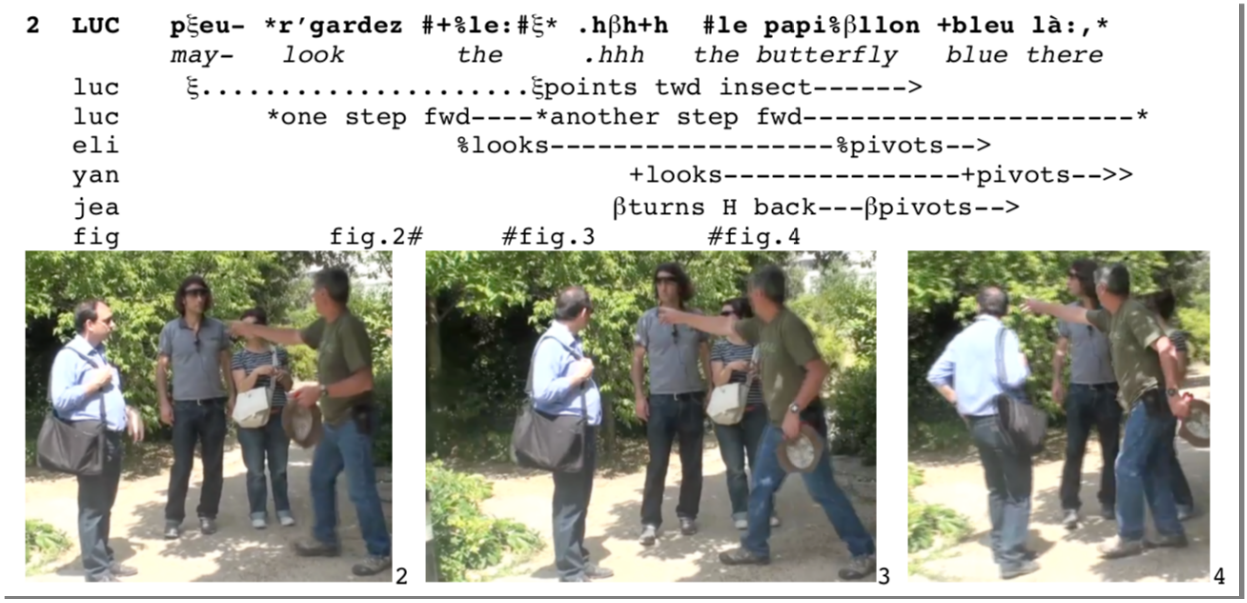
Lorenza.Mondada@unibas.ch

Conventions for multimodal transcription

(initial version: 2001; current version: 6.0.1, Dec. 2022)

How to refer to the conventions:
Please, when using these conventions, refer to the following paper:
Mondada, L. (2018). Multiple Temporalities of Language and Body in Interaction: Challenges for Transcribing Multimodality, *Research on Language and Social Interaction,* 51:1, 85-106.
and to the following web site:
https://www.lorenzamondada.net/multimodal-transcription

This file contains two versions of the conventions:
• A short version (usable in the appendix of articles and chapters)
• A long version (usable as a tutorial)



Multimodal transcript conventions

(short version)

Embodied actions are transcribed according to the following conventions developed by Lorenza Mondada (see Mondada 2018 for a conceptual discussion).
https://www.lorenzamondada.net/multimodal-transcription

* *	Descriptions of embodied actions are delimited between
+ +	two identical symbols (one symbol per participant and per type of action)
Δ Δ	that are synchronized with correspondent stretches of talk or time indications.
*--->	The action described continues across subsequent lines
---->*	until the same symbol is reached.
>>	The action described begins before the excerpt's beginning.
--->>	The action described continues after the excerpt's end.
.....	Action's preparation.
----	Action's apex is reached and maintained.
,,,,,	Action's retraction.
ric	Participant doing the embodied action is identified in small caps in the margin.
fig	The exact moment at which a screen shot has been taken
#	is indicated with a sign (#) showing its position within the turn/a time measure.

Reference:

Mondada, L. (2018). Multiple Temporalities of Language and Body in Interaction: Challenges for Transcribing Multimodality, *Research on Language and Social Interaction,* 51:1, 85-106.

Multimodal transcript conventions

(long version 1.1.2023)

https://www.lorenzamondada.net/multimodal-transcription

CA's distinctive way of transcribing is strongly related to its fundamental tenets. In particular, this concerns a focus on situated action, as it is organized within social interaction and among various co-participants, as it unfolds sequentially, establishing retrospective relations to previous actions and projecting next actions, within a temporality organized in a continuous, emergent, incremental way, strongly shaped by sequentiality. These constitutive aspects inform both the way audio-video recordings of naturally-occurring interactional activities are produced and their fine-grained transcription.

Multimodal conventions propose a way to transcribe a diversity of vocal, verbal, and embodied resources that might be relevant for the organization of social interaction, crucially considering time and sequentiality. They integrate Jefferson's conventions for transcribing talk (Jefferson 2004) with Mondada's conventions for transcribing embodiment. The latter conventions are detailed in this appendix. They have been elaborated over the years, and are conceptually discussed in Mondada (2018) (see for further instructions https://www.lorenzamondada.net/multimodal-transcription).

1. Principles

These conventions have been developed to annotate all possibly relevant embodied actions, such as gesture, gaze, body posture, movements, object manipulations, etc. that happen simultaneously with talk, vocalizations, or silent moments—constituting together the complex temporal dynamics of multimodality. Given that embodied resources are not conventionally pre-defined, but openly shaped by the participants to social interaction situatedly engaging in action with a plurality and diversity of detailed, embodied practices (see Mondada 2014 on the local definition of resources), the conventions do not pre-define any (types of) resources. Rather, they are conceived as flexible and robust enough to enable the transcription of an open range of embodied phenomena, which have not to be pre-categorized in order to be transcribed.

Multimodal transcripts aim to show the ordered details of interactivity, temporality, and accountability of action. First, concerning *interactivity,* multimodal transcripts allow the researcher to show that and how all participants are possibly constantly participating in the current action, including not only speakers, but also (silent) doers engaged, e.g., in manual, embodied, mobile activity, as well as recipients not speaking, e.g., gazing, nodding, etc., expressing real time embodied responses, and silently displaying their (mis)understanding or (dis)alignment. Second, the *temporality* of multimodal conduct integrates within the transcription a continuous flow of multimodal resources, such as gesture, gaze, body postures and movements, along with talk. These resources emerge, unfold, and are retracted across time, in both simultaneous and successive ways, exhibiting their fine-grained mutual coordination and their responsiveness to previous actions. These relations between a plurality of simultaneous successivities constitute the complex temporality crucial for understanding multimodal sequentiality (Mondada 2018). Third, the *accountability* of actions is achieved by resources which acquire ordered, distinctive, recognizable forms and trajectories in time. Annotations concern the details providing for the intelligibility of actions—in particular the emerging shape of movements—and their perceptibility—e.g. the visibility of a gesture, the

noticeability of a gaze shift, the transformation of a body posture, etc. as orchestrated by a participant and seen, glanced at, or monitored by the co-participants.

Consequently, the convention for transcribing multimodality is based on three principles:

a) *The identification of the interactants,* as doers of embodied conduct: each embodied action is attributed to a participant, indicated in the margins, who can be the same as the one talking, but also be another. Multimodal transcripts thus enable us to integrate a diversity and plurality of participants engaged in the course of the interaction (see section 2).
b) *The representation of temporality*: Each embodied action is precisely temporally located within the course of the multimodal activity. Of particular relevance for the sequential interpretation of precise timing of interaction are the moment in which the action is initiated, is incipiently emerging, and then is fully deployed, as well as the moment in which the action comes to completion. Therefore, embodied actions are delimited in time (see sections 3-4).
c) *The characterization of the embodied action* as it is intelligibly produced and visibly recognized by others: Each embodied action that is temporally located within the course of the activity can be described in two ways: A concise textual description provides for a rather analytical representation (see section 5); the use of images extracted from the video data (screenshots) provides for a rather holistic representation (see section 6).

2. Identification of the participants doing the embodied action

The conventions are here exemplified on the basis of exchanges in food shops between sellers and customers. In order to facilitate the readability of these instructions, the target phenomenon is here highlighted in grey.

Every embodied movement is attributed to a participant, who is identified by their pseudonym or category and by a set of symbols consistently used for the embodied conduct of that same participant throughout the transcription.

(1) Example:

```
* delimits gestures done by SEL
+ delimits gestures done by CUS
(where SEL = seller and CUS = customer)
```

Sometimes, several lines are necessary for indicating different embodied actions done by one participant at the same time. In this case, different symbols are used for each line (although symbols formally similar can be selected in order to create a visual link between different actions of the same person).

(2) Example:

```
* for gestures done by SEL
• for gaze by SEL
+ for gestures done by CUS1
± for gaze by CUS1
$ for gestures done by CUS2
£ for gaze by CUS2
etc.
```

If the embodied action is done by the current speaker, its description is generally *not* preceded by their identification in the margins.

(3) Example:

```
1  CUS     +EH s'il vous plaît+ un bout d'gruyère
            eh please           a piece of Gruyère
           +points at gruyère-+
```

If the embodied action is done by another participant, it will be identified in the margins. Capital letters are used for the identification of the speaker and lowercase letters for the identification of the participant doing the embodied action.

(4) Example:

```
1  CUS     •°ah oui° je• sais
            °oh yes° I know
   sel     •looks CUS-•
```

Multiple relevant embodied movements done by a participant roughly at the same time are described on different lines (bracketed by different symbols):

(5) Example:

```
1  CUS     uhm, ±(0.2) I like blue chee±ses. what +you recommend? ±+
               ±inspects fridge-------±looks at SEL--------------±
                                                  +circular gesture+
```

Note that, in some cases, when there are numerous lines referring to several participants, and for the sake of clarity, all of the participants doing embodied actions can be identified, including the speaker:

(6) Example:

```
11 SEL     $>ossau< $é um queijo* de ove£::lha,* +(0.8)+ li£:ndo
             ossau is a sheep cheese              (0.8) gorgeous
   sel                          *,,,,,,,,,,,,,,*puts piece on table->>
   cus2    $opens wallet$
   cus2                                  £looks at CUS1-----£
   cus1                                           +nods+
```

3. Temporality of the ongoing embodied actions

Every embodied action has a temporal trajectory that is delimited by two identical symbols, one indicating when the action begins and the other one when it ends. These two symbols are also located either in the line of talk or in a line with measurements of time, in order to allow a synchronization of the verbal/silent conduct and the embodied conduct. These symbols are spatially vertically aligned, one above the other, in order to represent their simultaneous unfolding. The description of the action is inserted between the two delimiting symbols.

(7a) Example:

```
1  CUS     que±ria un pasie±go
           I'd like a Pasiego
              ±points------±
```

(7b) Example

```
1  SEL     was darf's noch sein bitte?
           what else please?
```

```
2          (0.5) ± (1.5)±
   cus           ±points±
3  CUS     ah probier ich so `n Camembert
           oh I try PRT a Camembert
```

If embodied actions unfold in silence, then their annotations are related to measurements of time (Mondada 2019). This enables the transcription of silent actions as well as actions made by non-human agents, like animals (Mondada & Meguerditchian, 2022).

(8a) Example (customer tasting a sample and seller looking at the customer):

```
1          (2.3)                  + (2.0)*(1.1)± (2.2) ±+  •(1.2)•
   cus     >>puts sample in mouth+chews vigorously-----+swallows->
   cus                                          ±big nods±
   sel                                    *turns to CUS*
   sel                                                   •nods•
```

In this case, line 1 totalizes 8,8 seconds, which are segmented into smaller units depending on the annotations.

(8b) Example (three customers looking at products at a market stall)

```
1          (4.9)*(1.1)#(2.0)+(0.3)$±(0.3)•#(0.5)•(1.2)$±(1.1)#(0.5)
   cus1    >>crosses street-------$grasps cress-------$inspects->>
   cus2                     +approaches from middle of street->>
   cus2                                                ±looks->>
   cus3         *walks along the stand->>
   cus3                                  •glances•looks at veg->>
   fig                  #fig.1             #fig.2              #fig.3
```

If an embodied action begins on a line and continues either on the next line or some lines later, its description is followed by an arrow pointing to the direction of the next symbol that indicates its end. In this way, the arrow works as an instruction for the reader to search, in the following lines, for the next arrow pointing at the same symbol, closing that annotation.

(9) Example:

```
1  CUS     +°eh oui°
            °eh yes°
           +points->
2          (0.5)
3  CUS     j'aimerais+ un bout de Saint-Nectaire
           I would like a bit of Saint-Nectaire
                  ->+
```

If an embodied action begins in the middle of a pause or silence, then the pause has to be segmented into smaller temporal fragments in order to insert the symbol.

(10) Example:

```
1  CUS     °eh oui°
            eh yes
```

```
2          (0.2) + (0.3)
   cus           +points->
3  CUS     j'aimerais+ un bout de saint-nectaire
           I would like a bit of Saint-Nectaire
                  ->+
```

Note that, if an embodied action is timed with(in) a pause, there is always an identification of the participant doing it in the margin.

If an embodied action begins before the beginning of the excerpt and/or continues afterwards, its description is preceded and/or followed by a double arrow. In this case, there will be no first/second symbol opening/closing the temporal span of the action (and this is precisely indicated by the double arrow).

(11) Example of embodied action that begins before the beginning of the excerpt:

```
1  CUS     e mi +da' una bufala +di quella± là
           and you give one buffalo (Mozzarella) of those there
           >>looks at the Mozzarella------±
                +points---------+
```

(12) Example of embodied action that continues after the end of the excerpt:

```
2  SEL     vuole una bufala?
           you want a buffalo (Mozzarella)
3  CUS     sì.  quella soli±ta
           yes  the usual one
                           ±looks at a passer-by->>
```

4. Trajectories of embodied actions

Embodied actions have a temporal trajectory, which can be roughly described by distinguishing: (a) a preparatory phase; (b) a recognizable shape of the action; and (c) a retraction or withdrawal phase. This annotation is inspired by conventions used by Kendon (1990) for gestures and Goodwin (1981) for gaze. For all embodied movements, the trajectory is indicated in the following way:

... *dots* indicate that the embodied action is emerging,
,,, *commas* indicate that the embodied action is withdrawing, retracting.

The embodied action itself is described when it has reached its recognizable shape, which can also be maintained for some time (this maintenance is also visible in the dashes that fill the spatio-temporal segment corresponding to the action).

(13) Example:

```
1  CUS     donc bah j'vou- +j'voudrais+ bien du vieux+ gruyère+
           so   PRT I'd    I would like some old Gruyère
                           +.........+points--------+,,,,,,,,+
```

5. Descriptions of embodied actions

Descriptions of embodied actions are bracketed by the symbols delimiting their length in time. They are concisely described, and abbreviations can be used as well:

(14) Example

```
1  CUS     e mi +da' una bufala +di quella± là
           and you give one buffalo (Mozzarella) of those there
           >>looks at the Mozzarelle------±gz SEL->
                +points w wallet+
2  SEL     vuole una *bufa±la?*
           you want a buffalo (Mozzarella)
                     *points w knife*
   cus                    ±lks Mozzarella->>
3  CUS     sì.  +quella so+lita
           yes  the usual one
                +pts w RH+
```

Because of the spatial constraints (the descriptors have to fit into the space corresponding to a limited temporal segment), descriptors are generally as concise as possible. In this sense, the following abbreviated prepositions and adverbs can be used:

w	with
twd	towards
fwd	forward

The following actions are often abbreviated in the following way:

gz	gazes
lks	looks
pts	points
wks	walks

The following initials are often used:

R	right
L	left
H	hand
RH	right hand

The description of embodied conduct is an analytical issue: multimodal transcription addresses the local relevances of action and its contingencies (in this sense, *transcribing* is not *coding*: it does not refer to a finite set of describing categories; quite the opposite, the indexicality of action is represented by an open set of descriptors adjusted to the context). In the selection of descriptors to use, several issues should be considered, such as the use of emic descriptors (avoiding both attributions of intentionality or a reduction of movements to their physiology), and the issue of granularity (how fine-grained the characterization of the conduct is, how the conduct is segmented, etc.; see Mondada, 2018, 2019).

6. Figures

Multimodal transcripts are hybrid objects that rely on text and images, which are both integrated in the transcription. Text and images enable us to characterize embodied conduct in different and complementary ways. While textual annotations provide for an analytical description of the movements, visual images, typically in the form of screenshots (or filtered/drawn images of them), provide for a more holistic representation. Images offer a

view of an instant (a point in time) in which several multimodal features are visibly coordinated.

Figures are temporally positioned within the ongoing action. The exact moment to which the figure refers to with respect to the relevant line of talk/silence and its temporality is indicated by the symbol #. The symbol is placed both on the line of the talk/of the measured time and on the line dedicated to the figure (indicated by `fig` in the margins).

(15) Example:

```
1  SEL     #eh:*:: piensa* que# eso que es* muy líquido#
           eh:::  be aware of this that it's very liquid
               *takes w RH*2H around-----*palpates->>
   fig     #fig.1              #fig.2               fig.3#
```

The caption is generally minimal, and includes the number of the figure. Longer, more descriptive captions run the risk of being redundant with the transcript annotation and with the analytical commentary in the text.

The figures can be presented in different ways: they can correspond to the original frame of the camera view vs. they can be cropped in order to focus on the relevant details; they can be reproduced as a screen-shot vs. as a drawing vs. as a filtered image (also depending on the participants' authorizations concerning the use and reproduction of the images); they can be visually annotated with circles, arrows, etc.

References

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

Jefferson, G. (2004). Glossary of transcript symbols with an introduction. In G. H. Lerner (ed.) *Conversation Analysis: Studies from the First Generation*. Amsterdam: Benjamins, 13-31.

Kendon, A. (1990). *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Cambridge: Cambridge University Press.

Mondada, L. (2000-). Multimodal transcription conventions (constantly updated). https://www.lorenzamondada.net/multimodal-transcription

Mondada, L. (2014). The local constitution of multimodal resources for social interaction. *Journal of Pragmatics, 65,* 137-156.

Mondada, L. (2018). Multiple temporalities of language and body in interaction: challenges for transcribing multimodality. *Research on Language and Social Interaction*, 51(1), 85-106.

Mondada, L. (2019). Transcribing silent actions: a multimodal approach of sequence organization. *Social Interaction. Video-Based Studies of Human Sociality*, 2(2).

Mondada L, & Meguerditchian, A. (2022). Sequence organization and embodied mutual orientations: openings of social interactions between baboons. *Philosophical Transactions of the Royal Society B* 377: 20210101.