados interacionais audiovisuais: a disponibilidade de recursos multimodais nas interações. **DELTA: Documentação de Estudos em Linguística Teórica e Aplicada**, v. 35, 2019. Disponível em: <https://doi.org/10.1590/1678-460X2019350404>. Acesso em: 16 ago. 2024.

DEPPERMAN, A. Multimodal interaction from a conversation analytic perspective. **Journal of Pragmatics**, v. 46, n. 1, p. 1-7, 2013.

DIX, C. GAT2 trifft das International SignWriting Alphabet (ISWA). *In: SCHWARZE, C.; GRAWUNDER, S. (org.). **Transkription und Annotation gesprochener Sprache und multimodaler Interaktion: Konzepte, Probleme, Lösungen**. Tübingen: Narr Francke Attempto, 2022. p. 103-131.*

DUE, B. L.; LANGE, S. B. Body part highlighting: Exploring two types of embodied practices in two sub-types of showing sequences in video-mediated consultations. **Social Interaction. Video-Based Studies of Human Sociality**, v. 3, n. 3, 2021. Disponível em: <https://doi.org/10.7146/si.v3i3.123836>. Acesso em: 24 ago. 2024.

DUE, B.; LICOPPE, C. Video-mediated interaction (VMI): Introduction to a special issue on the multimodal accomplishment of VMI institutional activities. **Social interaction. Video-based studies of human sociality**, 3. Jg., Nr. 3, 2021. Disponível em: <https://doi.org/10.7146/si.v3i3.123836>. Acesso em: 24 ago. 2024.

EKMAN, P.; FRIESEN, W. V. **Facial Action Coding System**: A technique for the measurement of facial movement. Palo Alto: Consulting Psychologists Press, 1978.

FIDELIS DE PAULA, F. **A multimodalidade em aulas síncronas**: um estudo sobre a construção de afiliação e alinhamento no contexto de ensino remoto. 2023. Tese (Doutorado em Estudos Linguísticos) – Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, 2023.

FLEWITT, R.; HAMPEL, R.; HAUCK, M.; LANCASTER, L. What are multimodal data and transcription? In: JEWITT, C. (org.). **The Routledge handbook of multimodal analysis**. New York: Routledge, 2017. p. 44-59.

FRIESEN, N. Telepresence and tele-absence: A phenomenology of the (in) visible alien online. **Phenomenology & Practice**, v. 8, n. 1, p. 17-31, 2014.

GOODWIN, C. Environmentally coupled gestures. In: DUNCEN, S.; CASSELL, J.; LEVY, E. (org.). **Gesture and the dynamic dimensions of language**. Amsterdam, The Netherlands: John Benjamins, 2007. p. 195-212.

GOODWIN, C. Why multimodality? Why co-operative action? **Social Interaction. Video-Based Studies of Human Sociality**, v. 1, n. 2, 2018. DOI: <https://doi.org/10.7146/si.v1i2.110039>.

GOODWIN, M.; GOODWIN, C. Emotion within situated activity. **Communication: An arena of development**, v. 33, 2000. p. 33-54.

GROß, A.; DIX, C.; RUUSUVUORI, J.; PERÄKYLÄ, A. Facial gestures in social interaction: Introduction to the special issue. **Social Interaction. Video-Based Studies of Human Sociality**, v. 6, p. 2-11, 2023. DOI: <https://doi.org/10.7146/si.v6i3.142894>.

HAYASHI, M.; RAYMOND, G.; SIDNELL, J. (org.). **Conversational repair and human understanding**. Cambridge: Cambridge University Press, 2013.

HEATH, C.; HINDMARSH, J.; LUFF, P. **Video in qualitative research**. London: Sage, 2010.

HJULSTAD, J. Practices of organizing built space in videoconference-mediated interactions. **Research on Language and Social Interaction**, v. 49, n. 4, p. 325-341, 2016.

JEFFERSON, G. Glossary of transcript symbols with an introduction. *In*: LERNER, G. H. (org.). **Conversation analysis**. Studies from the first generation. Amsterdam, Philadelphia: John Benjamins, 2004. p. 13-31.

KAUKOMAA, T.; PERÄKYLÄ, A.; RUUSUVUORI, J. Turn-opening smiles: Facial expression constructing emotional transition in conversation. **Journal of Pragmatics**, v. 55, p. 21-42, 2013.

KENDON, A. Some functions of gaze-direction in social interaction. **Acta Psychologica**, v. 26, n. 1, p. 22-63, 1967. DOI: [https://doi.org/10.1016/0001-6918\(67\)90005-4](https://doi.org/10.1016/0001-6918(67)90005-4)

KENDON, A. **Gesture**: Visible action as utterance. Cambridge: Cambridge University Press, 2004.

LICOPPE, C.; MOREL, J. Video-in-interaction: “Talking heads” and the multimodal organization of mobile and Skype video calls. **Research on Language & Social Interaction**, v. 45, n. 4, p. 399-429, 2012.

LIU, S.; KINGINGER, C. The sociocultural ontogenesis of international students’ use of pragmatic strategies in ELF academic communication: Two contrasting case studies. **Journal of Pragmatics**, v. 186, p. 364-381, 2021.

LOENHOFF, J. Interactive technologies and the function of the senses. *In*: NORRIS, S. (org.). **Multimodality in practice**. Investigating theory-in-practice-through-methodology. New York, London: Routledge, 2012. p. 20-35.

LYONS, A. Multimodality. *In*: ZHU, H. (org.). **Research Methods in Intercultural Communication**: A Practical Guide. Malden, Oxford: John Wiley & Sons, 2016. p. 268-280.

MANSTEAD, A. S.; LEA, M.; GOH, J. Facing the future: Emotion communication and the presence of others in the age of video-mediated communication. *In*: KAPPAS, A.; KRÄMER, N. C. (org.). **Face-to-face communication over the internet**. Emotions in a web of culture, language and technology. Cambridge: Cambridge University Press, 2011. p. 144-175.

MONDADA, L. Conversation analysis: Talk and bodily resources for the organization of social interaction. *In*: MÜLLER, C.; CIENKI, A.; FRICKE, E.; LADEWIG, S. H.; MCNEILL, D.; TEßENDORF, S. (org.). **Body – language – communication**. An international handbook on multimodality in human interaction. Volume 1, Berlin, Boston: De Gruyter Mouton, 2013. p. 218-227.

MONDADA, L. Contemporary issues in conversation analysis: Embodiment and materiality, multimodality and multisensoriality in social interaction. **Journal of Pragmatics**, v. 145, p. 47-62, 2019. DOI: <https://doi.org/10.1016/j.pragma.2019.01.016>.

MONDADA, L. Transcription in linguistics. *In*: LITOSSELITI, L. (org.). **Research methods in linguistics**. London: Bloombury, 2018. p. 85-114.

NEVILE, M. The embodied turn in research on language and social interaction. **Research on Language and Social Interaction**, v. 48, n. 2, p. 121-151, 2015.

NORRIS, S.; PRINI, J. Communicating knowledge, getting attention, and negotiating disagreement via video conferencing technology: A multimodal analysis. **Journal of Organizational Knowledge Communication**, v. 3, n. 1, p. 23-48, 2016.

SALOMÃO, A. C. B.; FREIRE JUNIOR, J. C. (org.). **Perspectivas de Internacionalização em casa: intercâmbio virtual por meio do Programa BRaVE / UNESP**. São Paulo: Cultura Acadêmica, 2020.

SCHMIDT, T.; WÖRNER, K. EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research. **Pragmatics**, v. 19, p. 565-582, 2009.

SCHMIDT, T.; WÖRNER, K. EXMARaLDA. *In*: MÜLLER, C.; CIENKI, A.; FRICKE, E.; LADEWIG, S. H.; MCNEILL, D.; TEßENDORF, S. (ed.). **Body – Language – Communication: An International Handbook on Multimodality in Human Interaction**. Berlin/Boston: De Gruyter Mouton, 2014. v. 1, p. 402-418.

SCHRÖDER, U. **Co-constructing intercultural space: An embodied approach** [Mouton Series in Pragmatics]. Berlin, New York: De Gruyter, 2025.

SCHRÖDER, U.; NASCIMENTO, T. S. N.; SILVA, A. R. A. Reflexões metodológicas sobre transcrição e a escolha de GAT 2 como sistema de transcrição para o NUCOI. *In*: SCHRÖDER, U.; CARNEIRO MENDES, M. (org.). **Comunicação (inter)cultural em interação**. Belo Horizonte: Editora UFMG, 2019. p. 115-153.

SCHRÖDER, U.; CARNEIRO MENDES, M.; PIRES, C. C.; ALVES DA SILVA; D. H.; NASCIMENTO, T. C.; FIDELIS, F. P. com revisão técnica de P. C. GAGO (UFJF/UFRJ). GAT 2. Um sistema para transcrever a fala-em-interação, traduzido e adaptado do original da Selting, Margret *et al.* **Veredas**, v. 20, n. 2, p. 6-61, 2016.

STIVERS, T. Stance, alignment, and affiliation during storytelling: When nodding is a token of affiliation. **Research on Language and Social Interaction**, v. 41, n. 1, p. 31-57, 2008.

STIVERS, T.; TIMMERMAN, S. Always look on the bright side of life: making bad news bivalent. **Research on Language and Social Interaction**, v. 50, n. 4, p. 404-418, 2017.

STREECK, J. The emancipation of gestures. **Interactional Linguistics**, v. 1, n. 1, p. 90-122, 2021.

SUTTON, V. **The SignWriting Alphabet**. Read and Write any Sign Language in the world. ISWA Manual 2010. The SignWriting Press. Disponível em: <http://www.movementwriting.org/symbolbank/>. Acesso em: 17 ago. 2024.

Received on September 7, 2024

Approved on October 18, 2024

Appendix

Summary of the transcription conventions of GAT 2 as adapted by the ICMI Research Center

| SEQUENTIAL STRUCTURE | |
|--|---|
| . | falling to low final pitch movement of IU |
| ^SO | rising-falling accent pitch movement |
| ~SO | falling-rising accent pitch movement |
| ´SO | rising accent pitch movement |
| `SO | falling accent pitch movement |
| ↑ | pitch upstep |
| ↓ | pitch downstep |
| <<l>> | lower pitch register |
| <<h>> | higher pitch register |
| LOUDNESS AND TEMPO CHANGES | |
| <<f>> | forte, loud |
| <<p>> | piano, soft |
| <<all>> | allegro, fast |
| <<len>> | lento, slow |
| <<cresc>> | crescendo, increasingly louder |
| <<dim>> | diminuendo, increasingly softer |
| <<acc>> | accelerando, increasingly faster |
| <<rall>> | rallentando, increasingly slower |
| CHANGES IN VOICE QUALITY | |
| <<creaky>> | glottalized |
| <<whispery>> | change in voice quality as stated |
| DESCRIPTION OF FURTHER VOCAL AND NON-VOCAL ACTIONS AND OTHER CONVENTIONS | |
| ((laughs)) | non-verbal vocal actions and events |
| <<crying>> | non-verbal vocal actions and events with scope of accompanying speech |
| (may i) | assumed wording |
| (i say/let"s say) | possible alternatives |

| | |
|---|---|
| (xxx xxx) | two unintelligible syllables |
| SEQUENTIAL STRUCTURE | |
| . | falling to low final pitch movement of IU |
| ^SO | rising-falling accent pitch movement |
| ~SO | falling-rising accent pitch movement |
| ´SO | rising accent pitch movement |
| `SO | falling accent pitch movement |
| ↑ | pitch upstep |
| ↓ | pitch downstep |
| <<l>> | lower pitch register |
| <<h>> | higher pitch register |
| LOUDNESS AND TEMPO CHANGES | |
| <<f>> | forte, loud |
| <<p>> | piano, soft |
| <<all>> | allegro, fast |
| <<len>> | lento, slow |
| <<cresc>> | crescendo, increasingly louder |
| <<dim>> | diminuendo, increasingly softer |
| <<acc>> | accelerando, increasingly faster |
| <<rall>> | rallentando, increasingly slower |
| CHANGES IN VOICE QUALITY | |
| <<creaky>> | glottalized |
| <<whispery>> | change in voice quality as stated |
| DESCRIPTION OF FURTHER VOCAL AND NON-VOCAL ACTIONS AND OTHER CONVENTIONS | |
| ((laughs)) | non-verbal vocal actions and events |
| <<crying>> | non-verbal vocal actions and events with scope of accompanying speech |
| (may i) | assumed wording |
| (i say/let"s say) | possible alternatives |
| (xxx xxx) | two unintelligible syllables |

Source: Schröder (2025, p. 103-104)