Routledge
Taylor & Francis Group

Research on Language and Social Interaction

ISSN: 0835-1813 (Print) 1532-7973 (Online) Journal homepage: https://www.tandfonline.com/loi/hrls20

Multiple Temporalities of Language and Body in Interaction: Challenges for Transcribing Multimodality

Lorenza Mondada

To cite this article: Lorenza Mondada (2018) Multiple Temporalities of Language and Body in Interaction: Challenges for Transcribing Multimodality, Research on Language and Social Interaction, 51:1, 85-106, DOI: 10.1080/08351813.2018.1413878

To link to this article: https://doi.org/10.1080/08351813.2018.1413878

RESEARCH ON LANGUAGE AND SOCIAL INTERACTION
2018, VOL. 51, NO. 1, 85–106
https://doi.org/10.1080/08351813.2018.1413878

Routledge
Taylor & Francis Group

Multiple Temporalities of Language and Body in Interaction: Challenges for Transcribing Multimodality

Lorenza Mondada[a,b]

[a]Department of Linguistics and Literature, University of Basel, Switzerland; [b]Centre of Excellence on Intersujbectivity in Interaction, University of Helsinki, Finland

ABSTRACT

The article focuses on the principles of multimodal CA, the way they can be operationalized in a transcription system, and the analytical and conceptual consequences of transcription choices. Elaborating on the foundations of multimodal CA and on the basis of video recordings of French and Swiss German encounters, as well as animal interactions, the article discusses classic and contemporary challenges for transcription and analysis, such as beyond gesture and gaze, body arrangements in interactional spaces, larger groups, material environments, mobile settings, silent activities, and animal encounters. It also highlights the diversity of multimodal practices involved: mobilizing occasioned material resources, movements not only of the upper (head, gesture) but also the lower (feet, legs, posterior) parts of the body, haptic contacts touching objects and coparticipants, and camera movements. The precise transcription of relevant details reveals complex arrangements of multimodal resources and gestalts. Their fine-grained, distinct, multiple temporalities constitute the basis of their sequential order—for sequentiality as a fundamental organizational principle of action. Data are in French and Swiss German.

Video recordings are becoming a common way of collecting data in Conversation Analysis (CA) as well as in other disciplines, occasioning increasing multimodal analyses that take into account language, gesture, gaze, and more globally, the entire body. However, a fuller exploitation of the richness of video data could yet be developed; this invites further reflections on what it is to analyze, but also to transcribe, multimodality. This article aims to make explicit the links between the conceptual and analytical foundations of the study of multimodality in CA and possible solutions that can be applied in transcribing video data. More particularly, after positioning multimodal analysis and multimodal transcription within the principles of CA, the article discusses a series of phenomena—starting with simple cospeech gestures and moving on to more complex multimodal gestalts, including both mobile and silent embodied activities—and the ways in which they can be transcribed. Using a system of multimodal conventions that I developed more than a decade ago, the article discusses transcription choices, practical solutions, and their analytical consequences.

Multimodality and its transcription

Although CA was far from being the first approach to make use of film, it features a tradition of analyses of audiovisual data that can be traced back to several important precursors. One crucial precursor was the *Natural History of an Interview* (McQuown, 1971)—a project in which an interdisciplinary team of scholars took part in a detailed study of a filmed session—which inspired

CONTACT Lorenza Mondada lorenza.mondada@unibas.ch Department of Linguistics and Literature, French Studies, University of Basel, Maiengasse 51, CH 4056 Basel, Switzerland.

Color versions of one or more of the figures in the article can be found online at http://www.tandfonline.com/hrls.

a series of important early approaches to embodiment in interaction including kinesics (Birdwhistell, 1970). Even though the first decades of CA's history focused on detailed analysis of audio recordings for studying social interaction, the use of film characterized the discipline very early on. Its chief pioneers were Charles and Marjorie Goodwin, who, in the USA, produced films of everyday life as early as the 1970s (see Goodwin, 1981), and Christian Heath in the UK, who initiated a powerful tradition of studies of institutional and workplace settings with a special focus on medicine (see Heath, 1986). Early studies also involved the founders of CA: Gail Jefferson was fundamental in supporting these early attempts, joining the Goodwins and Heath in data sessions and offering training, for the former in Philadelphia (where she was officially a postdoctoral researcher transcribing data for W. Labov) and for the latter in England (where she occupied various positions). Furthermore, Schegloff and Sacks soon became interested in filmed data too (see Schegloff, 1984; Sacks & Schegloff, 2002). More recently, there has been a real boom in video studies (see the recent collective volumes edited by Depperman, 2013; Rasmussen, Hazel, & Mortensen, 2014; Streeck, Goodwin, & LeBaron, 2011). This has made it possible to speak of a "visual," "embodied," or "multimodal" turn in the discipline (Mondada, 2016a; Nevile, 2015).

What is distinctive about CA's use of video is the careful and precise attention paid to temporally and sequentially organized details of actions that account for how coparticipants orient to each other's conduct and assemble it in meaningful ways, moment by moment. Even where talk has been the central focus of CA, its primary object is not language (see Sacks, 1984) but rather action, for which language is an important, but neither the only nor the essential, resource. This has consequences for the importance of videos in CA, for how they are specifically produced, transcribed, and analyzed. This article focuses on the relation between the principles of CA video analysis and the requirements they imply for transcription and vice versa; that is, it also focuses on the consequences of transcription choices for the CA's understanding of multimodality. These issues are discussed on the basis of conventions that I have developed and which I apply here to a range of phenomena that represent contemporary challenges for the study of multimodality.

Multimodality

Video has allowed scholars to work on a fundamental dimension of human action: its multimodality. Multimodality includes all relevant resources that are mobilized by participants to build and interpret the public intelligibility and accountability of their situated action: grammar, lexicon, prosody, gesture, gaze, body postures, movements, manipulations of artifacts, etc. Multimodal resources are characterized by a series of features. First, they always relate to the organization of action but do not make sense out of it.[1] Second, the notion of multimodality includes linguistic and embodied resources as well, treating them *in principle* in the same way, without supposing a priori the priority of one over the other. Third, multimodal resources refer not only to conventional resources, such as grammar and some types of gesture, but also to situatedly occasioned resources depending on the local characteristics of the ecology of the activity—both enabling and constraining what participants treat as a meaningful resource.[2] Fourth, they are characterized by a specific temporality that combines multiple successive and simultaneous lines of conduct. Fifth, they are combined in various configurations, or multimodal gestalts, depending on the activity, its ecology, and its material constraints (Mondada, 2014a). The indexical constitution and mobilization of resources generates considerable variability in the way human action can be accountably formatted. For instance, in some activities language might play a crucial role, while in others, alternative

[1]This distinguishes how "multimodality" is defined in CA versus other approaches, where it can be seen as concerning the characteristics of texts, semiotic signs (e.g., visual vs. written), and digital interfaces (see Mondada, 2014a, p. 138). For a discussion on multimodality and transcription from a semiotic perspective, see Bezemer and Mavers (2011). My aim here is not to compare different transcription systems, which would require another article (see Ayass, 2015).

[2]For instance, iconic gestures can be "environmentally coupled" (Goodwin, 2007), they often relate to sedimented manual actions (Streeck, 2009), and they can be shaped in specific ways by objects working as prosthetic extensions (Mondada, 2014a).

resources might be privileged. The prioritization of one resource over another is not a matter that can be decided a priori but is an empirical issue that depends on the type of situated activity and how participants format it. Thus, to study the respective calibration of multimodal resources, the diversification of settings and activities to analyze is crucial in order to go beyond face-to-face conversation and explore a variety of multimodal praxeological configurations, including interactions without talk.

Within the last decade, the burgeoning study of multimodality has been tackled from different perspectives in EMCA—a broader field that takes in CA and ethnomethodological (EM) studies of work and workplace studies. Some analyses focus primarily on the organization of specific *settings*, mainly institutional ones, in order to understand complex spatio-material contexts for action including the use of technologies and artifacts (e.g., Broth, 2008; Heath & Luff, 2000). Other approaches tackle multimodality in relation to the organization of *turns, sequences, and actions*, in order to understand how action is made intersubjectively and publicly accountable and intelligible (e.g., Goodwin, 2000; Goodwin & Goodwin, 1987; Heath, 2013; Mondada, 2007), including activities with very little or even no talk (Ivarsson & Greiffenhagen, 2015; Lerner & Zimmerman, 2002). Finally, multimodality has also been integrated into the study of *grammar, syntax, and lexicon* in interaction, in order to expand our view of how language works in an embodied way (e.g., De Stefani, 2010; Keevallik, 2013).

These studies delineate multimodality in a variety of ways. Some focus on a restricted number of multimodal details in a given organizational phenomenon (a linguistic form, a nod, a type of gesture, gaze or face, e.g., Mondada, 2007; Peräkylä & Ruusuvuori, 2009; Rossano, 2012; Stivers, 2008). Others deal with a richer and holistically intertwined array of details constituting methodical practice (Goodwin, 2000; Mondada, 2014c). Both face the problem of how to go beyond single case analyses to produce analyses of "collections"—that is, how to systematize findings—an issue that becomes even more challenging when an increasing diversity of situated multimodal details, constituting complex *multimodal gestalts*, is considered together (Mondada, 2014b, 2016a).

These aspects constitute a challenge for multimodal transcription, which faces the multiple issues of re-presenting complex spatio-material ecologies constraining and making sense of the relevant embodied resources, details of embodied conduct articulated through talk and grammar, and systematic recurrent features as well as highly contingent situated ones.

Multimodal transcripts

Transcribing is a fundamental research practice that relates to the importance of "inscriptions" in science (Latour, 1986), producing "immutable mobiles" that allow scholars both to perform analyses, e.g., to find patterns of order, and to circulate them—thereby providing both the evidence permitting them to demonstrate their claims and results and the original materials allowing them to discuss their analyses. Transcription is used in a variety of disciplines and has been recognized for its theoretical dimension (Ochs, 1979). Nonetheless, it has a specific role in CA, where it provides a textual representation that stabilizes the fleeting flow of talk (Hepburn & Bolden, 2017). Thanks to the precision work of Jefferson (1973, 1985, 2004) being particularly sensitive to the temporal orders of details in talk, transcription in CA emerged early on as a distinctive practice that could respond to the requirements of CA's mentality: the relevance of detail, the notion of order at all points, the importance of the question "why that now?" for participants, the centrality of temporality, and sequentiality.

Transcripts constitute at the same time a (proto)analysis, a representation and annotation, and an embodied practice involving specific forms of professional vision (Goodwin, 1994) and

[3]Because of lack of space, I do not discuss the position that refutes the importance of transcribing multimodality here. Basically, the main argument defended here is that transcribing is indispensable for a fine-grained analytical investigation of temporally ordered details. This also holds true in the face of arguments claiming that video clips would solve all the problems and make multimodal transcription obsolete: even within a—highly desirable—editorial model of scientific articles that include clips in the analytical text, transcripts would still be needed for precise temporal and sequential analysis.

professional listening that are technologically supported.[3] They have the paradoxical properties of inscribing spoken words in a textual form, of spatializing time, and of stabilizing dynamic flows (Bergmann, 1985).[4]

Whereas verbal transcripts constitute a standard procedure with (relatively) shared conventions, multimodal transcripts are still characterized by a variety of practices. Whereas the former are based on (adapted) orthographic conventions specific to our writing culture, which linearize and segment talk in recognizable units (although this can be problematic for phonetic and prosodic analyses), the latter cannot appeal to a similar tradition for the notation of embodied conducts that constitute a continuous and gradual flow of actions. Whereas it is possible to produce a relatively homogeneous, basic transcript for an entire recording of talk, it is almost impossible to do the same for multimodality. Multimodal transcripts remind one that transcribing is always a selective activity (for speech too), depending on the objectives of the analysis, the granularity of the transcript, the private in-progress versus publicly edited status of the version, the recipient-oriented/reader-friendly character of the final version, and so on. Although *selectivity* can vary depending on the analytical focus (e.g., a systematic study zooming in on one detail for the analysis of a collection versus a comprehensive, single-case analysis considering the richest array of details organizing the situated actions), as well as on editorial and rhetorical strategies (e.g., favoring readability vs. technicality), it ultimately depends on the central issue of *relevance*. The relevance of resources is locally achieved and established by the participants themselves in and for their situated action, exploiting and orienting to them as publicly available, meaningful, and providing the accountability of their actions. This constitutes the fundamental *emic* dimension of multimodal details, consistent with the emic view on language, action, and social interaction characteristic of EMCA: The relevance of details is always indexical; it cannot be decided a priori and once and for all.[5]

These features generate a set of requirements for multimodal transcripts. They have to be flexible, relying neither on a canonical set of pregiven forms nor on their a priori hierarchization. They must be able to accommodate a variety of resources, including unique, ad hoc, and locally situated ones, besides more conventional ones. In other words, they must be able to represent the specific temporal trajectories of a diversity of multimodal details, including talk where this is relevant, but also silent embodied action when talk is not the main resource or activity.

The transcription of the body has attracted the attention of various disciplines for a long time: For instance, treatises on gesture have experimented with textual and visual descriptions of hand movements in vivid illustrations for centuries (e.g., Bulwer, 1644); Laban invented a notation for dancers and choreographers (1928); and Birdwhistell proposed a complex notation when founding his approach of kinesics (1970). In the tradition of interactional studies, specific notations[6] were proposed early on by Kendon (1990) for the trajectory of gesture, Goodwin (1981) for gaze, and Heath (1986; see Luff & Heath, 2015; for a recent account) and Streeck (1993) for both. The notations that I have developed over more than a decade build on

[4]Aligning software like ELAN or CLAN allows one to closely align the original recorded data and the transcript but does not fundamentally solve this problem. Because of lack of space, I do not discuss the impact on transcribing here. However, most excerpts have been transcribed using ELAN; ELAN transcripts are fully convertible into the conventions presented here.

[5]This points at the difference between *transcribing*—which depends on a preliminary analysis of what is made locally relevant by the participants, which in turn depends on the way they configure their embodied and verbal actions within its situated ecology—and *coding*. The former relies on descriptions adapted to the specificity of the particular movement targeted and the latter on standard labels selected from a predefined list of labels (a coding scheme) and homogeneously used throughout the corpus.

[6]The need for specific conventions for embodiment relates to limitations associated with the use of verbal conventions for annotating the body. For example, the use of [brackets] for indicating the placement of embodied conduct within talk would be treating it as having the same properties as overlapping turns, which is not the case. The temporality of embodied cues has different affordances and constraints than talk (e.g., extended simultaneity is a general feature of embodiment but a problematic feature for talk). The use of ((double parentheses)) for embodied conducts is also problematic because it reduces them to comments that are inserted in the flow of talk, ignoring their precise temporal location in the ongoing action and their specific temporal trajectory.

[7]See the website indicated at the end of this article. Although I have developed and used the system since the beginning of the 2000s (e.g., Mondada, 2007) and the convention is increasingly being used by other scholars, there is no publication as yet that sets out its specific rationale and principles of use. This article fills that gap.

these traditions. They are integrated in a convention[7] that is explicit, robust, systematic, and flexible. Moreover, this convention is characterized by the fact that it is not limited to gesture or particular types of phenomena but is adaptable to multiple and diverse types of embodied conducts and particularly careful in preserving and highlighting their finely detailed temporal orders. This convention is increasingly being used in the EMCA literature; it is not merely a formal notation for a range of phenomena but a way of seeing and discovering their temporal details, identifying ordered multimodal gestalts, and generating systematic analyses (e.g., of collections). In this article I discuss the use and the rationale of this convention. I both illustrate the practical implementation of the notations and discuss their analytical and conceptual implications, with a special focus on multiple temporalities typical of multimodality. The remainder of the article considers these issues through a discussion of several challenges for the analysis and transcription of multimodality. It starts with a relatively simple case, cospeech gesture, to introduce the basic principles of the convention system; it then moves on to complex multimodal gestalts including the entire body, considering not only static but also mobile activities, and integrating issues relating to embodied interactions without any talk. This shows the variety of uses of the convention, as well as their analytical and conceptual consequences.

From cospeech gesture to bodies in action/multimodal gestalts

There is a long tradition of studies of gesture (Kendon, 2004, chapter 3), and some early work on Conversation Analysis converges with its findings (Schegloff, 1984). In particular, gesture studies have shown the precisely timed relationship between movements and talk (Condon, 1971), and cospeech gestures and speech (Kendon, 1980; McNeill, 1992). The study of temporal relations of gesture and talk also crucially depends on how their emergent trajectory is considered, as pinpointed by Kendon's distinctions between a gesture's preparation, stroke, and retraction (1980). In CA, the issue is to show how gestures feature in the organization of social interaction (and not only in speakers' conduct) and how they occur not alone but alongside a range of other embodied resources.

Hence, gestures are an obvious phenomenon to start with when engaging in multimodal transcription. In this section, I use gesture to explicate the fundamental principles of the convention illustrated here. I also show how its transcription—and consequent analysis—can be extended to an array of other resources, and the conceptual consequences this has for temporality and sequentiality in interaction.

The first example is a request sequence, video-recorded in a food shop in France.[8]

Extract 1a (FRO_F_PAR_1007-31-15_oeufs)

```
1  CUS     °ben° j'vais vous prendre des oeu:fs, °et [pi:s,°
           °well° I will take from you some eggs °and [then°
2  SAL                                                [mh:
```

The customer requests some eggs (1), and the salesperson responds positively (2).

The action of the customer features a pointing gesture toward the product, which can be transcribed in the following way.

[8]All extracts were collected during ethnographic fieldwork with the agreement of the participants, who were informed about the use of the video recordings in analyses featuring selected moments of interaction and represented in transcripts involving textual and visual representations. In the textual part of the transcript, the identity of the participants has been anonymized (e.g., the names identifying them are all pseudonyms); however, the participants did authorize the use of unanonymized photographic images (i.e., screen shots that have been neither pixelated nor covered with black stripes over the faces), which are important for the analysis of gaze and other facial expressions.

Extract 1b

```
1  CLI     °ben° j'vais †vou†s+  pr+en#d†re†  d+es +oeu:fs,
           °well° I will    take from you     some eggs
                              +....+points eggs+,,,+
                         †...†turns head-†,,†
   fig                                 #fig.1
```

fig.1

The customer points towards the eggs (Figure 1). In the transcript, his pointing is precisely related to its position in time and in the ongoing turn, thanks to a symbol (+) that delimits not only the beginning and end of the gesture but also its preparation (indicated by the convention ….) and withdrawal (indicated by„„„).[9] The gesture's preparation and withdrawal occur rather quickly, since the "home position" (Sacks & Schegloff, 2002) of the pointing hand is on the backpack's strap on the customer's chest. This detail, visible in the image, and which could also possibly be textually described, shows the importance of the local ecology of the body for the position, shape, and temporality of the hand. This produces a quick pointing gesture. Interestingly, the gesture does not come alone: It is slightly preceded by a head movement toward the same product, which corresponds to gazing at it. This movement is transcribed using the same convention, delimited with another symbol (†). The gesture and the head movement are described ("points eggs," "turns head") in a concise way that can be expanded in the analytic text (cf. *supra*); moreover, their description is completed by a picture, a screen shot (Figure 1). Importantly, the picture is precisely related to the specific instant during the talk (thanks to the symbol #, meaning that Figure 1 refers to what the customer does in the middle of the word "pren#dre" 1) and relative to the trajectories of the other movements. This locates the screenshot in relation to other annotations and shows how various streams of embodied and verbal action are articulated together at a specific point (the image has been chosen to show a moment in which both are oriented to the same direction, and indeed this lasts less than one syllable: "+en#d†"). Images allow us to produce a *synthetic* view of the multimodal gestalt, which is *analytically* articulated in different, albeit temporally coordinated, lines in the transcript. In this sense, images are not only contributing to the representation of the movements textually described in the transcript but also to their holistic composition and the ecology in which they happen (their position relatively to relevant objects, the interacting bodies, the environment and its materiality, etc.). In this sense, images are a powerful and indispensable complement to what textual transcription can do.[10]

As shown by this simple initial example, multimodal annotations concern two fundamental aspects of embodied conducts that are not limited to gesture but concern all kinds of movement: (a) their *temporality*, that is their emergent and unfolding trajectory, including preparation and retraction, precisely situated within the turn and the action; and (b) their *shape*, that is what makes the movement recognizable and describable. This latter point raises the practical and analytical question of how to describe these movements, with the aim of *relevantly* capturing what the person is doing and what the coparticipant can see her/him doing, within an *emic* perspective (that is, the perspective of the participants). Within the framework of CA, this description avoids two opposing pitfalls: imputing intentions or cognitive states, and reducing actions to

[9]This article focuses on the conceptual rationale underlying these conventions; for detailed instructions on how to implement them in precisely formatted transcripts, see the web reference to a tutorial at the end of the article.

[10]Transcripts are visual-textual hybrids. The diversity of types of images and their role in transcripts are an important point that cannot be discussed here for lack of space. However, see Mondada (2016b) for an extended discussion on different uses of images in transcripts and their analytical and theoretical consequences.

physiological movements. Practical constraints affect these descriptions: Very limited space within transcripts invites the choice of short lexemes. However, these descriptors can be freely expanded within the analytical text. Moreover, and importantly, these descriptors are present not only in *textual* form but also in *iconic* form: the screenshots. The analyst must decide how to distribute these descriptions between transcript, images, and analytical text. In this sense, glosses in the transcript are shorthand for something that is elaborated elsewhere; in turn, the specific information the transcript provides concerns the temporal details, positions, trajectories, and arrangements of the movements.

Both aspects, time and shape, are necessary to understand what the body is doing. Within this conception of multimodality, the meaning of a movement is not reducible to its *form* but is related to the *moment* in which it is produced; a moment that is meaningful in relation to its sequential environment and its position in ongoing action.

The multimodal annotations in transcript 1b selectively represent only the speaker's turn and a small part of his upper body. It is possible, using the same convention, to annotate a larger array of details where these are made relevant by the participants. For instance, the following expanded transcript allows us to better understand not only the referential act performed by the individual speaker but also the request sequence of the customer addressing the salesperson.

Extract 1c

```
1          ǂ (0.4) #
   cus     †one lateral step twd eggs-->
   cus     >>looks at SAL---->
   sal     >>looks at CUS---->
   fig             #fig.2
```

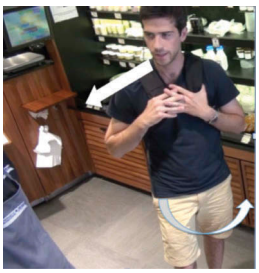


fig.2

```
2  CUS     °ben° j'vais †vou†s+  pr+end†re†ǂ d+es +oeu•:fs,
           °well° I will    take from you    some eggs
                               +....+points eggs+,,+
                     -->†....†turns head†,,†looks in front->
                                        ->ǂpivots twd hard cheese->
   sal                                          ---->•looks eggs->>
3          °et [pi:*†s,° #ǂ
           °and [then°
4  SAL          [Øm*†hØ  #ǂ
                 ØnodsØ
                   *walks tw eggs------->>
   cus           -->†turns head tw eggs-->>
   cus                    ǂwalks tw eggs-->>
   fig                   #fig.3
```

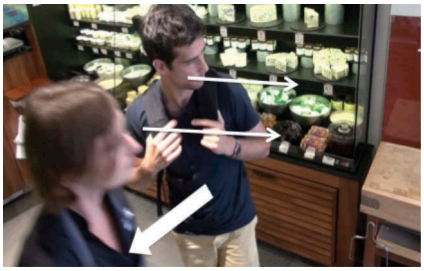


fig.3

Even before the beginning of his request, the customer positions his body in relation to the location of the requested product, by taking a step aside (1). This produces an early projection of the request. This step aside begins at line 1 (1: ‡one lateral step–>) and continues until the middle of the following line (2: –>‡). This notation indicates movements that begin on one line and end at a subsequent one: The arrow (–>) is an instruction to search for the subsequent arrow pointing at the same symbol.[11] This is important because most movements do not begin and end on the same line: They do not coincide with the graphical representation of the action and, most importantly, they do not match with the turn. Even if they are finely coordinated with talk, embodied temporalities have their own trajectory, which is neither synchronous (i.e., beginning and occurring at the same time) nor isochronous (i.e., occurring during the same time span) with it.

So, by taking a lateral step, one part of the customer's body is already oriented toward the product. Another part is oriented toward the salesperson: He looks at her until the beginning of the request (before he utters the second person pronoun *vous* ["you"] 2). Figure 2 shows these two orientations—a form of body torque (Schegloff, 1998)—made visible by several arrows in the figure, which constitute a form of visual annotation enhancing the readability of the transcript.

The salesperson orients to the customer too: She looks at him until the precompletion of his request (2: –>•). In the middle of the form "oeu•:fs"/"eggs" (2), she shifts her gaze from him toward the eggs, displaying a quick understanding of his action. This constitutes the earlier part of a composite multimodal response to the request, characterized by different temporalities: After an early glance toward the eggs (2), she produces an immediate verbal response ("ØmhmØ" 4), co-occurring with a ØnodØ; furthermore, reorienting her body she begins to walk (4), projecting compliance with the request and fetching the product. The first types of response (gaze, nod, and "mhm") are given in turn precompletion position; the second (walking and fetching the product) take more time to be completed. They mobilize different parts of the body, at different moments. Thus, different forms of response mobilize different multimodal resources within different temporalities. Finally, the customer aligns with the salesperson and turns again to the eggs (Figure 3).

One final observation can be made on the basis of this enhanced transcript. The customer displays an early orientation toward not only the currently requested product but also the next one: Even before mentioning the eggs, he already turns away from them and faces a cooler in which hard cheese is displayed. The actual request ends with a rising intonation on the stretched last syllable ("oeu:fs," 2), typical of lists in French, and is followed by "°et [pi:s°"/"°and then°" (3) in a lower voice, partially overlapped by the salesperson. These multimodal elements constitute an early projection of the next request (which will be realized immediately afterwards, by coming back to the cooler and targeting mimolette, a hard cheese located where he was looking). This shows how early embodied projections, which generally largely precede projections in talk, can be precisely documented. Including these preparatory movements radically changes our understanding of the ongoing action. In this case, it also has consequences for action formatting: The customer's action is not formatted as a single request but as a request in-a-series, part of a more global purchase (Mondada & Sorjonen, 2016). Moreover, this shows how action is embedded within the ecology of the activity and more precisely the local geography of the products in the shop—i.e., within the specific spatial distribution of types of products, working as a resource for both participants.

This first example highlights the importance of the precise temporality of participants' movements, demonstrating in particular the phenomenon of early projections; the multimodal annotation of these movements demonstrates their emergent orientations, which often largely precede emergent talk. This is significant for how sequentiality can be better understood thanks to multimodality.

These consequences for the conceptualization of temporality and sequentiality can be further expanded on the basis of another instance of a request for a product, this time in a convenience store in Freiburg/Switzerland where Swiss German is being spoken.

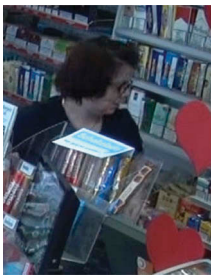
[11]The initial movement is signaled with a ≫, the last one, with a -≫ (4). These double arrows refer to movements beginning/continuing before/after the extract, which is important for their location within broader streams of action and temporal spans.

Extract 2 (KIO_CH_FRI_2-00–37 marylong filterBox)

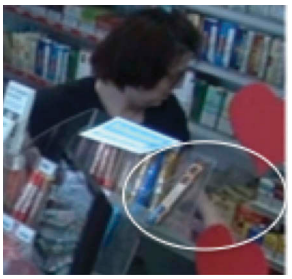
```
1  SAL     #hallo=
   fig     #fig.4
```

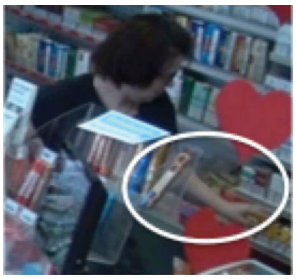
```
                fig.4
2  CUS     +=hallo
           +leans fwd--->
3          (0.3)+
              ->+


4  CUS     +ich *hätte+ gern zwöö#::+ *(0.3)+ #
           I would like to have two::  (0.3)
           +..........+points-------+,,,,,,,+
   sal          *withdraws from counter*turns back-->
```



```
                fig.5                    fig.6
5          marylong, *(1.0)#(0.5) >filter *box.< #
           marylong      (1.5)    >filter  box<
   sal             ->*moves Lhand---------*grasps cig--->>
   fig                    #fig.7                 #fig.8
```





```
          fig.7                fig.8
```

As soon as the salesperson shifts from the previous customer to the new one and greets her (Figure 4), the customer greets the salesperson in return while leaning over the counter (2). This change of position foreshadows the pointing gesture (Figure 5) deployed at the beginning of her request (4). The gesture reaches its full expansion on the numeral (“zwöö::”/“two::” 4) and begins to retract just after it. At this point, the request is recognizable on the basis of the verbal construction projecting the product’s name: The orientation of the body and the pointing indicate that the

product is located behind the salesperson, on the shelf containing cigarettes. This makes the request for cigarettes projectable, and indeed the salesperson begins to turn around at that point (Figure 6). In the remaining part of the customer's turn (who stops pointing, thus adjusting to the salesperson turning around), the customer mentions the brand and the type of packet she wants: This is followed (and reflexively adjusted to) step by step by an increasingly precise hand gesture by the salesperson, first moving toward the type of cigarettes (Figure 7) and then grasping the specific item requested (Figure 8). The ongoing responsive action is shaped in real time by the first action and in turn has a reflexive impact on its design, which can adjust to it in real time (cf. Goodwin, 1979).

The temporality of the first and second actions observed in these two extracts highlights issues concerning the articulation of temporality and sequentiality. These initial responses show that in embodied interaction, sequence organization is often temporally organized in such a way that once a first action is emerging, the second action largely anticipates its completion and begins quite early on. Contrary to the rather linear successivity of adjacency pairs characterizing turns at talk, where the next turn follows the previous one, here we observe an early responsive action of the recipient that largely exceeds what s/he can do in overlapping talk. The anticipation of gestures relative to their lexical affiliate is well known in gesture studies (Kendon, 2004; Schegloff, 1984). Nevertheless, the study of multimodality reveals other forms of anticipation in which an action (here, a turn) can be responded to by multiple embodied, interactionally relevant, and publicly visibly displayed actions oriented to as such, realized in different temporalities and by different parts of the body.

This has consequences for our conception of sequentiality. Multimodal analysis confirms that sequentiality is the fundamental principle of social interaction. But it also reveals that temporal relations between a first action and a second action in response can be much tighter and complex. The fact that responsive next actions emerge as the first one is still unfolding, moment by moment, and reflexively adjusting to it, has consequences for the way we conceptualize the relation between successivity and simultaneity. Even when two (or more) actions unfold more or less at the same time (in a simultaneous way), what is relevant is *when* they begin to emerge (in a successive way, which marks the distinction between initiating and responding actions). Thus, multimodal action is organized in several parallel temporalities, which can be characterized as a plurality of *sequentially ordered simultaneities*. This conception of time in interaction is crucial for considering *complex multimodal gestalts*—i.e., complex methodical arrangements of several resources distributed in time (Mondada, 2014a, 2014b)—that configure action's sequentiality. It is therefore important to represent these temporal arrangements in corresponding multimodal transcripts, which in turn makes analysis of their systematicity possible.

From static to mobile bodies

Having established the principles for multimodal analysis and the corresponding requirements for multimodal transcripts, as well as their implementation in a particular convention, the array of phenomena that can be handled in this way is very wide. This corresponds with current developments in multimodal CA, which has been expanding in at least two directions: (a) by taking into consideration further, previously neglected, multimodal details such as steps and feet position (Broth & Mondada, 2013; Mondada, 2014c); and (b) by considering phenomena of increasing complexity that were previously deemed difficult to handle in the absence of adequate video data, such as mobility (Haddington, Mondada, & Nevile, 2013), multiactivity (Haddington, Keisanen, Mondada, & Nevile, 2015), and materiality, the use of technologies at work (Heath & Luff, 2000; Luff & Heath, 2015), objects in talk-in-interaction (Nevile, Haddington, Heinemann, & Raunionomaa, 2015), writing (Mondada & Svinhufvud, 2016), and typing practices (Luff & Heath, 2015), as well as sensoriality (Mondada, 2016a).

Multimodal details mobilized by participants to organize their actions are potentially infinite because they exceed conventional forms and include situated, ad hoc resources that depend on the type of activity and its specific ecology—for example, the orientation of a pen used for pointing, the hook of a surgeon used as a pointer instead of an operating tool, etc. (Mondada, 2014a, 2016a). Consequently, the multimodal transcription system has to be flexible enough to accommodate all possible kinds of locally made relevant details and at the same time represent them in a similar, coherent, and robust way that is essential for systematic analyses.

Walking is exemplary in this respect. It is a collective and finely coordinated practice (Ryave & Schenkein, 1974) that is sequentially organized, impinging on and reflexively revealing the emergent construction of turns, the organization of sequences, and the changing dynamics of participation (Mondada, 2014c). It also constitutes an interesting challenge for transcription since it involves embodied details previously ignored because of a focus on the upper body, such as steps, feet position, and bodily orientation. It invites us to consider the entire moving body, drawing dynamic multimodal gestalts. Moreover, several bodies moving together, in small groups and more so in larger groups, raises interesting questions about how they build dynamic and changing interactional spaces together (Mondada, 2009), which is crucial for the analysis of participation, and how they position their bodies within the material environment.[12]

The following fragment introduces a discussion on how turn and sequence organization are intertwined with the organization of activities such as standing and walking—involving larger groups of participants—and how to transcribe them. The fragment was video-recorded on a construction site in France, where the architect, Ligour, is presenting the work underway to citizens involved in its collaborative planning. We join the action as he is explaining the importance of having big trees in the park. He occupies the center of the interactional space, with the group looking at him (see Figure 9). To facilitate reading of the transcription, Ligour's movements are highlighted in gray, while collective movements are placed in a boxed text.

Extract 3 (cab21_douves_38_20)

```
1  LIG     eu::h globalement i vaut mieux avoir quelques grands
           ehm:: globally    it's better to have some big
2          arbres# quand-même. (.) hein ça aide un petit peu à pas
           trees nonetheless.  (.) PRT this helps a bit not to
   fig           #fig.9
```

fig.9

[12]For the representation of mobility, cartographic representations complementing other visual representations can be useful. This is especially relevant for longer mobile trajectories, such as those involving cars or bikes (McIlvenny, 2015), but might be difficult to implement for micromobilities, such as small steps.

```
3          s'faire poursuivre ©dans l'*parc quoi.*©
           be pursued in the park          PRT.
                                          *puts weight on L leg*
   cam                          ©moves slightly-----©
4          *(0.2)
           *1 step RF->
5  LIG     .h::::: on* •va# eu:+:*::h a±van© [•#cøer?† #
           .h::::: we'll eh::::m move ⌐on?
                   ->*1 step LF--*
                        •.....•points fwd-->
   ct1                          +looks at the indicated direction--->
   ct2                                 ±looks at indicated dir----->
   ct3                                              †looks ind dir->
   cam                                        ©moves to indic dir--->
   fig                    #fig.10              #fig.11  #fig.12
```

fig.10

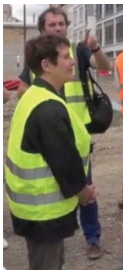


fig.11

fig.12

```
6  LOU                                    [(j'vøoulais)
                                          [(I wanted)
   lig                                         øturns H tw LOU-->>
7          d'mand‡e:[r,* eu]:•:h©• #
           to      as[k,  eh]m::
8  LIG               [ou†±i.] +
                     [yes.]
                        *1 step back--->
                           ->•,,•
   ct4           ‡looks at the indicated direction-->
   ct3                  ->†looks at LOU-->>
   ct2                   ->±looks at LOU-->>
   ct1                          +looks at LOU-->>
   cam                         --->©
   fig                                #fig.13
```

fig.13

fig.14

```
9          (0.2) * (0.3) ‡
   lig       --->*
   ct4                 ->‡
10 LOU     pour les peupliers, ©vous n'avez pa::s© peur (0.5) #
           for the poplars,     you are not   afraid   (0.5)
   fig                                                  fig.14#
   cam                          ©moves, making LOU visible©
```

Ligour's joking explanation is brought to completion syntactically, prosodically, and with the closing particle *quoi* (3). Moreover, on the two last words he begins to move: He first puts his weight on his left leg (3), projecting a step, then steps forward (4) (see his movements in gray). Turn and sequence completion is achieved not only linguistically but also in an embodied way, by his preparing to move forward (Broth & Mondada, 2013).

What comes next is explicitly instructed of the participants: Ligour utters a verbal invitation to move forward (5), while taking a further step. The direction of the walk is indicated by his body stepping (Figure 10) and a pointing gesture (Figure 11). This is responded to in an aligned way by several citizens, who look in the direction he indicates (boxed text with a continuous line, 5, Figure 12).

This movement is aligned to by the movements of the camera too, slightly moving on turn completion (3) and then moving in the indicated direction (5). In the transcript the response of the cameraperson is treated in the same way as the actions of other participants are and transcribed accordingly (cam: ©moves->, 5, Figure 4). Thus, the practice of video recording is both integrated into the event and considered a relevant detail for analysis (Mondada, 2014d, 2016a). The image here (Figure 12, and see also Figure 14) is used not only for documenting a multimodal gestalt but also for showing how the gestalt is framed and how its visual availability is achieved through the work of the cameraperson. While images are generally treated as a *transparent* tool serving multimodal analysis —an open window on human behavior—here they are also treated as having an opaque and reflexive dimension. In this way, images become a *topic for* multimodal analysis rather than just a *resource* (cf. Broth, Laurier, & Mondada, 2014).

At this point, a participant, Lournès, asks a question: She prefaces it ("I wanted to ask ehm::" 6–7) in a hesitant way, manifesting it as a dispreferred action. Ligour responds immediately, turning his head toward her (6), then producing a go-ahead (8) while simultaneously taking a step back (8) and withdrawing his pointing gesture (8) (Figure 13). This multimodal gestalt shows his suspension of the previous action and his realignment with her. It also generates the realignment of the citizens, who were turned in the direction of the instructed walk and who now look back at her (boxed text with a discontinuous line, 8, Figure 13)—with the exception of one participant (ct4) who still looks in the direction initially indicated (boxed text with a continuous line, 8). This shows how people in groups might align but also how others might either delay or resist alignment.

The camera also aligns with this reconfiguration of the participation framework and interactional space, to frame Lournès and Ligour in a more visible way (10, Figure 14).

This excerpt shows how mobility can be treated in multimodal transcripts. Steps are part of a complex multimodal gestalt characterizing turn construction and sequence organization in closing environments within a mobile interactional space, involving the entire bodies of participants. Consideration of the entire body, not only of the speaker but of the coparticipants too, shows the relevance of how several bodies move together, revealing ongoing projections and (dis)aligned coparticipants' responses. Interacting bodies define dynamic interactional spaces that are established, transformed, and dissolved along the activity sequentially emerging and unfolding (Mondada, 2009).

The excerpt also shows how it is possible to make sense of larger groups of participants. Here not all the participants are transcribed (which would be impossible anyway), only a subset (5, 8): This makes it possible to observe how several persons visibly and silently successively align, realign, and disalign with the unfolding action, through the reconfiguration of their trajectory occasioned by the question. Thus, the ongoing action concerns not just the guide and the person asking the question but the entire group. Moreover, the treatment of the "group" as a series of individual "participants" allows us to differentiate their responses, which are not in unison or undifferentiated but are characterized by some persons responding earlier, another later, eventually following the former (looking where they look), or not. Big groups constitute a challenge for detailed multimodal analysis: The way they are represented in transcripts has consequences for their treatment as a homogeneous mass or as individualized entities.

From embodied talk to silent interaction

A further challenge for the study of multimodality, and hence for transcribing it, is the consideration of silent embodied activities. Whereas CA has been developed on the basis of activities in which talk constitutes the main resource, video data allow CA scholars to tackle interactions in which talk might not be the main course of action but are instead organized by a diversity of other, embodied, resources. This raises questions about whether the principles of sequential analysis also hold in the absence of talk, as well as challenges for how to transcribe such interactions. In the last part of the article I discuss these issues on the basis of two sets of video data: workplace activities and primate interactions.

Manual activities of participants at work

Workplace studies were significant in the early development of video analysis (Goodwin & Goodwin, 1996; Heath & Luff, 2000; Luff & Heath, 2015). Workplaces, with their complex spatial configurations, the pervasiveness of technologically mediated activities, and the presence of documents and other artifacts, could only be investigated through video recordings. They are an exemplary setting in which participants often engage in manual activities while interacting with few words. These activities present interesting challenges for multimodal transcription and invite us to take a step further in the elaboration of multimodal annotations that do not depend on talk.

The next fragment was recorded during the preparation of an art exhibition in a museum and follows the work of a team carrying several big and heavy paintings from one room to another. We join the action as the team moves toward a painting that is lying horizontally against the wall (Figure 15) and which Michel proposes to situate vertically (1). The fragment shows how the team carries the painting for the first half of this maneuver, which consists of moving it to the front (Figure 16) then pivoting it by putting down one of its corners on a piece of foam to protect it (Figure 17), in order to reposition it vertically (Figure 18). This simple operation is performed by three coworkers engaged in different coordinated movements.

Extract 4a (ART_Day2_AM_TAT_1-03–36)

```
1  MIC   #on va le redress[er
         we will replace it straight
         >>standing on the right side of the painting->
   bru   >>standing on the left side of the painting->
   guy   >>looking at the foam, on the floor->
   fig   #fig.15
```


```
            fig.15
2  BRU?                    [(   )?
3         (0.5)
4  MIC    %ouais.* vas+-y %sors
          yeah.    go     come out
                *carries->
   bru             +carries->
   guy    %takes foam-----%brings foam under painting->
5         (1.1) # (0.6) +% (0.2) *%+  (2.3)
   mic                        ->*holds-------->
   bru               ->+holds-----+rearranges foam->
   guy               -->%deposits f%
   fig          #fig.16
```

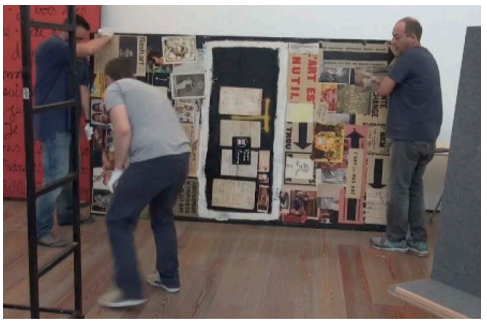


fig.16

```
6   GUY    >pardon<+
           excuse me
    bru           ->+
7          +(0.2) %(0.5)*+ (0.7) # (0.6) % (0.8)* # (3.7)
    bru    +poses--------+pulls------------------------->>
    mic                -->*raises----------------*pushes-->>
    guy             %walks twd MIC---------%helps MIC------>>
    fig                          #fig.17           #fig.18
```

fig.17

fig.18

Michel instructs the next action (1), which is already projected by their disposition around the painting (Figure 15): He is standing on the right, Bruno on the left, while Guy is looking around toward some foam, which is indispensable for laying down any art piece in a safe way. Michel utters a directive (4) while lifting up the painting. Bruno aligns with him immediately, lifting it up too. These two actions are perfectly coordinated (4).

The timed position of the *manual* action, with respect to the *verbal* directive (4), shows that the former is initiated by both men *before* the latter—although it has been generally announced before (1). These temporal details indicate that carrying is not the action made in response to the directive, but rather the coordinated action initiated by Michel lifting up his side of the painting is—a manual action that Bruno can sense *haptically* even before the directive is uttered. Thus, the verbal directive confirms this action, rather than initiating it. This shows that sequence organization can be initiated in an embodied way, even when a first action is verbally produced: What is decisive is the temporality of these actions' emergence. The way this is achieved also points to the difficulty for the transcriber in treating an action that the participants orient to for its (haptic) *sensorial* features, rather than its *visual* aspect documented by the video camera—although the two are not incompatible (Mondada, 2016a).

While Michel and Bruno are engaged in lifting up the painting, Guy bends down to the ground grasping a piece of foam (Figure 16), which he puts underneath the painting. His two colleagues hold

the painting, waiting for the foam to be deposited under its left corner (5). This simple operation is not performed smoothly: As soon as Guy deposits the foam on the ground, Bruno rearranges it with his feet (5), correcting its position on the floor. This delays the next action, occasioning further waiting by Bruno and Michel still holding the painting. Although they do not say a word, Guy notices the problem and apologizes (6).

As soon as the foam is correctly placed on the floor, Bruno positions the left corner of the painting on the foam (7). He and Michel begin the next action simultaneously, in perfectly timed and silent coordination: Michel lifts the painting, and Bruno pulls it (Figures 17–18). Guy also orients to this action by moving toward Michel (7, Figure 3) and helping him (7, Figure 18). The timing of his contribution—joining the action of lifting and pushing the piece, after it has been initiated, without being asked or waited for—builds its accountability as "giving a helping hand" or "assisting" and doing so as an optional facilitating action relatively to the ongoing maneuver. Differentiated temporalities captured in the multimodal transcription (initiating and then responding vs. being simultaneously coordinated vs. joining an already initiated action late) are crucial for interpreting these actions (as a request, a collective movement, an offer of help) and for distinguishing two broader types of actions: those that are sequentially organized versus those that are organized in unison.

The transcript of this short moment shows how silent actions can be annotated by positioning and ordering them either directly on a timeline (lines 5 and 7, which can be further segmented depending the location of embodied conducts) or relative to talk (lines 1, 4, 6). Talk constitutes a natural, linear, emergent, and progressive stream of action segmented into recognizable sounds, which provides useful temporal landmarks for locating other actions in time. Alternatively, chronometric indications of time can be used, in verbal as well as multimodal transcripts. They are even more necessary when talk is neither the main ongoing activity nor the main resource used, as with here in the second part of the excerpt. Even where they are often combined in (verbal and multimodal) transcripts, these are two different types of time. The unfolding of talk provides for a praxeological time, a relative time defined by the pace of the ongoing action—a form of *emic* time that I identify with *kairos*. In contrast, the timeline provides an abstract, measured, homogeneous time—a form of *etic* time that I identify with *chronos*. Given that transcripts need a segmentable time for ordering different temporal lapses, embodied action itself cannot be used as a primary praxeological reference because the synthetic description of its continuous flow is not analytically segmented. This is part of the paradox of inscribing time in a visual-spatial representation, making a pure emic representation of time difficult to achieve.

Moreover, alternation between talk serving as the temporal reference for embodied details and measured nontalk lapses of time can vary depending on local relevancies and consequent interpretive choices. The precise role of talk cannot be decided a priori but depends on how participants situatedly organize their actions; moreover, this distribution of multimodal resources might change during the course of the activity. When talk is neither the main resource nor the main ongoing activity, another representation can be used in which talk is subordinated to the time of the embodied course of action. For example, in order to show how the directives of a surgeon instructing his assistant are not only embedded in the ongoing surgical action but are also initiated and partially carried out by his gesture, Mondada (2014e, pp. 285–286) transcribes the verbal directive in a *subordinate* line with respect to other lines devoted to embodied conducts: Talk is no longer the primary activity or the resource. This is facilitated by the use of ELAN, which offers a continuous chronometric timeline, but it can easily be converted into the convention used here. For instance, if we go back to transcript 6a, Bruno's apology (6) could also be transcribed in the following way.

Extract 4b (detail of lines 4–6)

```
4  MIC    %ouais.* vas+-y %sors
                 *carries painting---->
   bru              +carries----->
   guy    %takes foam-----%brings foam under painting-->
5         (1.1) # (0.6) +% (0.2) *%+  (2.3) & (0.3)+ &
   mic                      ->*holds------>
   guy                 -->%deposits f%
   bru                 ->+holds-----+rearranges foam+
   guy →                                 &>pardon<&
```

This way of transcribing inserts the verbal turn ">pardon<" into a course of action initiated by Guy positioning the foam on the floor, then corrected by Bruno rearranging it, and consequently apologized for by Guy. The fast pace of the apology orients to the temporality of the silent actions, since it is formatted in such a way as to immediately respond to Bruno's correction and to be completed when the latter is also completed, thereby definitively closing the sequence.

In sum, the representation of time is central in multimodal transcripts; it is organized in relation to either talk or a timeline constituted by segments of measured time.[13] Both can account for successively ordered conducts that might unfold simultaneously. Sequentiality is fundamentally achieved through the methodically emergent and unfolding organization of actions and their ordered distribution in time. Each multimodal stream of action has different temporalities, but they are finely coordinated. Their sequential order constitutes a fundamental organizational dimension of human action.

Interactions among primates

An extreme case of interactions without words is offered by interactions among nonhuman animals. Interactions among animals present an interesting challenge for the study of interaction in general (Rossano, 2013): They are mostly embodied, although vocalizations sometimes play a role; they raise questions about their description, highlighting the risks of anthropomorphizing and attributing intentions related to the vocabulary of action used; and they provide for an interesting case of finely timed actions, raising issues of sequential organization. They also represent a further challenge for multimodal transcription. Solutions to these challenges could have important consequences not only for our understanding of interactions among animals but more generally for the analysis of embodied sequentiality in interaction.

The following fragment documents an encounter between two female baboons, Ava and Bin, engaging in what has been described as a sequence of greetings taking the form of a ritualized sexual act (Meguerditchian & Mondada, 2018; Smuts & Watanabe, 1990): One individual presents the posterior to the other, who in turn touches it and/or her genitalia. The extract[14] is transcribed using the same multimodal conventions as those used in the previous ones.

[13]An interesting alternative notation used by Luff and Heath (2015; see also Goodwin, 1981) represents segments of 0.1 seconds with dashes (-): Even though this avoids a numbered quantification of time, it relies on the measure of the segment's length, it has the advantage of showing the emergent progression of time, but it still segments it in homogeneously measured units. Hence, this alternative notation still relies on *chronos* rather than *kairos*.

[14]The extract belongs to a video corpus assembled in collaboration with A. Meguerditchian at the CNRS primatology center in Rousset (France).

Extract 5 (B8P2_2.1_Zoom_pa_5.18–5.40)

```
1          (1.4) + # (1.8)•* (0.2) + # (0.3) * (0.1)
   Ava     >>walks twd BIN•pivots, posterior twd BIN->
   Ava                    *looks back at BIN*
   Bin           +looks at AVA-----+looks at AVA's posterior--->
   fig             #fig.19           #fig.20
```

fig.19

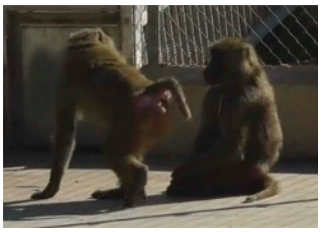


fig.20

```
2          †(0.2)† (0.2) •* (0.2) # (0.3) ‡ (0.2) ‡• (0.5)   #  ‡(0.8)‡
   Bin     †.....†holds AVA's Rposterior side-------------------->
   Ava                --->•bends knees--------------•lowers pst--->
   Ava                    *looks back at BIN------------>
   Bin                                  ‡.......‡touches genit‡.....‡
   fig                          #fig.21                    #fig.22
```

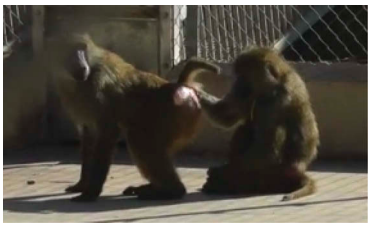


fig.21

fig.22

```
3          ‡ (0.1) * (0.7)*  (0.4) # + (0.2) ‡†• (0.4) •#
   Bin     ‡touches AVA's back---------------‡
   Bin                                  ----->†
   Bin                              ->+looks away--->>
   Ava           ->*,,,,,,*looks away--------------->>
   Ava                                      •,,,,,,,•walks away->>
   fig                              #fig.23              #fig.24
```

fig.23

fig.24

Ava walks towards Bin (1), who watches her approaching (1, Figure 19). Ava's movement is accountably designed from a distance, visibly projecting a friendly approach, and Bin's gaze shows that it is perceived as such. Once in close proximity to Bin, Ava pivots and turns her posterior toward her, looking back during this movement (1, Figure 20).

Bin in turn looks at Ava's posterior (1, Figure 20). She also responds by extending her right hand to Ava's right posterior side and holding it (2, Figure 21). Ava complies with Bin's grasp, bending her knees and further lowering her posterior when Bin's left hand extends and touches Ava's genitalia (2, Figure 22). During this haptic moment, Ava continues to look back at Bin, monitoring her.

Next, Bin places her left hand on Ava's back (3, Figure 23). Ava looks forward (3) and the sequence—as well as the encounter—is brought to completion: Bin retracts both hands and looks away; Ava walks away (Figure 24).

This brief encounter shows how interactions among primates are mutually organized: They look at each other, following each other on the move, monitoring the other's gestures and touch, and exchanging glances. Moreover, when one initiates a sequence of action (presenting the posterior), the other responds (grasping the posterior, looking at it, and touching). Touching the genitalia might not just be a form of "greeting"; it relies on a form of trust—offering a vulnerable part of one's body to another, trusting that the response will not be aggressive. This trust relies on a shared definition of the context of the encounter, which is actively established by the initial exchange of glances, mutual monitoring of the approach, and aligned bodily displays (vs. disaligned running away). It is further sustained by a constant monitoring (i.e., by Ava, constantly turning toward Bin during their haptic contact). Multimodal transcripts show that these conducts are methodically, sequentially, and temporally smoothly ordered. They offer evidence of how baboons interact socially, intersubjectively establish a common definition of the context, and achieve and reflexively sustain this context across sequences of actions initiated and responded to in an accountable way.

This enriches our understanding of social interaction among nonhuman primates, documenting the orderliness of its constitutive and relevant details. It also enriches our understanding of how embodied interactions without words among human primates can be documented, transcribed, and analyzed—especially with respect to local evidence of coparticipants' mutual orientations, the responsiveness of their movements, and the local accountability of action formation.

Conclusions

This article has shown the relation between (a) the principles of multimodal CA, (b) the way they can be operationalized in a particular transcription system, (c) the analyses that are made possible by these transcriptions, and (d) more generally the analytical and conceptual importance of transcription choices. Transcription is a practice that responds to conceptual and analytical issues and implements the solutions offered to them. In this sense, transcribing cannot be divorced from analyzing.

Transcripts and their conventions are essential tools for multimodal analysis. They articulate two crucial aspects: on the one hand, a diversity of multimodal conducts involving the bodies of participants in their entire and articulated complexities—as they are oriented to and made relevant by them as accountable and describable; on the other, coordinated but distinct temporal trajectories that can be initiated at distinct, sequentially relevant, moments, possibly subsequently unfolding in parallel and simultaneous ways. The particular convention I use aims to implement the principles of multimodal analysis in a systematic, coherent, robust, and explicit way. It allows the transcription of an unlimited range of embodied conducts; the annotation of their detailed relation to talk, if there is any; the explicit and precise representation of their relative temporal positioning and unfolding trajectories; and their synthetic description in images precisely located within the temporality of action.

This article has demonstrated how transcripts can respond to the requirements of multimodal analysis by discussing a number of phenomena on the basis of their annotation using the multimodal transcription conventions I have developed. The first empirical section showed the importance of multiple embodied details and multilayered temporalities for understanding a specific form of sequentiality typical of multimodality: *sequentially ordered simultaneities*. The next section focused on mobility as an exemplary case of practices mobilizing the entire body in ways that are specifically tied to the local ecology. The final section focused on interactions at work, in which language is not always the main resource: It showed how it is possible to transcribe stretches of action in which participants are talking very little or not at all. Finally, the last extract shows a transcription of interactions without words, using animal interactions as an extreme case. These empirical discussions highlight the crucial importance of temporality in the arrangement of multimodal gestalts and how this elaborates on the notion of sequentiality.

By discussing the foundations of multimodal analysis and transcription, this article has tackled classic as well as contemporary challenges facing multimodality, such as beyond gesture and gaze, interactions in interactional spaces, larger groups, material environments (using tools, orienting to and carrying objects), mobile settings, silent activities, and animal communities. It has also highlighted the diversity of multimodal practices involved, including occasioned resources such as forms of grasping hands, manipulations of artifacts, movements involving not only the upper part of the body (head, gesture) but also the lower parts (feet, legs, posterior), haptic contacts touching objects and coparticipants, and camera movements.

The arrangement of multimodal resources and gestalts confirms that temporality and sequentiality are the fundamental organizational principles of action among humans and possibly nonhuman primates. Sequentiality is a key concern for CA; multimodal studies reveal its complexity, raising issues of the *emic* time of actions, multiple temporalit*ies* of multilayered conducts, the interplay of successivity and simultaneity, and the coordination of related but distinct components of action. *Multiple temporalities articulated in sequentially ordered simultaneities* show that within an emergent and unfolding interactional activity several sequential orders can be achieved simultaneously by coparticipants, who can attend to all or part of them. Thus, embodied interactions are characterized by several projections going on at the same time, initiated at various moments, and responded to in ordered ways, earlier or later. In particular, responsive actions can be produced during initiating actions, which in turn reflexively adjust to them; other forms of mutual adjustments, moment by moment, remain to be described. These phenomena all show how fascinating and challenging the study of sequential organization continues to be and the crucial contribution of multimodal transcription in discovering and demonstrating it.

References

Ayass, R. (2015). Doing data: The status of transcripts in conversation analysis. *Discourse Studies*, *17*(5), 505–528. doi:10.1177/1461445615590717

Bergmann, J. (1985). Flüchtigkeit und methodische fixierung sozialer wirklichkeit: Aufzeichnungen als daten der interpretativen soziologie [Fleetingness and methodological fixation of social reality]. In W. Bonss & H. Hartmann (Eds.), *Entzauberte Wissenschaft* (pp. 299–320). Göttingen, Germany: Schwarz.

Bezemer, J., & Mavers, D. (2011). Multimodal transcription as academic practice: A social semiotic perspective. *International Journal of Social Research Methodology*, *14*(3), 191–206. doi:10.1080/13645579.2011.563616

Birdwhistell, R. L. (1970). *Kinesics in context: Essays on body motion communication*. Philadelphia, PA: University of Pennsylvania Press.

Broth, M. (2008). The studio interaction as a contextual resource for TV-production. *Journal of Pragmatics*, *40*(5), 904–926.

Broth, M., Laurier, E., & Mondada, L. (Eds.). (2014). *Studies of video practices. Video at work*. London, England: Routledge.

Broth, M., & Mondada, L. (2013). Walking away. The embodied achievement of activity closings in mobile interactions. *Journal of Pragmatics*, *47*, 41–58. doi:10.1016/j.pragma.2012.11.016

Bulwer, J. (1644). *Chirologia*. London, England: Harper.

Condon, W. S. (1971). Speech and body motion synchrony of the speaker-hearer. In D. L. Horton & J. J. Jenkins (Eds.), *Perception of language* (pp. 150–173). Columbus, OH: Merrill.

De Stefani, E. (2010). Reference as an interactively and multimodally accomplished practice. In M. Pettorino, A. Giannini, I. Chiari, & F. Dovetto (Eds.), *Spoken communication* (pp. 137–170). Newcastle, England: Cambridge Scholars Publishing.

Deppermann, A. (Ed.). (2013). Special issue: Conversation analytic studies of multimodal interaction. *Journal of Pragmatics*, *46*(1). doi:10.1016/j.pragma.2012.11.014

Goodwin, C. (1979). The interactive construction of a sentence in natural conversation. In G. Psathas (Ed.), *Everyday language: Studies in ethnomethodology* (pp. 97–121). New York, NY: Irvington.

Goodwin, C. (1981). *Conversational organization: Interaction between speakers and hearers*. New York, NY: Academic Press.

Goodwin, C. (1994). Professional vision. *American Anthropologist*, *96*, 606–633. doi:10.1525/aa.1994.96.issue-3

Goodwin, C. (2000). Action and embodiment within situated human interaction. *Journal of Pragmatics*, *32*, 1489–1522. doi:10.1016/S0378-2166(99)00096-X

Goodwin, C. (2007). Environmentally coupled gestures. In S. Duncan, J. Cassell, & E. Levy (Eds.), *Gesture and the dynamic dimensions of language* (pp. 195–212). Amsterdam, The Netherlands: John Benjamins.

Goodwin, C., & Goodwin, M. H. (1987). Concurrent operations on talk: Notes on the interactive organization of assessments. *Pragmatics*, *1*(1), 1–55.

Goodwin, C., & Goodwin, M. H. (1996). Seeing as a situated activity: Formulating planes. In D. Middleton & Y. Engestrom (Eds.), *Cognition and communication at work* (pp. 61–95). Cambridge, England: Cambridge University Press.
Haddington, P., Keisanen, T., Mondada, L., & Nevile, M. (Eds.). (2015). *Multiactivity in social interaction*. Amsterdam, The Netherlands: John Benjamins.
Haddington, P., Mondada, L., & Nevile, M. (Eds.). (2013). *Interaction and mobility: Language and the body in motion*. Berlin, Germany: De Gruyter.
Heath, C. (1986). *Body movement and speech in medical interaction*. Cambridge, England: Cambridge University Press.
Heath, C. (2013). *The dynamics of auction: Social interaction and the sale of fine art and antiques*. Cambridge, England: Cambridge University Press.
Heath, C., & Luff, P. (2000). *Technology in action*. Cambridge, England: Cambridge University Press.
Hepburn, A., & Bolden, G. B. (2017). *Transcribing for social research*. London, England: Sage.
Ivarsson, J., & Greiffenhagen, C. (2015). The organization of turn-taking in pool skate sessions. *Research on Language and Social Interaction*, *48*(4), 406–429. doi:10.1080/08351813.2015.1090114
Jefferson, G. (1973). A case of precision timing in ordinary conversation: Overlapped tag-positioned address terms in closing sequences. *Semiotica*, *9*, 47–96. doi:10.1515/semi.1973.9.1.47
Jefferson, G. (1985). An exercise in the transcription and analysis of laughter. In T. A. Van Dijk (Ed.), *Handbook of discourse analysis volume 3* (pp. 25–34). New York, NY: Academic Press.
Jefferson, G. (2004). A sketch of some orderly aspects of overlap in natural conversation (1975). In G. Lerner (Ed.), *Conversation Analysis: Studies from the first generation*. Amsterdam, The Netherlands: John Benjamins.
Keevallik, L. (2013). The interdependence of bodily demonstrations and clausal syntax. *Research on Language and Social Interaction*, *46*(1), 1–21. doi:10.1080/08351813.2013.753710
Kendon, A. (1980). Gesture and speech: Two aspects of the process of utterance. In M. R. Key (Ed.), *Nonverbal communication and language* (pp. 207–277). The Hague, The Netherlands: Mouton.
Kendon, A. (1990). *Conducting interaction: Patterns of behavior in focused encounters*. Cambridge, England: Cambridge University Press.
Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge, England: Cambridge University Press.
Laban, R. (1928). *Schrifttanz; Methodik, Orthographie, Erläuterungen* [Written dance; Methodology, Orthography, Explanations]. Vienna, Austria: Universal.
Latour, B. (1986). Visualisation and cognition: Drawing things together. *Knowledge and Society*, *6*, 1–40.
Lerner, G. H., & Zimmerman, D. H. (2002). Action and the appearance of action in the conduct of very young children. In P. Glenn, C. LeBaron, & J. Mandelbaum (Eds.), *Studies in language and social interaction* (pp. 441–457). Mahwah, NJ: Lawrence Erlbaum.
Luff, P., & Heath, C. (2015). Transcribing embodied action. In D. Tannen, H. E. Hamilton, & D. Schiffrin (Eds.), *The handbook of discourse analysis* (pp. 367–390). New York, NY: John Wiley.
McIlvenny, P. (2015). The joy of biking together: Sharing everyday experiences of vélomobility. *Mobilities*, *10*(1), 55–82. doi:10.1080/17450101.2013.844950
McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago, IL: University of Chicago Press.
McQuown, N. (1971). *The natural history of an interview*. Chicago, IL: Microfilm Collection, University of Chicago. (Original work published 1955)
Meguerditchian, A., & Mondada, L. (2018). The systematic organization of greetings among baboons (*Papio cynocephalus anubis*). Manuscript in preparation.
Mondada, L. (2007). Multimodal resources for turn-taking: Pointing and the emergence of possible next speakers. *Discourse Studies*, *9*(2), 195–226. doi:10.1177/1461445607075346
Mondada, L. (2009). Emergent focused interactions in public places. *Journal of Pragmatics.*, *41*, 1977–1997. doi:10.1016/j.pragma.2008.09.019
Mondada, L. (2014a). The local constitution of multimodal resources for social interaction. *Journal of Pragmatics*, *65*, 137–156. doi:10.1016/j.pragma.2014.04.004
Mondada, L. (2014b). Pointing, talk and the bodies: Reference and joint attention as embodied interactional achievements. In M. Seyfeddinipur & M. Gullberg (Eds.), *From gesture in conversation to visible utterance in action* (pp. 95–124). Amsterdam, The Netherlands: John Benjamins.
Mondada, L. (2014c). Bodies in action: Multimodal analysis of walking and talking. *Language and Dialogue*, *4*(3), 357–403. doi:10.1075/ld
Mondada, L. (2014d). Shooting as a research activity: The embodied production of video data. In M. Broth, E. Laurier, & L. Mondada (Eds.), *Video at work* (pp. 33–62). London, England: Routledge.
Mondada, L. (2014e). Requesting immediate action in the surgical operating room. In B. Couper-Kuhlen & P. Drew (Eds.), *Requests in interaction* (pp. 271–304). Amsterdam, The Netherlands: John Benjamins.
Mondada, L. (2016a). Challenges of multimodality: Language and the body in social interaction. *Journal of Sociolinguistics*, *20*(2), 2–32. doi:10.1111/josl.1_12177
Mondada, L. (2016b). Zwischen text und bild: Multimodale transkription [Between text and image: Multimodal transcriptions]. In H. Hausendorf, R. Schmitt, & W. Kesselheim (Eds.), *Interaktionsarchitektur, sozialtopograhie und interaktionsraum* (pp. 111–160). Tübingen, Germany: Narr.

Mondada, L., & Sorjonen, M.-L. (2016). Making multiple requests in French and Finnish convenience stores. *Language in Society*, *45*, 733–765. doi:10.1017/S0047404516000646
Mondada, L., & Svinhufvud, K. (2016). Writing-in-interaction: Studying writing as a multimodal phenomenon. *Language and Dialogue*, *6*(1), 1–53. doi:10.1075/ld.6.1.01mon
Nevile, M. (2015). The embodied turn in research on language and social interaction. *Research on Language and Social Interaction*, *48*(2), 121–151. doi:10.1080/08351813.2015.1025499
Nevile, M., Haddington, P., Heinemann, T., & Rauniomaa, M. (Eds.). (2015). *Interacting with objects: Language, materiality, and social activity*. Amsterdam, The Netherlands: John Benjamins.
Ochs, E. (1979). Transcriptions as theory. In E. Ochs & B. Schieffelin (Eds.), *Developmental pragmatics* (pp. 43–72). New York, NY: Academic Press.
Peräkylä, A., & Ruusuvuori, J. (2009). Facial expressions and spoken utterances in assessing stories and topics. *Research on Language and Social Interaction*, *42*(4), 377–394. doi:10.1080/08351810903296499
Rasmussen, G., Hazel, S., & Mortensen, K. (Eds.). (2014). Special issue: A body of resources—CA studies of social conduct. *Journal of Pragmatics*, *65*.
Rossano, F. (2012). *Gaze behavior in face-to-face interaction*. Nijmegen, The Netherlands: MPI.
Rossano, F. (2013). Sequence organization and timing of bonobo mother-infant interactions. *Interaction Studies*, *14*(2), 160–189. doi:10.1075/is.14.2.02ros
Ryave, A. L., & Schenkein, J. (1974). Notes on the art of walking. In R. Turner (Ed.), *Ethnomethodology* (pp. 265–274). Harmondsworth, England: Penguin.
Sacks, H. (1984). Notes on methodology. In J. M. Atkinson & J. Heritage (Eds.), *Structures of Social Action* (pp. 21–27). Cambridge: Cambridge University Press. (Edited by Gail Jefferson from various lectures).
Sacks, H., & Schegloff, E. A. (2002). Home position. *Gesture*, *2*(2), 133–146. doi:10.1075/gest
Schegloff, E. A. (1984). On some gestures' relation to talk. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action* (pp. 266–296). Cambridge, England: Cambridge University Press.
Schegloff, E. A. (1998). Body torque. *Social Research*, *65*(3), 535–586.
Smuts, B. B., & Watanabe, J. M. (1990). Social relationships and ritualized greetings in adult male baboons. *International Journal of Primatology*, *11*, 147–172. doi:10.1007/BF02192786
Stivers, T. (2008). Stance, alignment and affiliation during storytelling: When nodding is a token of affiliation. *Research on Language and Social Interaction*, *41*(1), 31–57. doi:10.1080/08351810701691123
Streeck, J. (1993). Gesture as communication I: Its coordination with gaze and speech. *Communication Monographs*, *60*, 275–299. doi:10.1080/03637759309376314
Streeck, J. (2009). *Gesturecraft: The manufacture of understanding*. Amsterdam, The Netherlands: John Benjamins.
Streeck, J., Goodwin, C., & LeBaron, C. (Eds.). (2011). *Embodied interaction, language and body in the material world*. Cambridge, England: Cambridge University Press.

Appendix

Transcript conventions

Talk is transcribed with the conventions developed by Gail Jefferson.

Embodied actions are transcribed according to the following conventions developed by Lorenza Mondada (for a full version and a tutorial see https://franzoesistik.philhist.unibas.ch/fileadmin/user_upload/franzoesistik/mondada_multimodal_conventions.pdf):

* *	Descriptions of embodied movements are delimited between
+ +	two identical symbols (one symbol per participant's line of action)
	and are synchronized with corresponding stretches of talk/lapses of time.
*-->	The action described continues across subsequent lines
-->*	until the same symbol is reached.
≫	The action described begins before the extract's beginning.
-->>	The action described continues after the extract's end.
....	Preparation.
------	Full extension of the movement is reached and maintained.
,,,,,	Retraction.
ava	Participant doing the embodied action is identified when (s)he is not the speaker.
fig	The exact moment at which a screen shot has been taken is indicated
#	with a symbol showing its temporal position within turn at talk/segments of time.