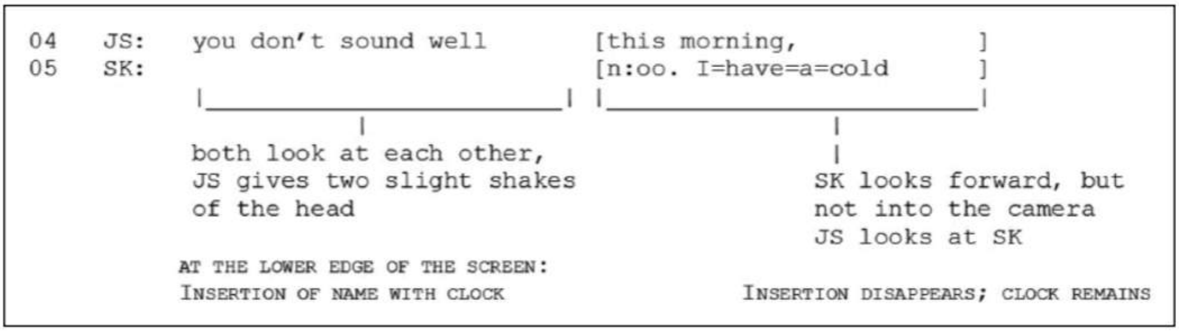
A Collection of Multimodal Transcription Methods

Organised alphabetically by first author, then by date of publication.

Ayaß, Ruth – 'Doing data: The status of transcripts in Conversation Analysis'



```
04   JS:   you don't sound well         [this morning,            ]
05   SK:                                [n:oo. I=have=a=cold      ]
           |___________________________| |__________________________|
                       |                              |
           both look at each other,                   |
           JS gives two slight shakes              SK looks forward, but
           of the head                             not into the camera
                                                   JS looks at SK
          AT THE LOWER EDGE OF THE SCREEN:
          INSERTION OF NAME WITH CLOCK          INSERTION DISAPPEARS; CLOCK REMAINS
```

Figure 1.1. Initial description: *Excerpt from Ayaß (2001: 237; our translation).*

Ayaß demonstrates a description style she presented in a 2001 text, which she describes as employing a 'logocentristic method on a number of levels'. The spoken word is the central element in this transcription method, and thus lines containing utterances are numbered, and non-verbal modalities are represented around the text. The non-verbal modalities included here include gaze and gesture, and the combination with conversation analysis (CA) techniques means that intonational information is encoded within the verbal transcription. Ayaß describes the method of transcribing 'non-linguistic' information as 'verbalized, that is, the shake of the head, the gaze, and so on'.

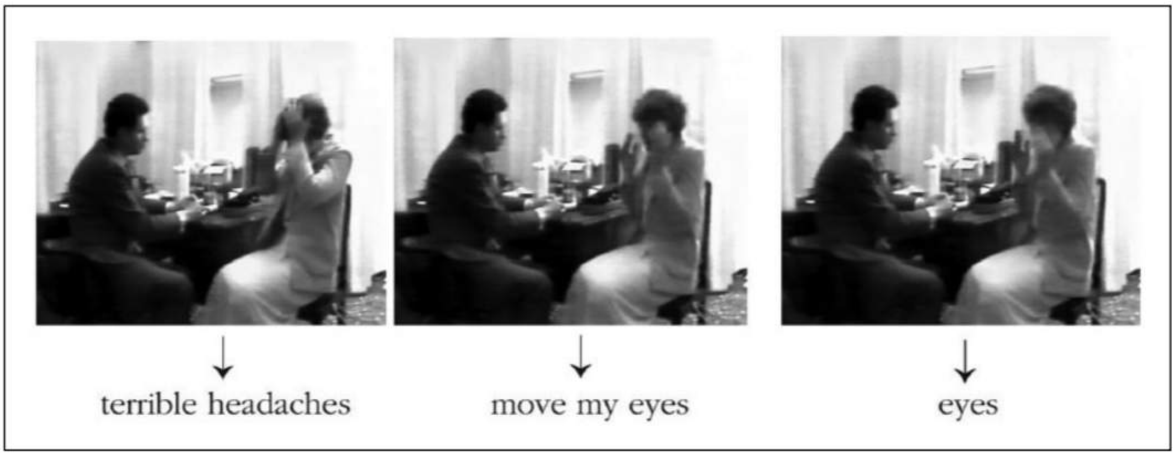
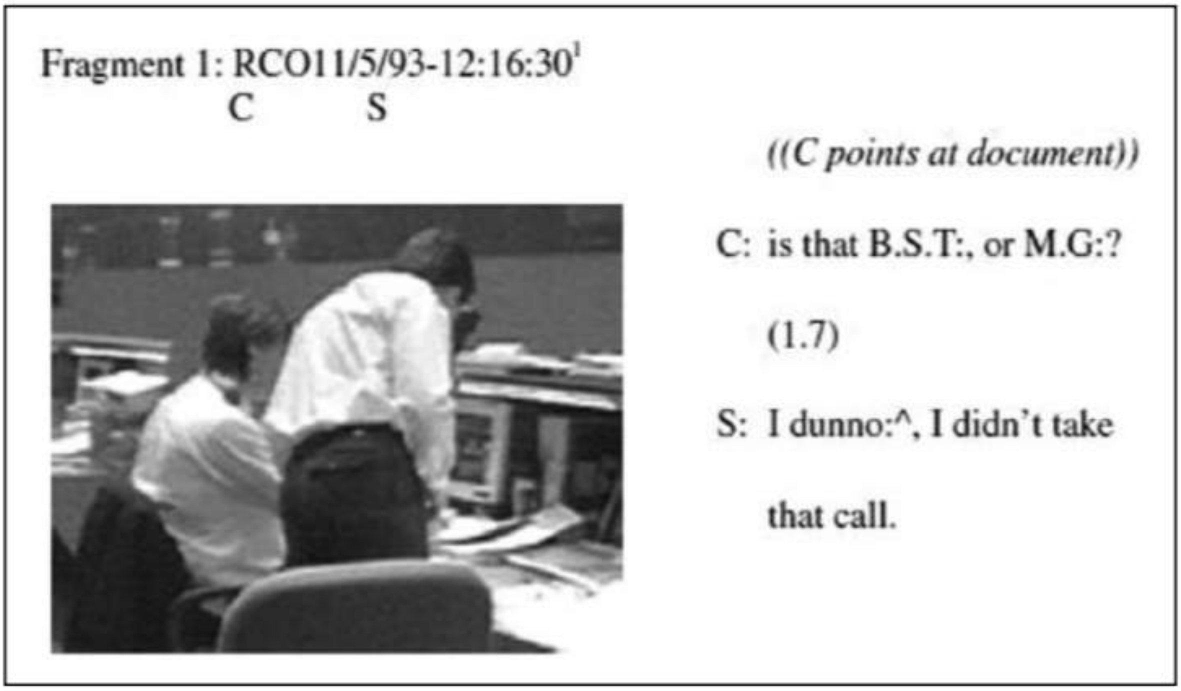


Figure 1.2. Initial description: *Excerpt from Heath's 'Demonstrative suffering' (2002: 600), reprinted with permission of the author.*

Ayaß's second transcription methods seeks to avoid the subordination of information she describes as 'non-linguistic' (that is, modalities other than speech). It is a text transcribing how a patient is likely to demonstrate the pain of a headache, using still-frame images from a recorded video. Instead of simply exemplifying the transcript, the images are the primary piece of information within the transcript. Whilst this method does have some merits, it is too minimalistic for extensive transcription of audiovisual data, and is not accessible to those unfamiliar with the method.



Fragment 1: RCO11/5/93-12:16:30[1]

((*C points at document*))

C: is that B.S.T:, or M.G:?

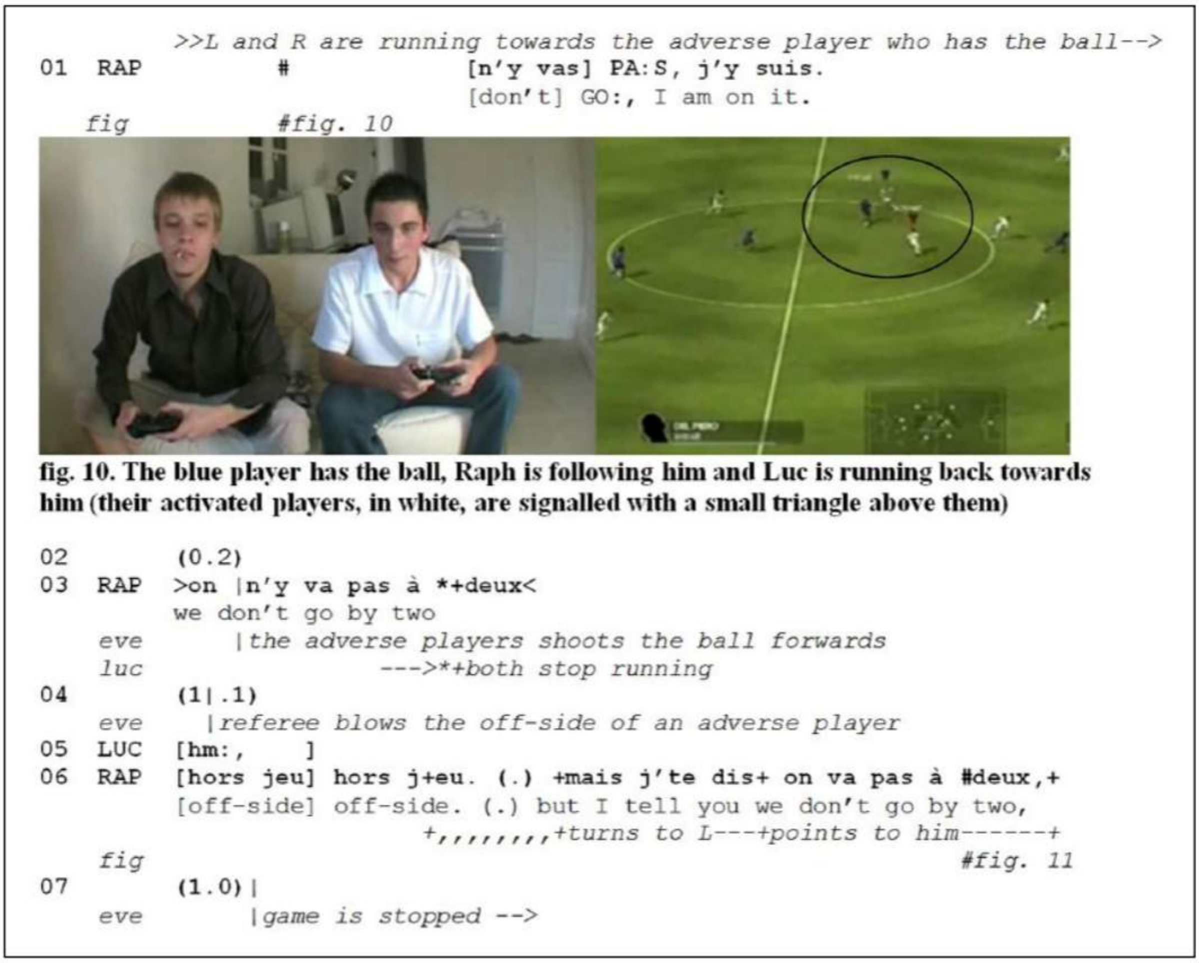
(1.7)

S: I dunno:^, I didn't take

that call.

Figure 1.3. Initial description: *Excerpt from Hindmarsh and Heath's 'Sharing the tools of the trade' (2000: 535), reprinted with permission of the authors*.

The third transcription method which Ayaß demonstrates is something of a combination of the previous two. The positioning of the still-frame image on the left of the transcript gives it prominence in Western reading styles, and it is supported by a vertical play-script style transcription drawing from CA techniques. There is no specific temporal information here aside from the date and timestamp of the fragment, so we are led to assume that 'C's action continues for the duration of the transcribed information. This is a simple yet effective method which will presumably continue to provide images alongside transcription for the duration of transcribed material.



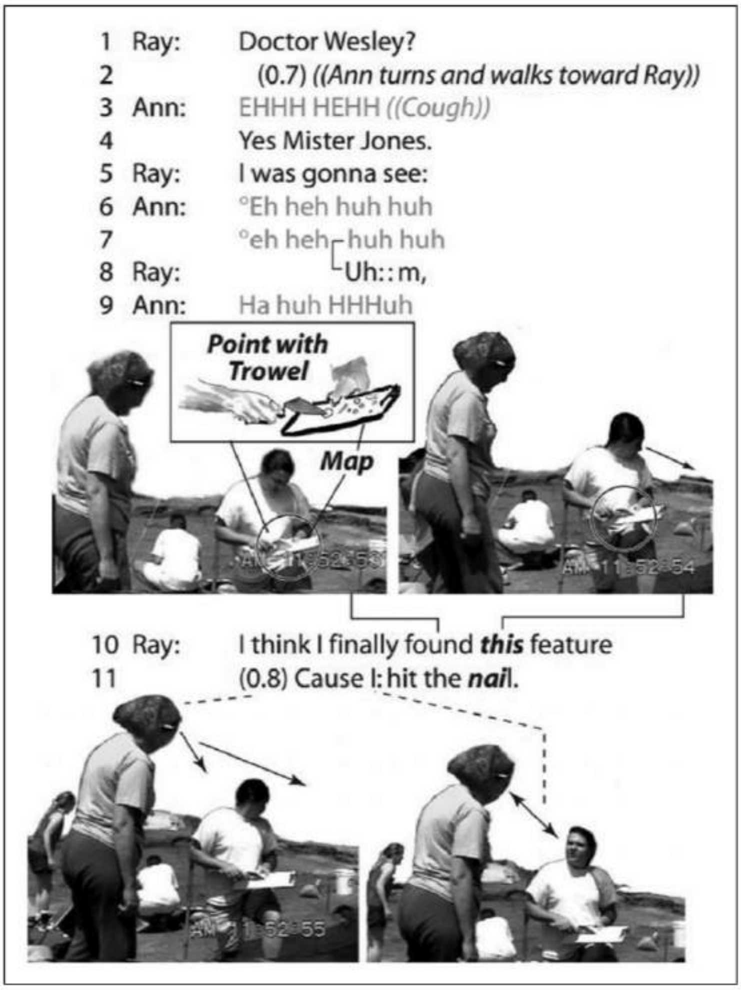
```
         >>L and R are running towards the adverse player who has the ball-->
01  RAP          #            [n'y vas] PA:S, j'y suis.
                              [don't] GO:, I am on it.
   fig          #fig. 10
```

fig. 10. The blue player has the ball, Raph is following him and Luc is running back towards him (their activated players, in white, are signalled with a small triangle above them)

```
02        (0.2)
03  RAP   >on |n'y va pas à *+deux<
          we don't go by two
    eve       |the adverse players shoots the ball forwards
    luc                  --->*+both stop running
04        (1|.1)
    eve     |referee blows the off-side of an adverse player
05  LUC   [hm:,     ]
06  RAP   [hors jeu] hors j+eu. (.) +mais j'te dis+ on va pas à #deux,+
          [off-side] off-side. (.) but I tell you we don't go by two,
                             +,,,,,,,,+turns to L---+points to him------+
    fig                                                         #fig. 11
07        (1.0)|
    eve        |game is stopped -->
```

Figure 1.4. Initial description: *Excerpt from Mondada's 'Coordinating action and talk-in-interaction in and out of video games' (2012: 246), reprinted with permission of the author.*

Ayaß's fourth method, reproduced from a text by Lorenza Mondada, shows a more complex transcript. Two participants playing a football video game are pictured on camera themselves, as well as their game represented alongside. Underneath this, a CA vertical transcript encodes what the participants say, their gestures to one another and the events occurring in-game. Gaze is presented through the first image, where both participants are looking facing the television screen. As no contrary information on gaze is given, we presume that this state does not change. The textual transcription is given in a conventional style created by Mondada, seen in Figure 10.1. Although Mondada's transcription method relies on prior knowledge of its conventions, this method is otherwise an effective means for transcribing audiovisual data.



1 Ray: Doctor Wesley?
2 (0.7) *((Ann turns and walks toward Ray))*
3 Ann: EHHH HEHH *((Cough))*
4 Yes Mister Jones.
5 Ray: I was gonna see:
6 Ann: °Eh heh huh huh
7 °eh heh⌈huh huh
8 Ray: ⌊Uh:: m,
9 Ann: Ha huh HHHuh

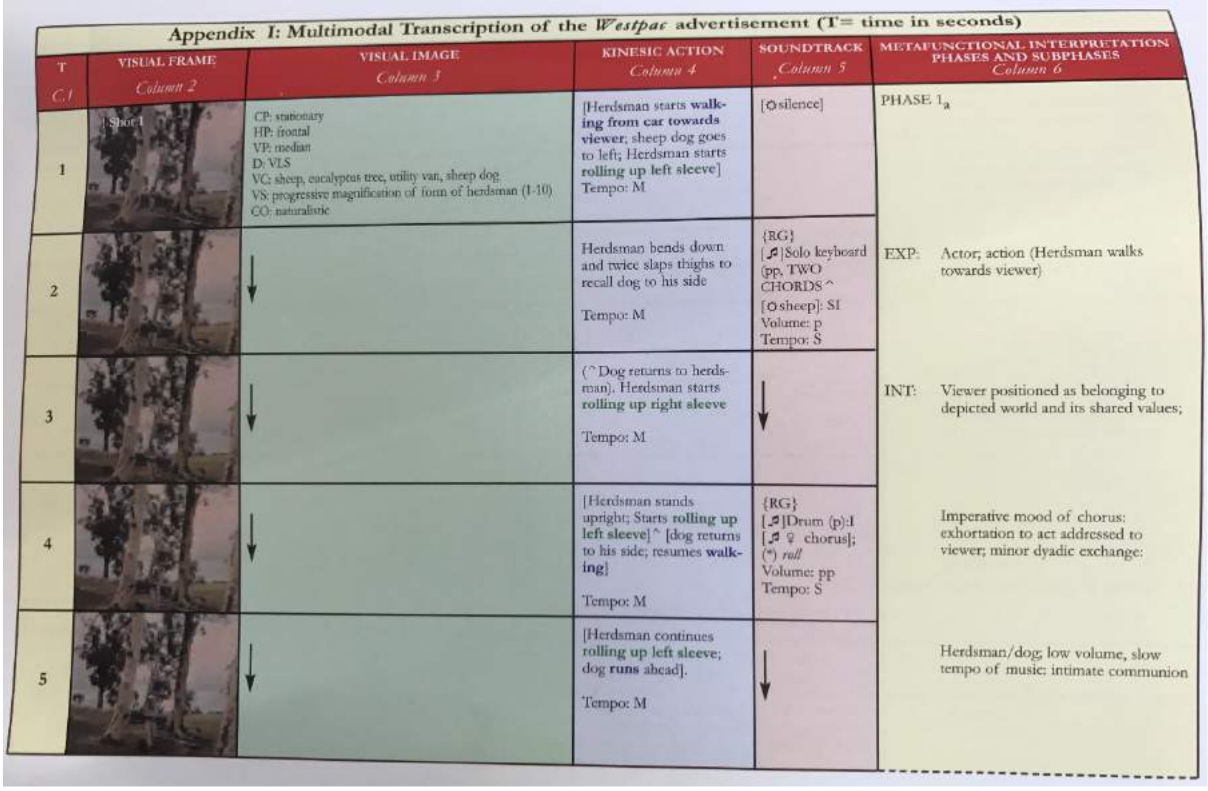
10 Ray: I think I finally found ***this*** feature
11 (0.8) Cause I: hit the ***nai***l.

Figure 1.5. Initial description:

Excerpt from Goodwin's 'Pointing as situated practice' (2003: 19), reprinted with permission of the author.

The final transcription method Ayaß exemplifies seeks to encode information on speech, gaze, and gesture. Dashed lines aim to connect people with spoken word, and directional arrows show gaze (bidirectional arrows show reciprocated gaze). The image exemplifies the situation, but it is important to note that the box containing the 'Point with Trowel' is another separate type of media, as it is an enlargement, likely through redrawing. The transcript functions adequately in portraying information in a sequential manner, but whilst the directional arrows for gaze are effective, the dashed lines indicating speech complicate the image. It seems unnecessary to attribute speech to speaker using their name in the transcript and also with dashed lines (though inconsistently).

<u>Baldry, Anthony and Thibault, Paul – ‘Multimodal Transcription and Text Analysis’</u>

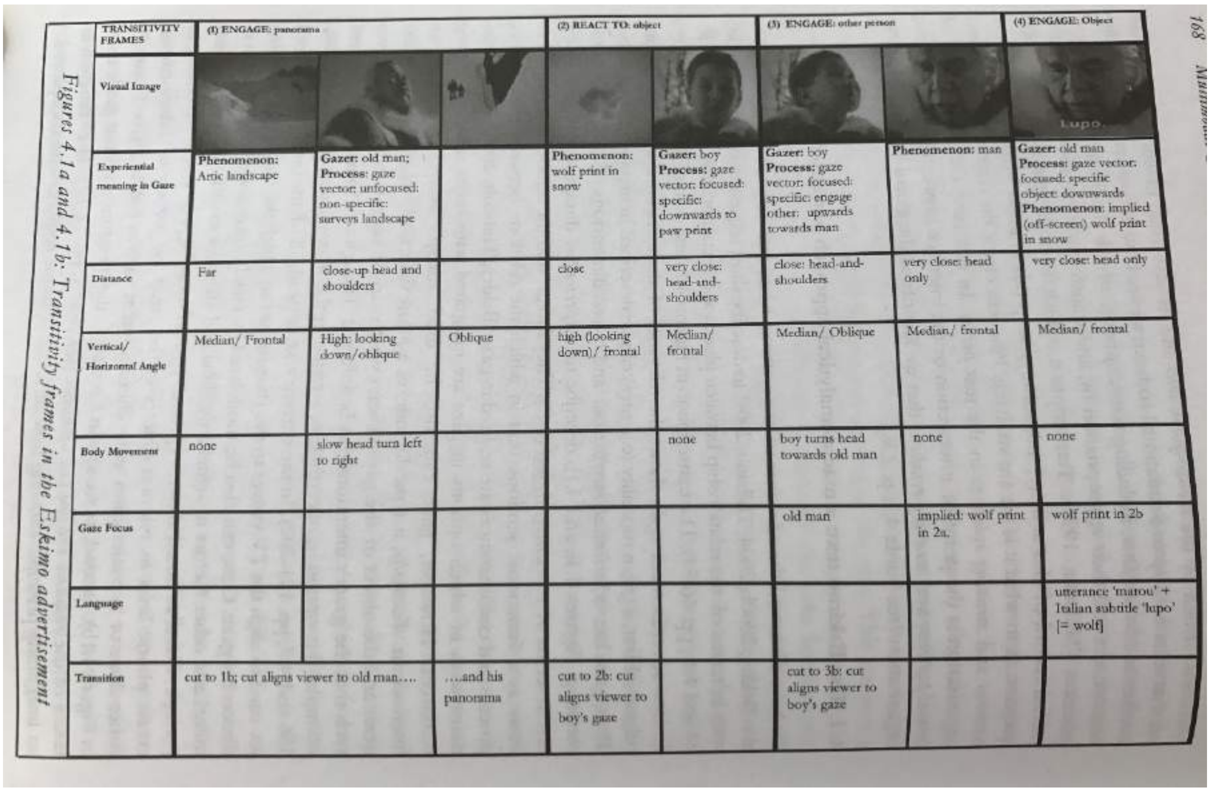


Appendix I: Multimodal Transcription of the *Westpac* advertisement (T= time in seconds)

T *C.1*	VISUAL FRAME *Column 2*	VISUAL IMAGE *Column 3*	KINESIC ACTION *Column 4*	SOUNDTRACK *Column 5*	METAFUNCTIONAL INTERPRETATION PHASES AND SUBPHASES *Column 6*
1	Shot 1	CP: stationary HP: frontal VP: median D: VLS VC: sheep, eucalyptus tree, utility van, sheep dog VS: progressive magnification of form of herdsman (1-10) CO: naturalistic	[Herdsman starts **walking from car towards viewer**; sheep dog goes to left; Herdsman starts **rolling up left sleeve**] Tempo: M	[○silence]	PHASE 1_a
2		↓	Herdsman bends down and twice slaps thighs to recall dog to his side Tempo: M	{RG} [♫]Solo keyboard (pp, TWO CHORDS ^ [○sheep]: SI Volume: p Tempo: S	EXP: Actor; action (Herdsman walks towards viewer)
3		↓	(^Dog returns to herdsman). Herdsman starts **rolling up right sleeve** Tempo: M	↓	INT: Viewer positioned as belonging to depicted world and its shared values;
4		↓	[Herdsman stands upright; Starts **rolling up left sleeve**] ^ [dog returns to his side; resumes **walking**] Tempo: M	{RG} [♫]Drum (p):I [♫ ♀ chorus]; (*) *roll* Volume: pp Tempo: S	Imperative mood of chorus: exhortation to act addressed to viewer; minor dyadic exchange:
5		↓	[Herdsman continues **rolling up left sleeve**; dog **runs** ahead]. Tempo: M	↓	Herdsman/dog, low volume, slow tempo of music: intimate communion

Figure 2.1. Initial description: *Appendix I: Multimodal Transcription of the Westpac advertisement (T= time in seconds)*

Baldry and Thibault’s first multimodal analysis method is in the form of a table, correlating time against a number of other elements of the text. The visual frame is portrayed at a rate of one frame per second, with a visual image column describing the image framing. Further comments encode information on kinesic action and sountrack, and finally metafunctional interpretations of the action. This method effectively portrays the information from the audiovisual text, and the tabular layout means it is easy to separate modalities. Furthermore, it allows a vertical reading to analyse the ongoing use of one modality, or a horizontal reading gaining all information regarding a fixed point. In all, this is a well-designed transcription method which provides useful ideas, but it is not perfect. In order to transcribe communicational cues for arrogance, we will need to focus more on modalities such as gaze and gesture, thus will need to give them more prominence within the transcription.



168

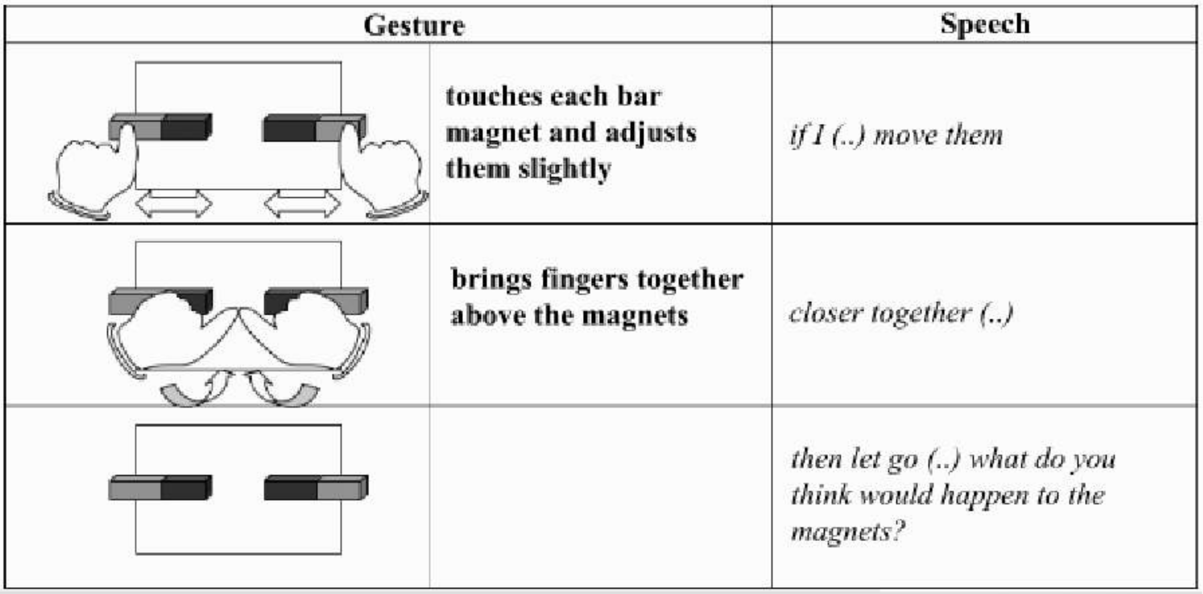
Figures 4.1a and 4.1b: Transitivity frames in the Eskimo advertisement

TRANSITIVITY FRAMES	(1) ENGAGE: panorama			(2) REACT TO: object		(3) ENGAGE: other person		(4) ENGAGE: Object
Visual Image								
Experiential meaning in Gaze	**Phenomenon:** Artic landscape	**Gazer:** old man; **Process:** gaze vector: unfocused: non-specific: surveys landscape		**Phenomenon:** wolf print in snow	**Gazer:** boy **Process:** gaze vector: focused: specific: downwards to paw print	**Gazer:** boy **Process:** gaze vector: focused: specific: engage other: upwards towards man	**Phenomenon:** man	**Gazer:** old man **Process:** gaze vector: focused: specific object: downwards **Phenomenon:** implied (off-screen) wolf print in snow
Distance	Far	close-up head and shoulders		close	very close: head-and-shoulders	close: head-and-shoulders	very close: head only	very close: head only
Vertical/ Horizontal Angle	Median/ Frontal	High: looking down/oblique	Oblique	high (looking down)/ frontal	Median/ frontal	Median/ Oblique	Median/ frontal	Median/ frontal
Body Movement	none	slow head turn left to right			none	boy turns head towards old man	none	none
Gaze Focus						old man	implied: wolf print in 2a.	wolf print in 2b
Language								utterance 'marou' + Italian subtitle 'lupo' [= wolf]
Transition	cut to 1b; cut aligns viewer to old man….		….and his panorama	cut to 2b: cut aligns viewer to boy's gaze		cut to 3b: cut aligns viewer to boy's gaze		

Figure 2.2. Initial Description: *Transitivity frames in the Eskimo advertisement*

In their second transcription method, Baldry and Thibault once again use a tabular layout, but rearrange somewhat. The still-image frames run horizontally, supporting western left-right reading patterns, thus they are portrayed temporally sequential. Instead of using one frame per second on a termporal axis, the audiovisual text is split into units according to on-screen content. These units are then titled as to their content. Far more information is transcribed within this table, including shot type and angle, gesture, gaze, and shot transition. Some units contain two still-frame images, and some three. This transcription method is much more thorough than their initial table, and especially more useful in terms of transcribing communicational cues for arrogance. The extra resources required, such as gaze and gesture, are laid out accessibly. Due to the horizontal sequence, reading one modality left to right gives its progression through the text. This is a more natural format for most western readers. Therefore, Baldry and Thibault's second transcription method, which they have named 'macro-analytical' is one of the most comprehensive and useful we have seen so far.

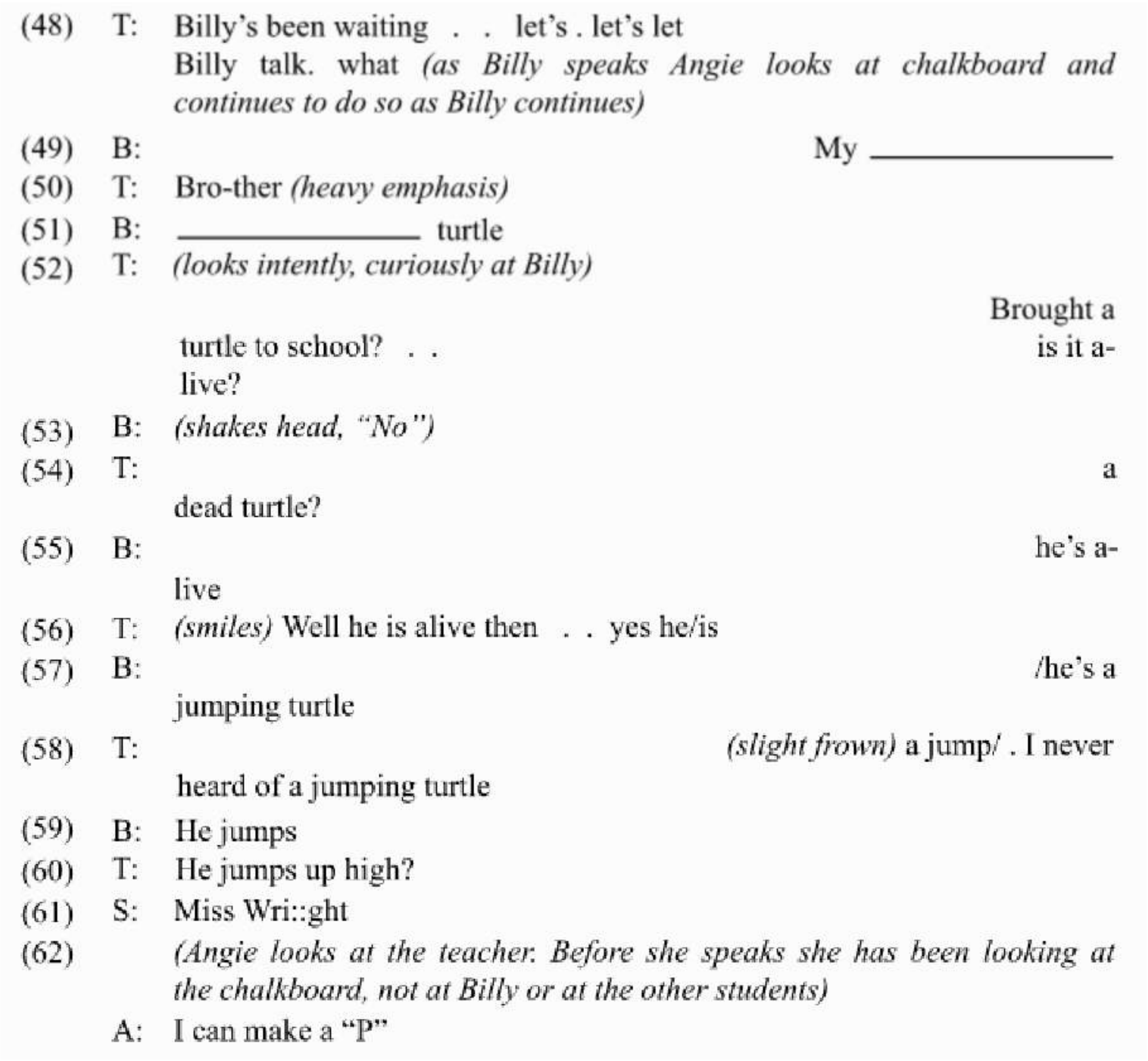
Bezemer, Jeff and Mavers, Diane – 'Multimodal transcription as academic practice: a social semiotic perspective'



Gesture		Speech
	touches each bar magnet and adjusts them slightly	*if I (..) move them*
	brings fingers together above the magnets	*closer together (..)*
		then let go (..) what do you think would happen to the magnets?

Figure 3.1. Initial description: *Excerpt from [Mavers (2009: p. 146)] Reprinted with permission of the author(s).*

The first transcription method that Bezemer and Mavers present comes from a previous text by Mavers herself. It is a minimalistic transcription method encoding gesture and speech designed to instruct students. Gesture is both describe linguistically and illustrated, whilst speech is transcribed orthographically. Whilst it is apparent that the transcription would be sufficient to portray the limited instructions the students require, it is not a comprehensive transcription method capable of dealing with other modalities such as gaze and intonation.



(48) T: Billy's been waiting . . let's . let's let
Billy talk. what *(as Billy speaks Angie looks at chalkboard and continues to do so as Billy continues)*
(49) B: My ________
(50) T: Bro-ther *(heavy emphasis)*
(51) B: ________ turtle
(52) T: *(looks intently, curiously at Billy)*
Brought a
turtle to school? . . is it a-
live?
(53) B: *(shakes head, "No")*
(54) T: a
dead turtle?
(55) B: he's a-
live
(56) T: *(smiles)* Well he is alive then . . yes he/is
(57) B: /he's a
jumping turtle
(58) T: *(slight frown)* a jump/ . I never
heard of a jumping turtle
(59) B: He jumps
(60) T: He jumps up high?
(61) S: Miss Wri::ght
(62) *(Angie looks at the teacher. Before she speaks she has been looking at the chalkboard, not at Billy or at the other students)*
A: I can make a "P"

Figure 3.2. Initial description: *Excerpt from [Erickson (2004: p. 58)] Reprinted with permission of the author(s).*

Another transcription method reproduced by Bezemer and Mavers opts from a play-script style. Speech is the most prominent modality in this transcription, whilst other information is given in italicised text. Modalities other than speech are not differentiated, but are described linguistically with little distinction. Information on gaze, gesture and intonation are all given above, but all in italics. This method may be of some use where there is little non-verbal communication in the interaction, but in the interests of finding a flexible and adaptive transcription method, this is too restrictive.

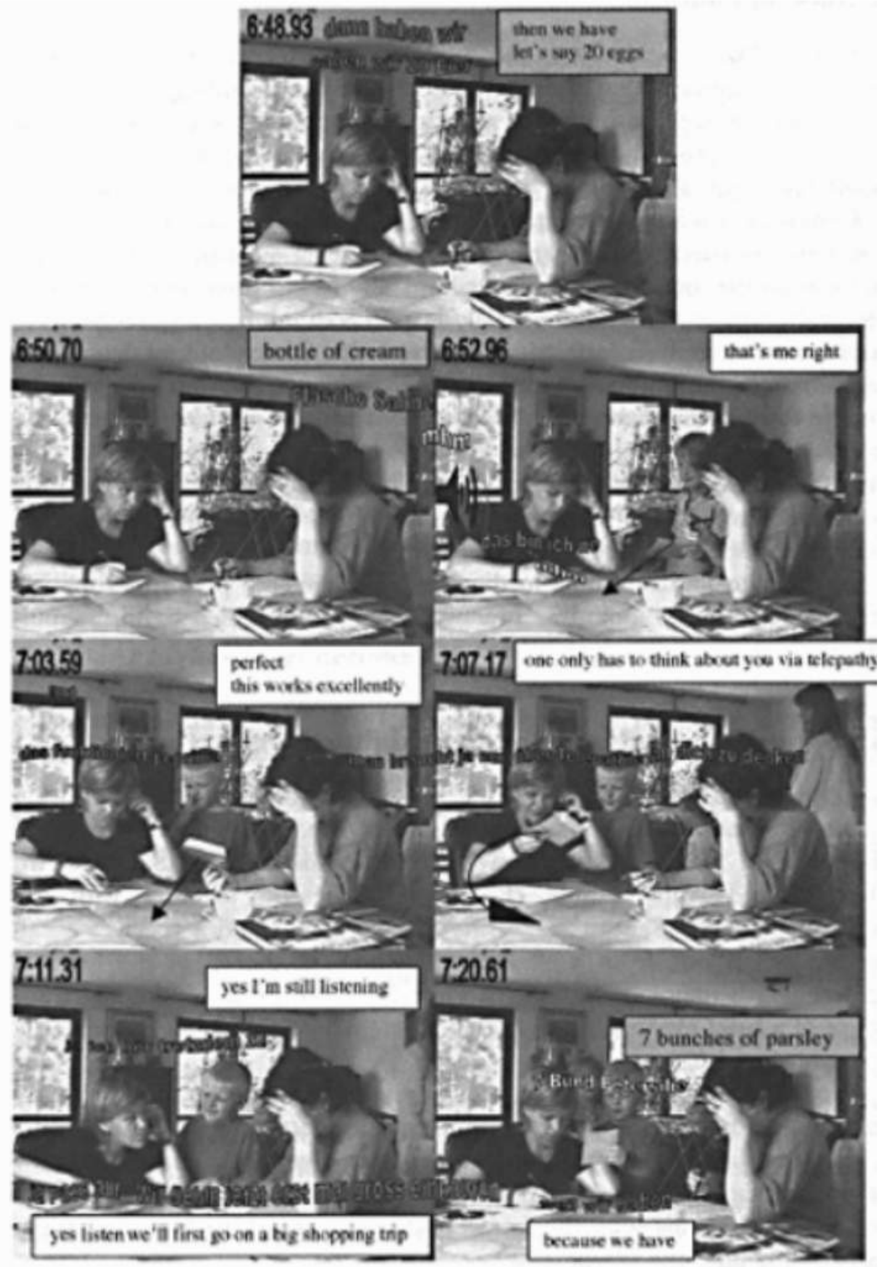


Figure 3.3. Initial description:

Excerpt of transcript from [Norris (2004: p. 102)] Reprinted with permission of the author(s).

Bezemer and Mavers exemplify a transcription method from a 2004 text. Several strategies are used within this transcription to express different information. Font and text layout is used to express intonation and stress, and different participants are given different shades to differentiate between them. English translation of initially German speech is given in text boxes, and gaze is shown using directional arrows. Gesture is evident through the still-frame images. Whilst this method does encode most of the information we will require, the layout can be difficult to make out. Considering this interaction only contains two participants, a larger discussion would lead to a very cluttered description. This method does have its merits, but it will be insufficient in transcribing larger group discussion.

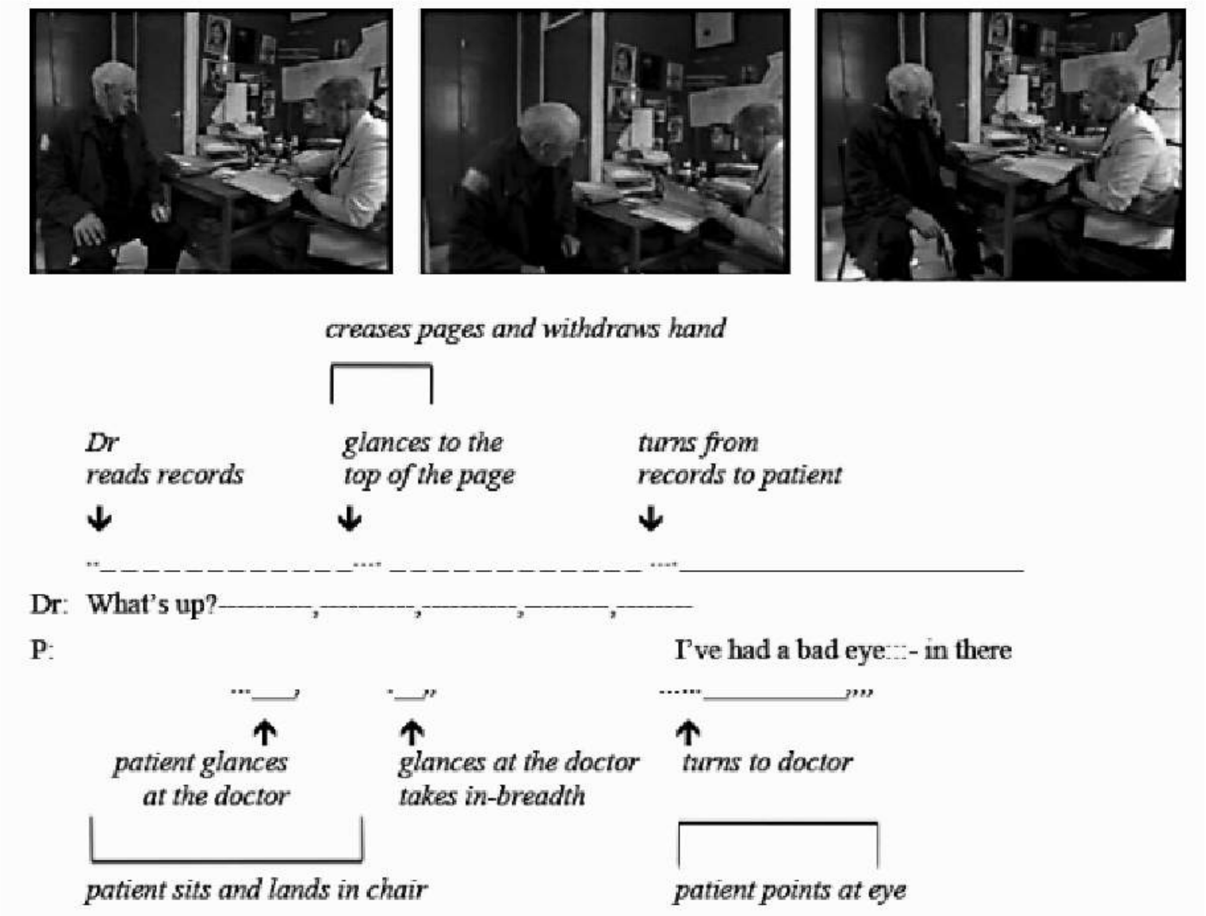

creases pages and withdraws hand
Dr reads records
glances to the top of the page
turns from records to patient
Dr: What's up?———,———,———,———,———
P: I've had a bad eye:::- in there
patient glances at the doctor
glances at the doctor takes in-breadth
turns to doctor
patient sits and lands in chair
patient points at eye

Figure 3.4. Initial description: *Excerpt of transcript from [Heath, Hindmarsh & Luff (2010: p. 70)] Reprinted with permission of the author(s).*

This transcription method is drawn from a 2010 text, and similarly to Figure 1.4 combines still-frame images with orthographic transcription. This transcription, however, uses a horizontal transcription of speech. This is possible due to the short fragment of data being transcribed – further transcription would require extending the transcript horizontally or representing another fragment in similar style below it. Speech is transcribed orthographically, and there are two layers of information which seem to be an overlap of gaze and gesture. The italicised information closest to speech focuses mainly on gaze, but also described how the participants orientate their bodies. The second layer describes other action within the audiovisual text, such as sitting and pointing. There is an attempt at transcribing the temporal length of actions, but close analysis of this requires understanding of the underlying conventions used. The horizontal layout of this transcription is effective, as it demonstrates how different actions are occurring simultaneous to, or between speech acts. It would benefit from some refining, and better separation of modalities like gaze and gesture.

Bezemer, Jeff – ‘How to Transcribe Multimodal Interaction?’ DRAFT

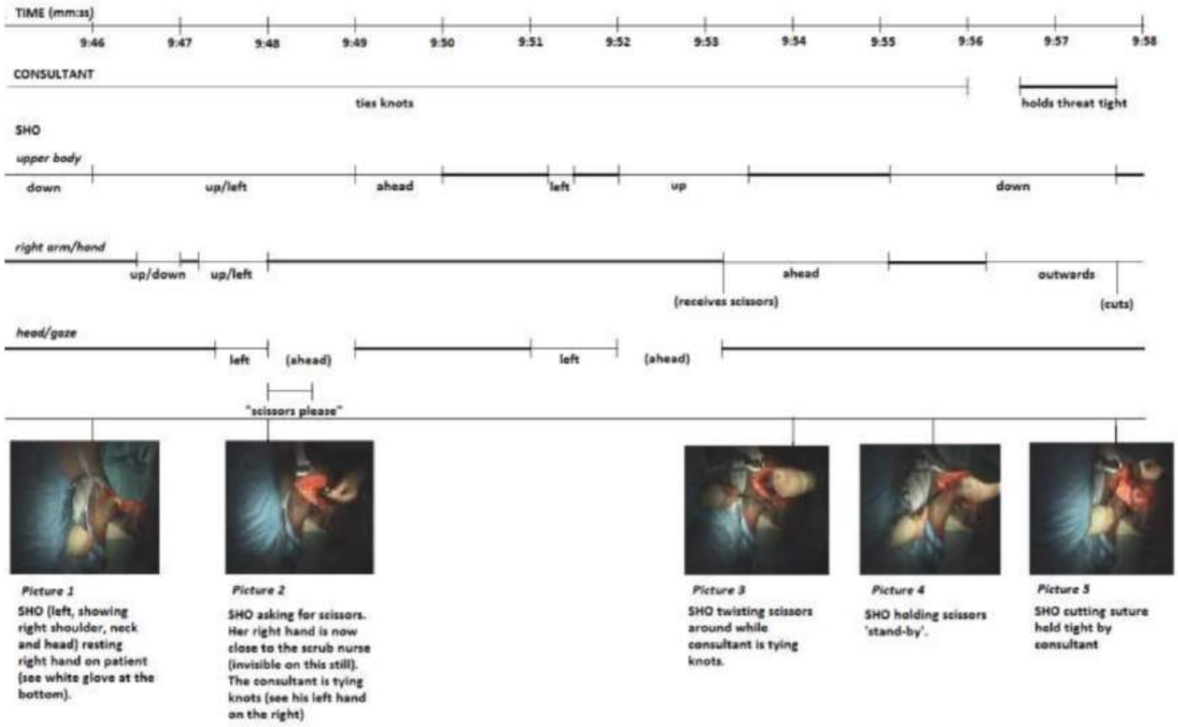
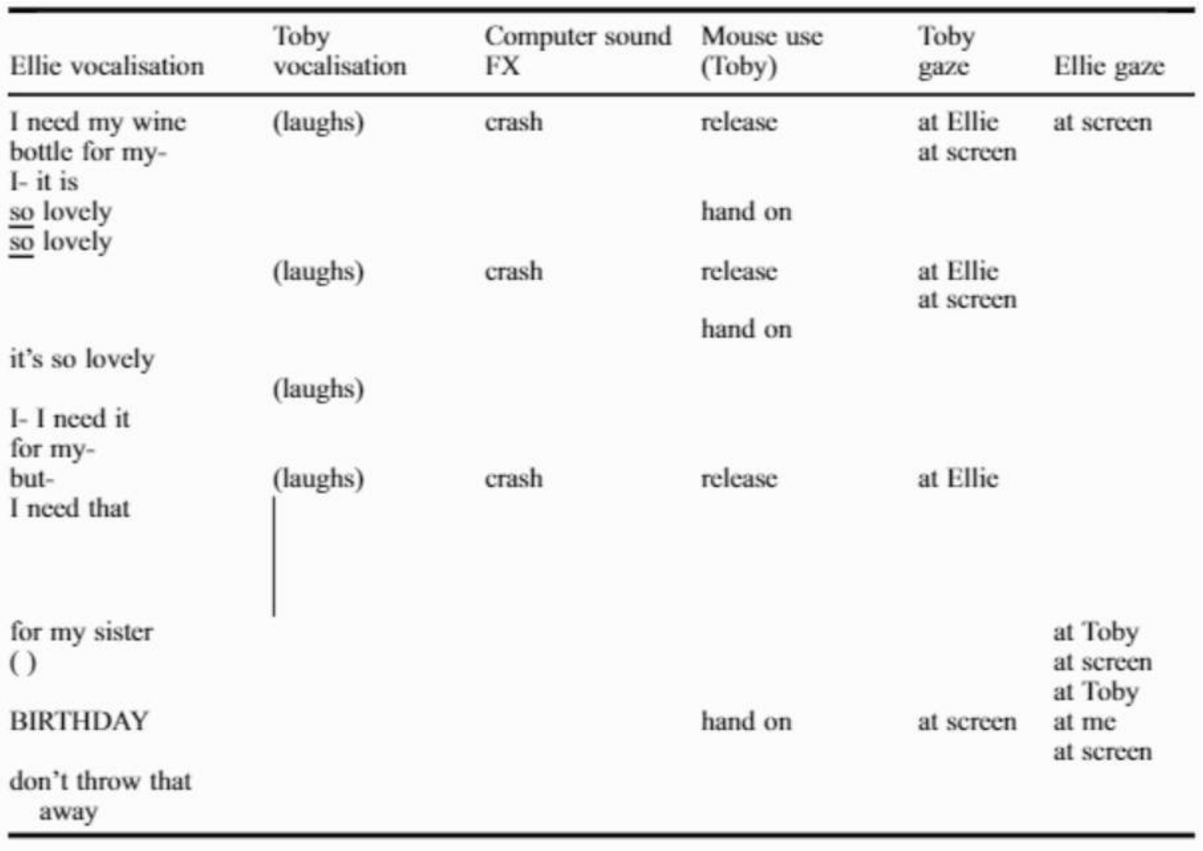

TIME (mm:ss)
9:46
9:47
9:48
9:49
9:50
9:51
9:52
9:53
9:54
9:55
9:56
9:57
9:58
CONSULTANT
ties knots
holds threat tight
SHO
upper body
down
up/left
ahead
left
up
down
right arm/hand
up/down
up/left
ahead
(receives scissors)
outwards
(cuts)
head/gaze
left
(ahead)
left
(ahead)
"scissors please"
Picture 1
SHO (left, showing right shoulder, neck and head) resting right hand on patient (see white glove at the bottom).
Picture 2
SHO asking for scissors. Her right hand is now close to the scrub nurse (invisible on this still). The consultant is tying knots (see his left hand on the right)
Picture 3
SHO twisting scissors around while consultant is tying knots.
Picture 4
SHO holding scissors 'stand-by'.
Picture 5
SHO cutting suture held tight by consultant

Figure 4.1. No initial description.

Bezemer arranges this transcription temporally along a horizontal axis. The timestamp of the multimodal text is given continually across the top of the transcription. There are two main participants, the consultant and the senior house officer (SHO). Each participant’s contributions are separated into their various modalities, though the consultant only enacts gestures. Each modality progresses horizontally along with the time, ad still-frame images capture what is happening in the surgical theatre simultaneously. The SHO’s gestures are split further into the body part which conducts them, thus ‘upper body’, ‘right arm/hand’ and ‘head/gaze’ are transcribed separately. The combination of ‘head/gaze’ implies that gaze is linked to where the head is pointing. Speech, though minimal, is given on its own horizontal access. This transcription method is effective as it is easy to read each modality as it progresses. Moreover, it is apparent which actions are coinciding, and the transition between gestures is provided in tandem with temporal information. Thus, the transcript remains uncluttered and manageable. Bezemer’s transcription method is one of the most effective documented so far, and will prove valuable in transcribing future multimodal texts.

Cowan, Kate – 'Multimodal transcription of video: examining interaction in Early Years classrooms'



Ellie vocalisation	Toby vocalisation	Computer sound FX	Mouse use (Toby)	Toby gaze	Ellie gaze
I need my wine	(laughs)	crash	release	at Ellie	at screen
bottle for my-				at screen	
I- it is					
so lovely			hand on		
so lovely					
	(laughs)	crash	release	at Ellie	
				at screen	
			hand on		
it's so lovely					
	(laughs)				
I- I need it					
for my-					
but-	(laughs)	crash	release	at Ellie	
I need that					
for my sister					at Toby
( )					at screen
					at Toby
BIRTHDAY			hand on	at screen	at me
					at screen
don't throw that away					

Figure 5.1. Initial description: *Multimodal transcript of video using a tabular layout.*

Cowan uses this multimodal transcription method so as to invite 'alternative non-temporal reading pathways'. Cowan separates vocalisations from Ellie and Toby, and also transcribes sounds from the computer program, Toby's mouse use (a form of gesture) and both of their gaze patterns. This transcription method was designed specifically for use in early years classroom usage where students are recorded using a computer program. Thus, gesture is restricted to 'Mouse use (Toby)' as this is seen as the only relevant gesture. A setback of this method is that no temporal information is encoded, aside for the sequence of events. It is impossible to tell how long vocalisations or actions take, except by comparison to simultaneous events. Nevertheless, the transcription is effective within the sphere of its own goals. Considering an interaction with more participants, it could become cumbersome to encode all of their speech, gaze and gesture separately. Thus, whilst this method is effective for transcribe the minimal speech, gesture and gaze by two participants, it would struggle to support larger and more dynamic multimodal events.

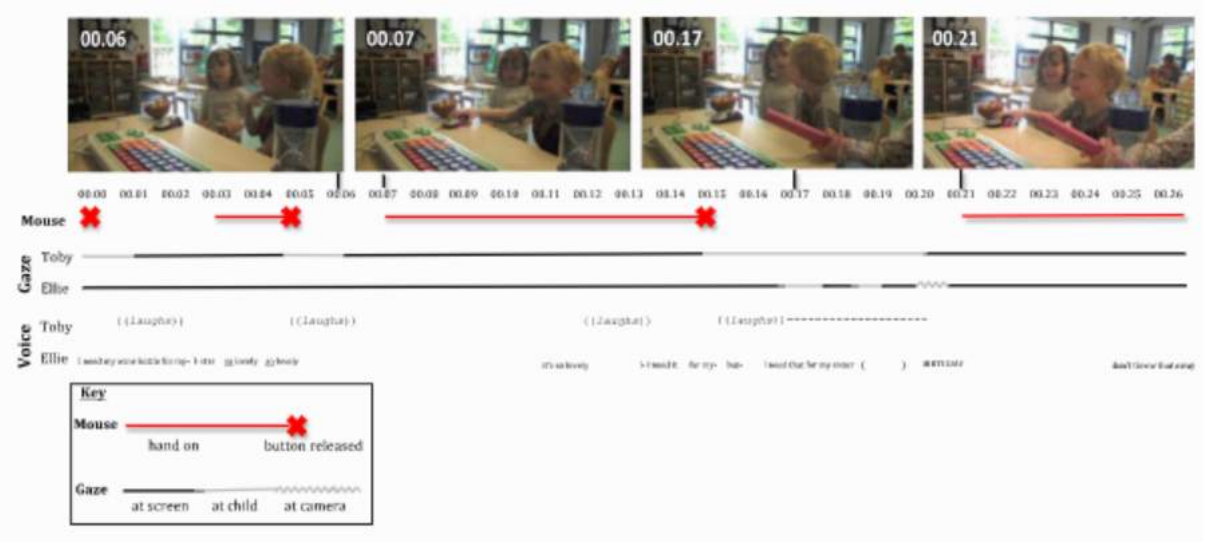
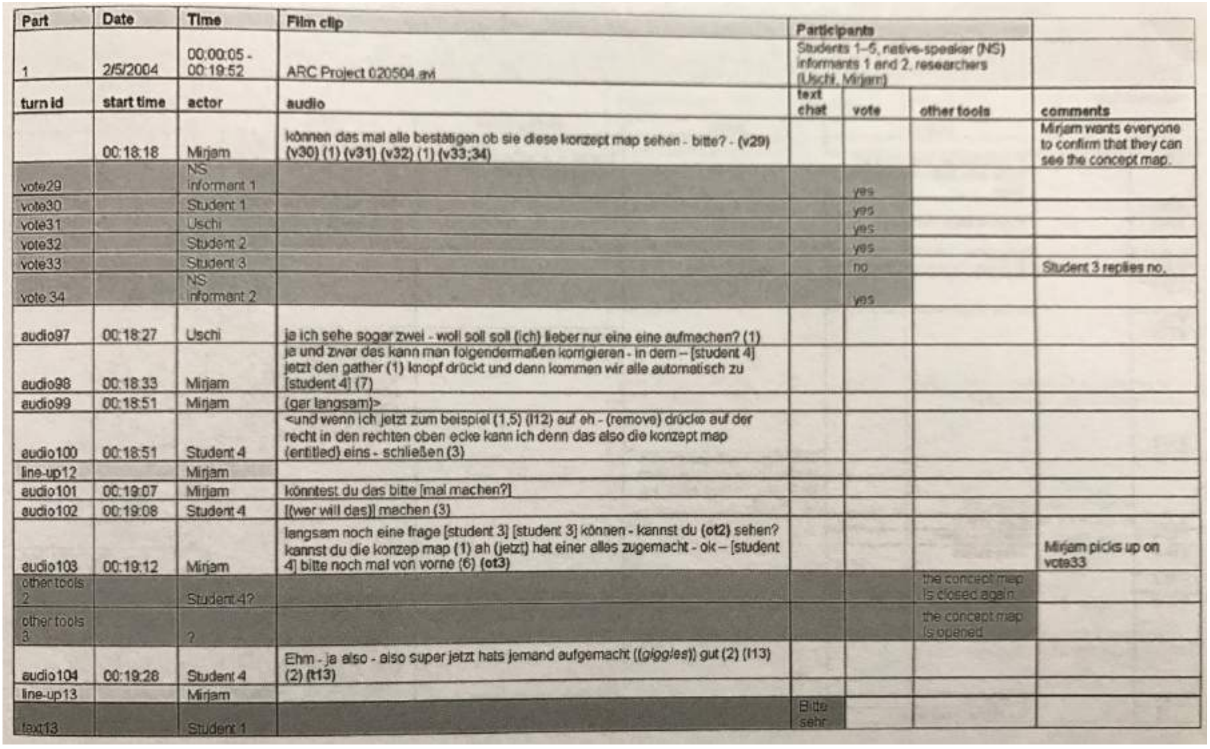


Figure 5.2. Initial description: *Multimodal transcript of video using a timeline layout.*

Cowan's second transcription style is similar to Bezemer's in Figure 4.1. The same information is encoded as in Figure 5.1, but it is far more effective in a timeline style layout. Information continues horizontally along a temporal access, and a key renders the text accessible even to a lay audience given some time to become familiar with it. In this manner, it is far more easy to see the length of events, and the addition of still-frame images allows a cross reference where any confusion may occur. It is apparent that this flexible style of transcription could accept more participants simply by adding their contributions as additional horizontal axis. Within the transcription of speech, some CA techniques are observed to encode information such as stress and intonation. In all, this is a flexible and effective transcription method.

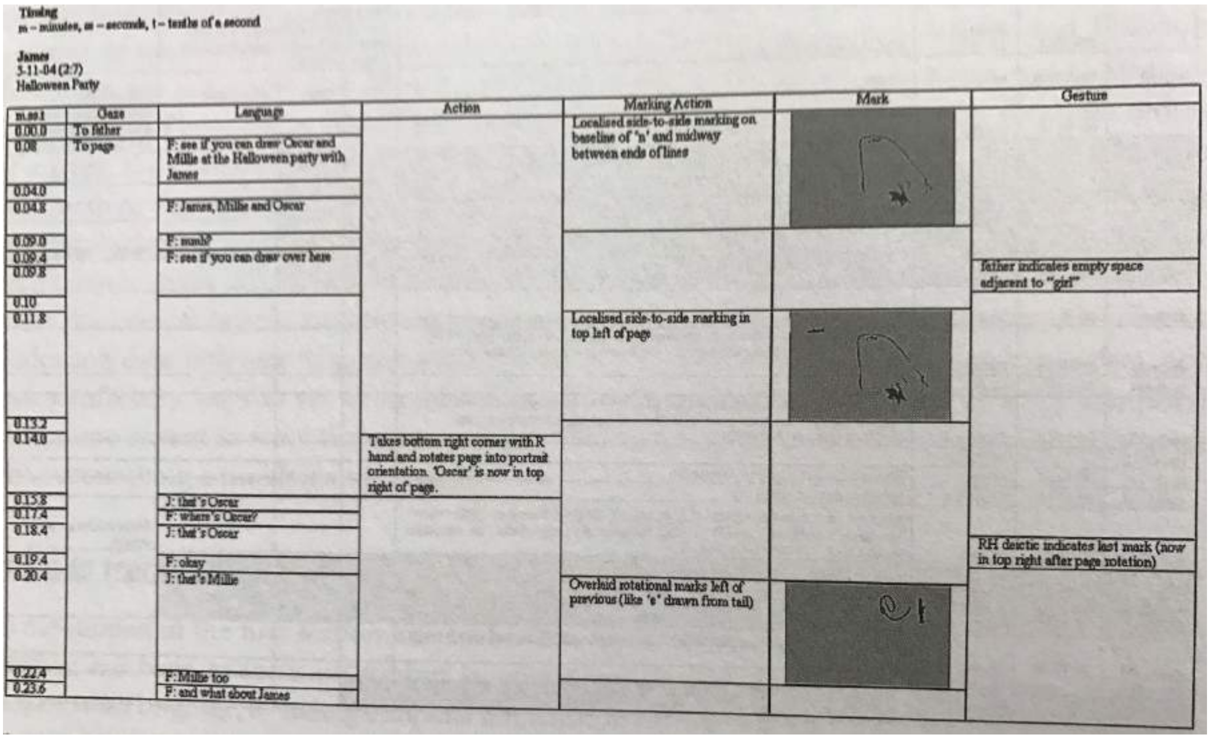
Flewitt, Rosie – 'What are multimodal data and transcription?'



Part	Date	Time	Film clip	Participants			
1	2/5/2004	00:00:05 - 00:19:52	ARC Project 020504.avi	Students 1–5, native-speaker (NS) informants 1 and 2, researchers (Uschi, Mirjam)			
turn id	**start time**	**actor**	**audio**	**text chat**	**vote**	**other tools**	**comments**
	00:18:18	Mirjam	können das mal alle bestätigen ob sie diese konzept map sehen - bitte? - (v29) (v30) (1) (v31) (v32) (1) (v33;34)				Mirjam wants everyone to confirm that they can see the concept map.
vote29		NS informant 1			yes		
vote30		Student 1			yes		
vote31		Uschi			yes		
vote32		Student 2			yes		
vote33		Student 3			no		Student 3 replies no.
vote 34		NS informant 2			yes		
audio97	00:18:27	Uschi	ja ich sehe sogar zwei - woll soll soll (ich) lieber nur eine eine aufmachen? (1)				
audio98	00:18:33	Mirjam	ja und zwar das kann man folgendermaßen korrigieren - in dem – [student 4] jetzt den gather (1) knopf drückt und dann kommen wir alle automatisch zu [student 4] (7)				
audio99	00:18:51	Mirjam	(gar langsam)>				
audio100	00:18:51	Student 4	<und wenn ich jetzt zum beispiel (1,5) (l12) auf eh - (remove) drücke auf der recht in den rechten oben ecke kann ich denn das also die konzept map (entitled) eins - schließen (3)				
line-up12		Mirjam					
audio101	00:19:07	Mirjam	könntest du das bitte [mal machen?]				
audio102	00:19:08	Student 4	[(wer will das)] machen (3)				
audio103	00:19:12	Mirjam	langsam noch eine frage [student 3] [student 3] können - kannst du **(ot2)** sehen? kannst du die konzep map (1) ah (jetzt) hat einer alles zugemacht - ok – [student 4] bitte noch mal von vorne (6) **(ot3)**				Mirjam picks up on vote33
other tools 2		Student 4?				the concept map is closed again	
other tools 3		?				the concept map is opened	
audio104	00:19:28	Student 4	Ehm - ja also - also super jetzt hats jemand aufgemacht ((*giggles*)) gut (2) (l13) (2) **(t13)**				
line-up13		Mirjam					
text13		Student 1		Bitte sehr			

Figure 6.1. Initial description: *Transcription of synchronous online interaction in audio conferencing environment.*

Flewitt's first example of multimodal transcription is drawn from a 2006 text, and is designed for transcribing audio conferencing. The transcription is tabular, with relatively basic information. The data is divided into turns, and these are allocated a timestamp. Audio from a film clip is transcribed orthographically, and there are also columns for text chat, and other actions within the call. The transcription is effective in collecting what the participants said within the discourse and their actions within the call (such as voting), but the method would not support a more complex multimodal transcription requirement. There are no resources for encoding external gaze and gesture. This transcription serves its purpose, but is restrictive due to the specificity of that purpose.



Timing
m – minutes, ss – seconds, t – tenths of a second

James
3-11-04 (2:7)
Halloween Party

m.ss.t	Gaze	Language	Action	Marking Action	Mark	Gesture
0.00.0	To father			Localised side-to-side marking on baseline of 'n' and midway between ends of lines		
0.08	To page	F: see if you can draw Oscar and Millie at the Halloween party with James				
0.04.0						
0.04.8		F: James, Millie and Oscar				
0.09.0		F: mmh?				
0.09.4		F: see if you can draw over here				
0.09.8						father indicates empty space adjarent to "girl"
0.10						
0.11.8				Localised side-to-side marking in top left of page		
0.13.2						
0.14.0			Takes bottom right corner with R hand and rotates page into portrait orientation. 'Oscar' is now in top right of page.			
0.15.8		J: that's Oscar				
0.17.4		F: where's Oscar?				
0.18.4		J: that's Oscar				
0.19.4		F: okay				RH deictic indicates last mark (now in top right after page rotation)
0.20.4		J: that's Millie		Overlaid rotational marks left of previous (like 'e' drawn from tail)		
0.22.4		F: Millie too				
0.23.6		F: and what about James				

Figure 6.2. Initial description: *Transcription of a section of an episode of naturally occurring interaction between adult and child around the construction of a text.*

This transcription also uses a tabular format, and is somewhat more extensive than the previous one. The transcription is once again separated into event turns, and each turn is given a timestamp, including those without spoken language. This method encapsulates all main transcription requirements: gaze, gesture, language and actions. This method was designed to transcribe a drawing exercise between an adult and child, thus there is a section containing still-frame images of the drawing as it progresses. Once again, we see gesture and action separated, with gesture relating to indicative or pointing actions, whereas action encodes physical occurrences beyond these limits. Spoken language is transcribed orthographically, with no input from CA. Thus, there is no information encoded on stress, intonation and similar as these are not required for the specific purpose of this transcription. This transcription is helpful insofar as providing another example of a tabular transcription method, and we can continue to evaluate what strengths and weaknesses this style holds.

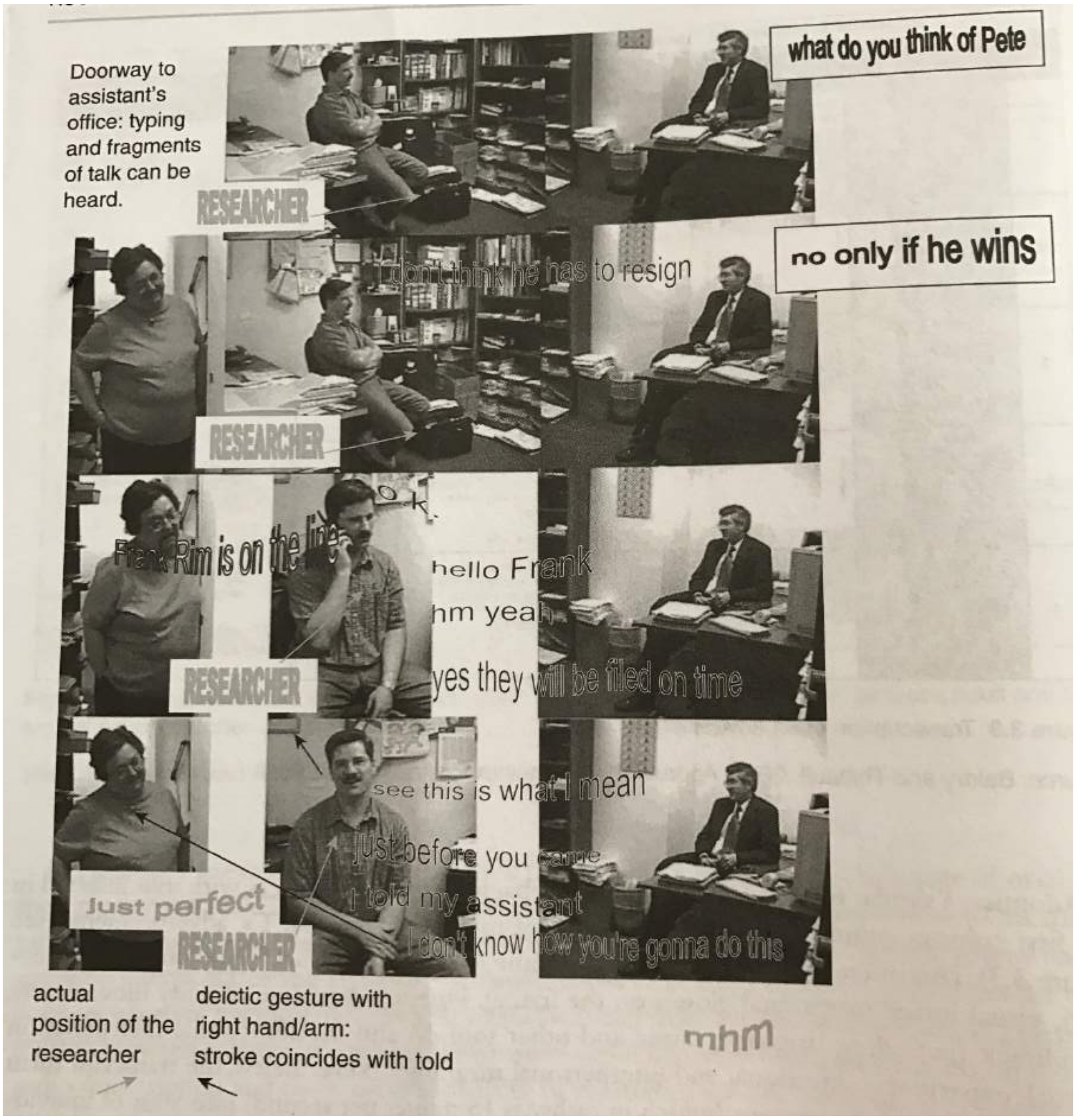
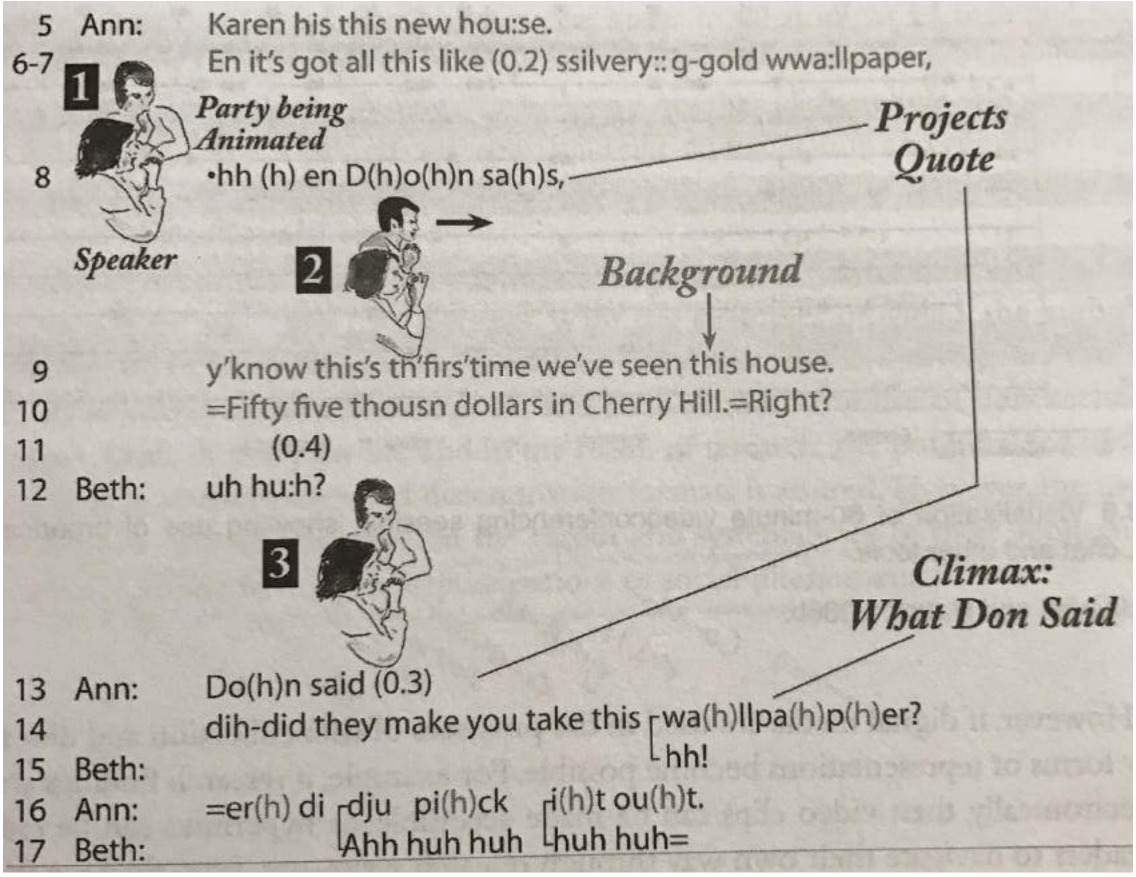

Doorway to assistant's office: typing and fragments of talk can be heard.
what do you think of Pete
RESEARCHER
no only if he wins
I don't think he has to resign
RESEARCHER
Frank Rim is on the line
o k
hello Frank
hm yeah
yes they will be filed on time
RESEARCHER
see this is what I mean
just before you came
I told my assistant
I don't know how you're gonna do this
just perfect
RESEARCHER
actual position of the researcher
deictic gesture with right hand/arm: stroke coincides with told
mhm

Figure 6.3. Initial description: *Transcription of multiparty interaction.*

Flewitt's third transcription method shows some similarity to that in Figure 3.3, using a still-frame image based style with further information overlaid. We see again the text presented in a curved script representing intonation. The image has been constructed from multiple images so as to show where the researcher was standing. Language from different participants is separated as one is in text boxes. Again, much of the gesture and gaze is inferred solely through the still-frame images, and directional arrows indicate the direction of gesture movement. This transcription method has its benefits in showing lots of visual information, but it can be difficult to deconstruct. Considering again the potential for multimodal data to involve more participants, this style of transcription will be excessively complex as more information is added.



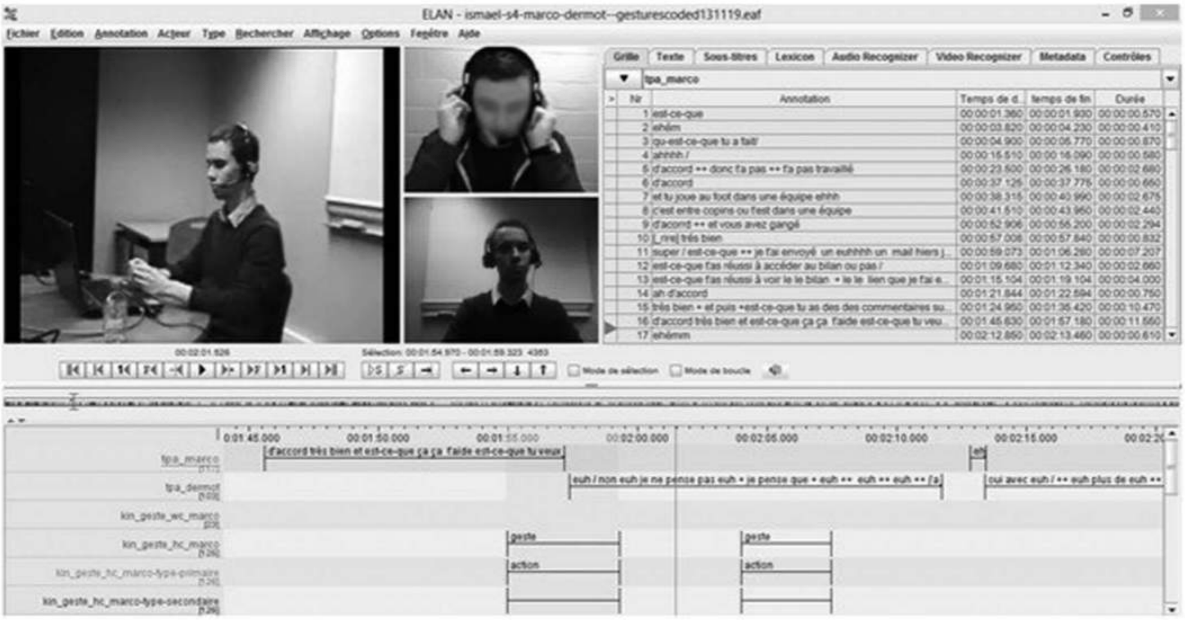
```
5     Ann:   Karen his this new hou:se.
6-7          En it's got all this like (0.2) ssilvery:: g-gold wwa:llpaper,
8            •hh (h) en D(h)o(h)n sa(h)s,
9            y'know this's th'firs'time we've seen this house.
10           =Fifty five thousn dollars in Cherry Hill.=Right?
11                (0.4)
12    Beth:  uh hu:h?
13    Ann:   Do(h)n said (0.3)
14           dih-did they make you take this ┌wa(h)llpa(h)p(h)er?
15    Beth:                                  └hh!
16    Ann:   =er(h) di ┌dju  pi(h)ck  ┌i(h)t ou(h)t.
17    Beth:            └Ahh huh huh   └huh huh=
```

1
Party being
Animated
Speaker
2
Background
3
Projects
Quote
Climax:
What Don Said

Figure 6.4. Initial description: *Drawings integrated into an extensive transcript of speech*.

Flewitt's final example transcription is based in conversation analysis techniques. It consists of a CA convention play-script style transcript with images integrated to encode further modalities. The images are a recreation of participants during the event. Directional arrows indicate gaze, and the images attempt to demonstrate their gestures. Due to the CA conventions used in the language transcript, intonation and stress are included within this transcription. Whilst the drawings are an acceptable medium, they are not nearly so accurate as still-frame images in transcription, and their placement disrupts the conventions of play-script style language transcription. This transcription method, therefore, would not be wholly effective in transcribing complex multimodal data.

Guichon, Nicolas and Wigham, Ciara – ‘A semiotic perspective on webconferencing-supported language teaching’

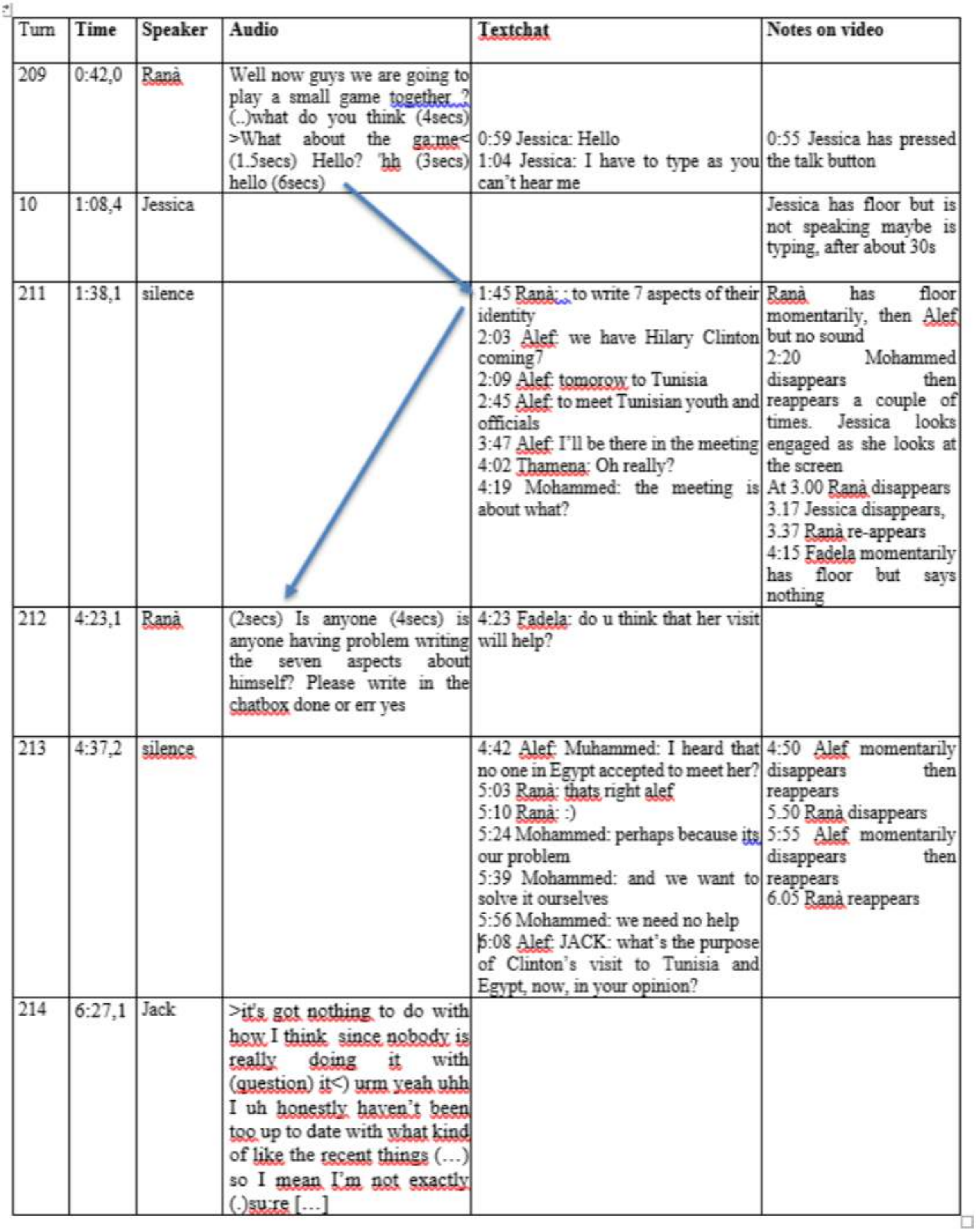

ELAN - ismael-s4-marco-dermot--gesturescoded131119.eaf
Fichier Edition Annotation Acteur Type Rechercher Affichage Options Fenêtre Aide
Grille Texte Sous-titres Lexicon Audio Recognizer Video Recognizer Metadata Contrôles
tpa_marco
00:02:01.526
Sélection: 00:01:54.970 - 00:01:59.323 4353
□ Mode de sélection □ Mode de boucle
0:01:45.000
00:01:50.000
00:01:55.000
00:02:00.000
00:02:05.000
00:02:10.000
00:02:15.000
tpa_marco
tpa_dermot
kin_geste_wc_marco
kin_geste_hc_marco
kin_geste_hc_marco-type-primaire
kin_geste_hc_marco-type-secondaire
d'accord très bien et est-ce-que ça ça t'aide est-ce-que tu veux
eh
euh / non euh je ne pense pas euh + je pense que + euh ++ euh ++ euh ++ j'a
oui avec euh / ++ euh plus de euh ++
geste
action

Nr	Annotation	Temps de d.	temps de fin	Durée
1	est-ce-que	00:00:01.360	00:00:01.930	00:00:00.570
2	ehèm	00:00:03.820	00:00:04.230	00:00:00.410
3	qu-est-ce-que tu a fait/	00:00:04.900	00:00:05.770	00:00:00.870
4	ahhhh /	00:00:15.510	00:00:16.090	00:00:00.580
5	d'accord ++ donc t'a pas ++ t'a pas travaillé	00:00:23.500	00:00:26.180	00:00:02.680
6	d'accord	00:00:37.125	00:00:37.775	00:00:00.650
7	et tu joue au foot dans une équipe ehhh	00:00:38.315	00:00:40.990	00:00:02.675
8	c'est entre copins ou t'est dans une équipe	00:00:41.510	00:00:43.950	00:00:02.440
9	d'acord ++ et vous avez gangé	00:00:52.906	00:00:55.200	00:00:02.294
10	[_rire] très bien	00:00:57.008	00:00:57.840	00:00:00.832
11	super / est-ce-que ++ je t'ai envoyé un euhhhh un mail hiers j...	00:00:59.073	00:01:06.280	00:00:07.207
12	est-ce-que t'as réussi à accéder au bilan ou pas /	00:01:09.680	00:01:12.340	00:00:02.660
13	est-ce-que t'as réussi à voir le le bilan + le le lien que je t'ai e...	00:01:15.104	00:01:19.104	00:00:04.000
14	ah d'accord	00:01:21.844	00:01:22.594	00:00:00.750
15	très bien + et puis +est-ce-que tu as des des commentaires su...	00:01:24.950	00:01:35.420	00:00:10.470
16	d'accord très bien et est-ce-que ça ça t'aide est-ce-que tu veu...	00:01:45.630	00:01:57.180	00:00:11.550
17	ehèmm	00:02:12.850	00:02:13.460	00:00:00.610

Figure 7.1. Initial description: *Alignment of trainee hors-champ video with trainee and student webcam videos in ELAN.*

Guichon and Wigham demonstrate a transcription using the ELAN software to match modalities such as speech and gesture to an audiovisual text along a temporal axis. Bars are used along these horizontal axis to show the duration of gestures and speech acts. This is less a method of transcription so much as an example of how data is configured by the ELAN software. Thus, this transcription is effective in that is shows the potential for using ELAN to process multimodal data.

Helm, Francesca and Dooly, Melinda – Challenges in Transcribing Multimodal Data: A Case Study

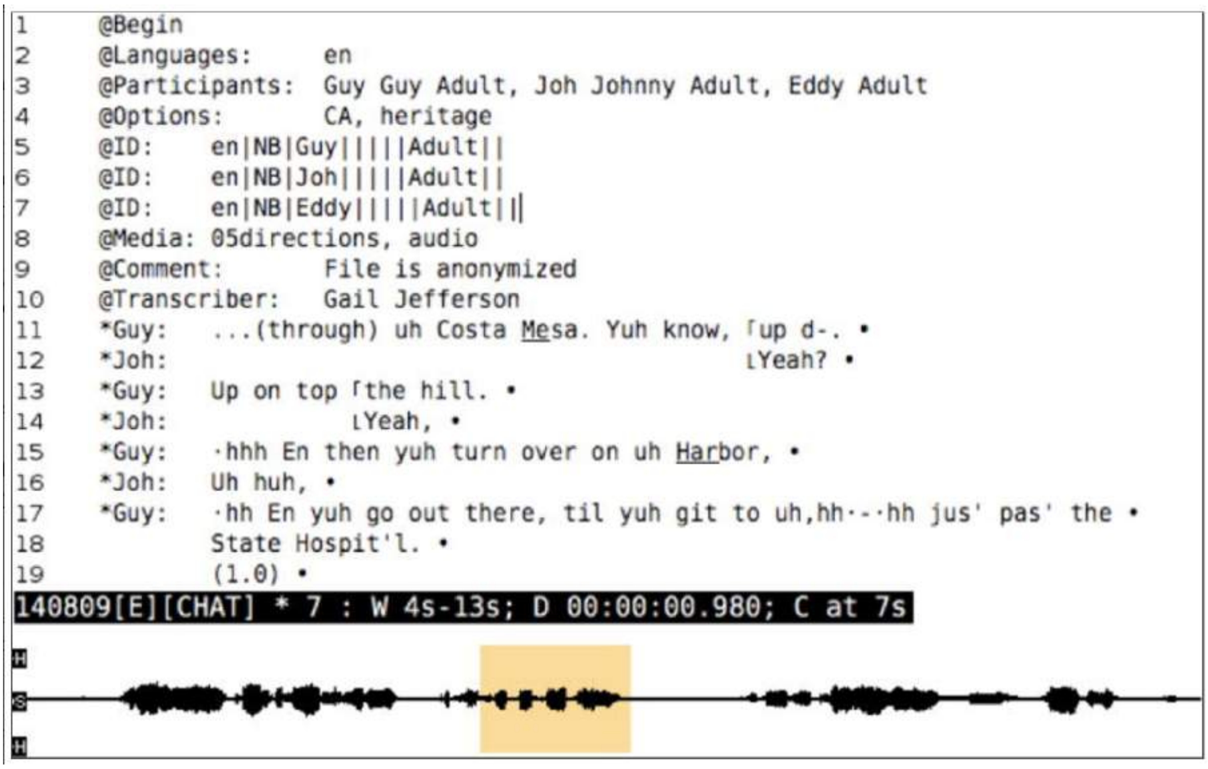


Turn	Time	Speaker	Audio	Textchat	Notes on video
209	0:42,0	Ranà	Well now guys we are going to play a small game together ? (..)what do you think (4secs) >What about the ga:me< (1.5secs) Hello? 'hh (3secs) hello (6secs)	0:59 Jessica: Hello 1:04 Jessica: I have to type as you can't hear me	0:55 Jessica has pressed the talk button
10	1:08,4	Jessica			Jessica has floor but is not speaking maybe is typing, after about 30s
211	1:38,1	silence		1:45 Ranà: : to write 7 aspects of their identity 2:03 Alef: we have Hilary Clinton coming7 2:09 Alef: tomorow to Tunisia 2:45 Alef: to meet Tunisian youth and officials 3:47 Alef: I'll be there in the meeting 4:02 Thamena: Oh really? 4:19 Mohammed: the meeting is about what?	Ranà has floor momentarily, then Alef but no sound 2:20 Mohammed disappears then reappears a couple of times. Jessica looks engaged as she looks at the screen At 3.00 Ranà disappears 3.17 Jessica disappears, 3.37 Ranà re-appears 4:15 Fadela momentarily has floor but says nothing
212	4:23,1	Ranà	(2secs) Is anyone (4secs) is anyone having problem writing the seven aspects about himself? Please write in the chatbox done or err yes	4:23 Fadela: do u think that her visit will help?	
213	4:37,2	silence		4:42 Alef: Muhammed: I heard that no one in Egypt accepted to meet her? 5:03 Ranà: thats right alef 5:10 Ranà: :) 5:24 Mohammed: perhaps because its our problem 5:39 Mohammed: and we want to solve it ourselves 5:56 Mohammed: we need no help 6:08 Alef: JACK: what's the purpose of Clinton's visit to Tunisia and Egypt, now, in your opinion?	4:50 Alef momentarily disappears then reappears 5.50 Ranà disappears 5:55 Alef momentarily disappears then reappears 6.05 Ranà reappears
214	6:27,1	Jack	>it's got nothing to do with how I think since nobody is really doing it with (question) it<) urm yeah uhh I uh honestly haven't been too up to date with what kind of like the recent things (…) so I mean I'm not exactly (.)su:re […]		

Figure 8.1. Initial description: *Working multimodal transcription of Session 1 Turns 209-214.*

This transcript was produced specifically for data from a videoconferencing program called 'Soliya', thus its purpose is restricted to the modalities of that platform. Nevertheless, in this transcription we see a fairly conventional tabular layout, separating different actions into turns and allocating each a timestamp. The modalities engaged through Soliya are a webcam, including audible speech, and a text chat box. However, the transcription only handles the recorded speech and text chat rather than gesture or gaze from the webcam video. Thus, whilst this transcript handles verbalised language and text language adequately, it does not seek to provide a complete transcription. The transcription serves its purpose, but we will learn more from one of the more comprehensive tabular layouts exemplified.

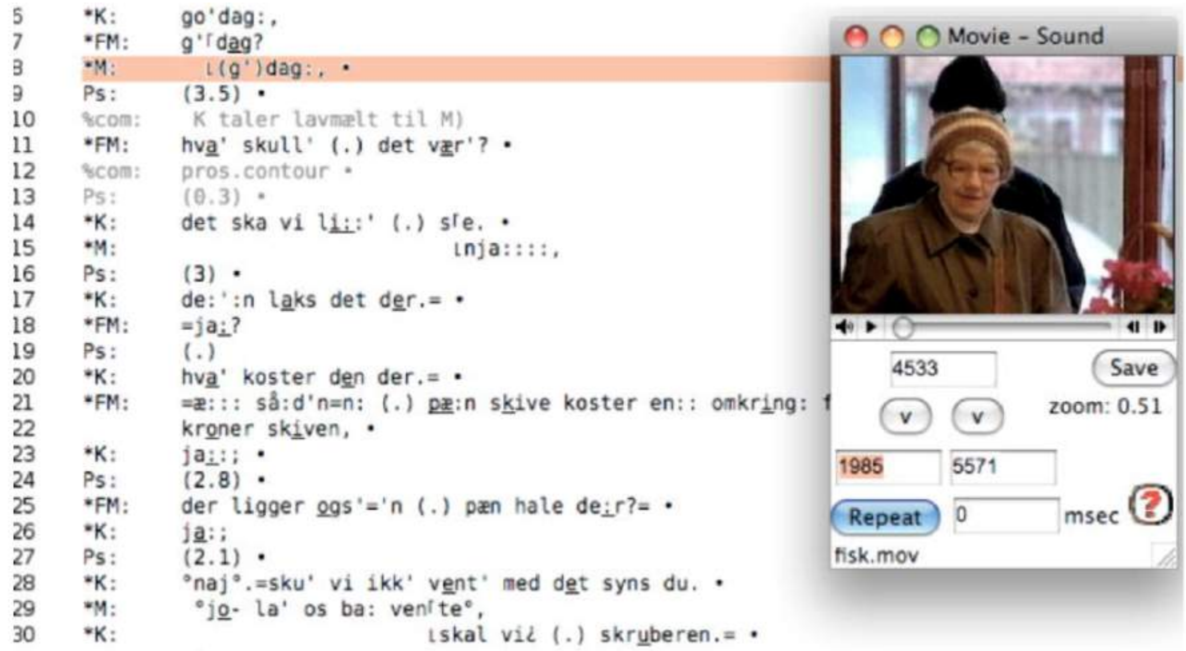
MacWhinney, Brian – ‘Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository



```
@Begin
@Languages:     en
@Participants:  Guy Guy Adult, Joh Johnny Adult, Eddy Adult
@Options:       CA, heritage
@ID:    en|NB|Guy||||Adult||
@ID:    en|NB|Joh||||Adult||
@ID:    en|NB|Eddy||||Adult||
@Media: 05directions, audio
@Comment:       File is anonymized
@Transcriber:   Gail Jefferson
*Guy:   ...(through) uh Costa Mesa. Yuh know, ⌈up d-. •
*Joh:                                        ⌊Yeah? •
*Guy:   Up on top ⌈the hill. •
*Joh:             ⌊Yeah, •
*Guy:   ·hhh En then yuh turn over on uh Harbor, •
*Joh:   Uh huh, •
*Guy:   ·hh En yuh go out there, til yuh git to uh,hh··hh jus' pas' the •
        State Hospit'l. •
        (1.0) •
140809[E][CHAT] * 7 : W 4s-13s; D 00:00:00.980; C at 7s
```

Figure 9.1. Initial description: *Audio Transcription from a Waveform*

MacWhinney first demonstrates how CLAN can present an audio transcription and match the transcription to the waveform of the original audio as it plays. This can allow for easier identification of parts of the transcription within the whole original text. However, this method of transcription only deals with the verbal modality, so it is insufficient for multimodal analysis.

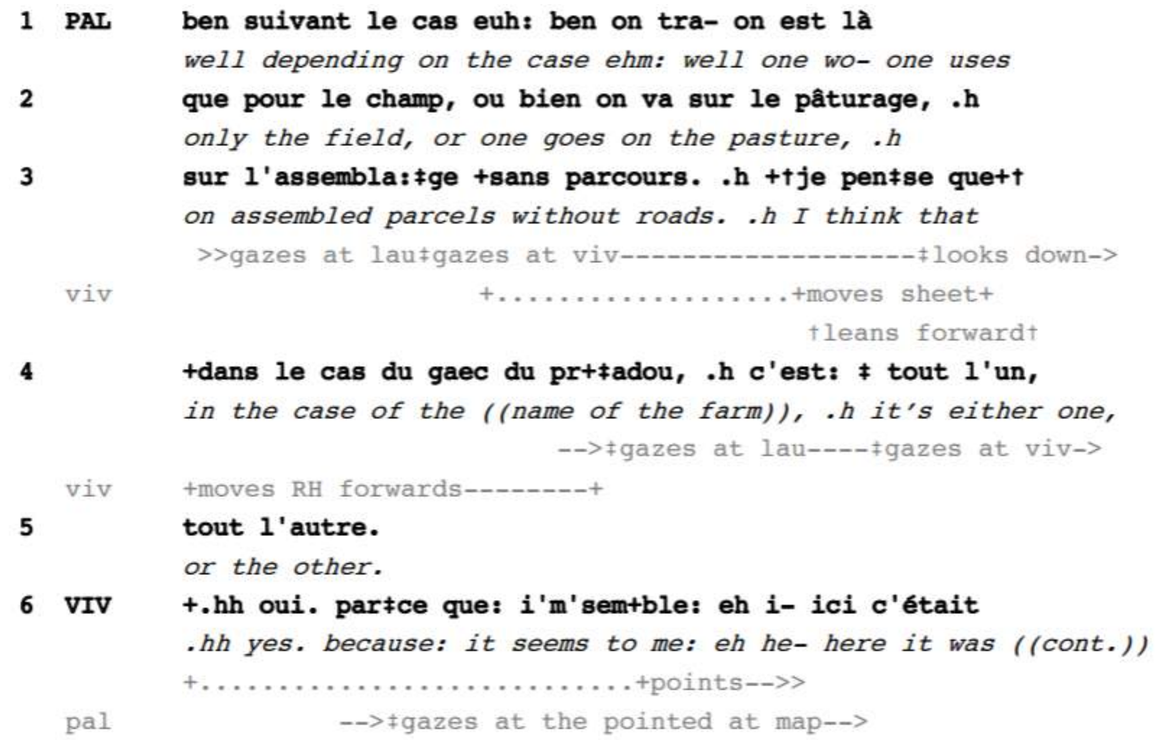


```
*K:      go'dag:,
*FM:     g'⌈dag?
*M:        ⌊(g')dag:, •
Ps:      (3.5) •
%com:     K taler lavmælt til M)
*FM:     hva' skull' (.) det vær'? •
%com:    pros.contour •
Ps:      (0.3) •
*K:      det ska vi li::' (.) s⌈e. •
*M:                           ⌊nja::::,
Ps:      (3) •
*K:      de:':n laks det der.= •
*FM:     =ja:?
Ps:      (.)
*K:      hva' koster den der.= •
*FM:     =æ::: så:d'n=n: (.) pæ:n skive koster en:: omkring: f
         kroner skiven, •
*K:      ja::; •
Ps:      (2.8) •
*FM:     der ligger ogs'='n (.) pæn hale de:r?= •
*K:      ja::
Ps:      (2.1) •
*K:      °naj°.=sku' vi ikk' vent' med det syns du. •
*M:       °jo- la' os ba: ven⌈te°,
*K:                         ⌊skal vi¿ (.) skruberen.= •
```

Figure 9.2. Initial description: *Video transcription*

MacWhinney also demonstrates how CLAN can link a transcript to the video it came from, highlighting the transcribed text as it is spoken in the video. Only verbal speech is transcribed through CA conventions, and all other information is represented by the original multimodal data being displayed. Therefore, this is not a complete transcription, but a means of linking some transcribed parts of the data temporally to the media itself. An effective transcription method for communicational cues to arrogance will need to encode data such as gesture and gaze specifically, rather than rely on its demonstration through the original media.

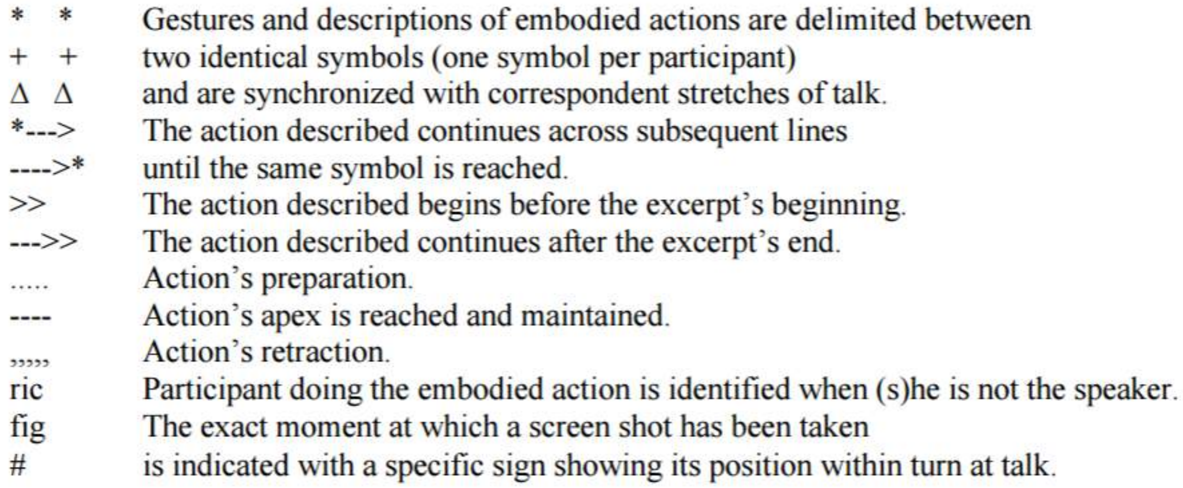
Mondada, Lorenza – ‘Covnentions for multimodal transcription’



```
1  PAL    ben suivant le cas euh: ben on tra- on est là
          well depending on the case ehm: well one wo- one uses
2         que pour le champ, ou bien on va sur le pâturage, .h
          only the field, or one goes on the pasture, .h
3         sur l'assembla:‡ge +sans parcours. .h +†je pen‡se que+†
          on assembled parcels without roads. .h I think that
           >>gazes at lau‡gazes at viv-------------------‡looks down->
   viv                        +..................+moves sheet+
                                                   †leans forward†
4         +dans le cas du gaec du pr+‡adou, .h c'est: ‡ tout l'un,
          in the case of the ((name of the farm)), .h it's either one,
                                -->‡gazes at lau----‡gazes at viv->
   viv    +moves RH forwards--------+
5         tout l'autre.
          or the other.
6  VIV    +.hh oui. par‡ce que: i'm'sem+ble: eh i- ici c'était
          .hh yes. because: it seems to me: eh he- here it was ((cont.))
          +...........................+points-->>
   pal              -->‡gazes at the pointed at map-->
```

Figure 10.1. Initial description: *Example*

Through her paper, Mondada introduces new conventions for multimodal transcription, and refines her transcription method as she continues. This transcription is the best example of her final methodology. Mondada provides a key of the textual symbols she uses, as well as further explanation of how these work flexibly within a larger transcription.

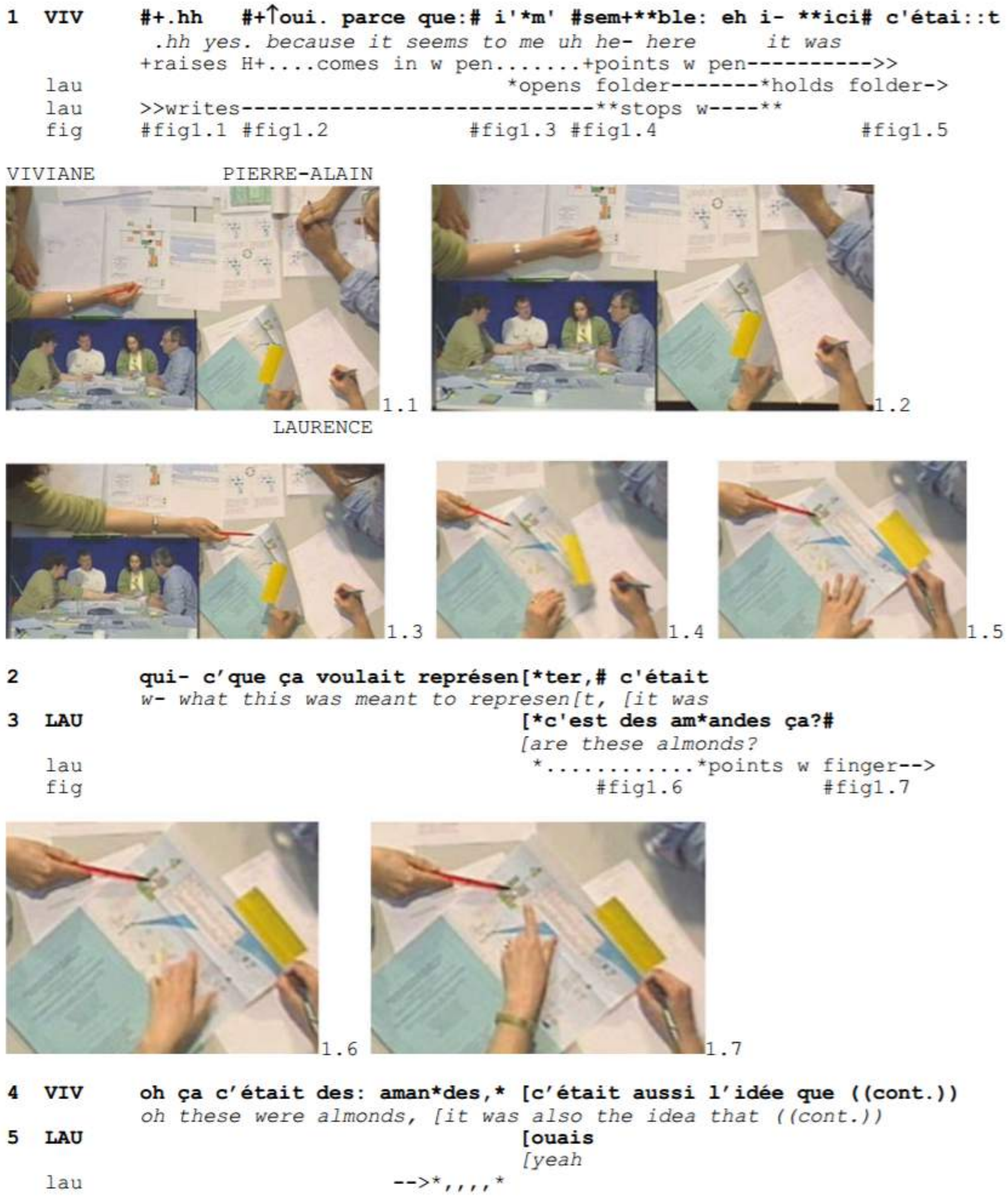


```
*   *     Gestures and descriptions of embodied actions are delimited between
+   +     two identical symbols (one symbol per participant)
Δ   Δ     and are synchronized with correspondent stretches of talk.
*--->     The action described continues across subsequent lines
---->*    until the same 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 when (s)he is not the speaker.
fig       The exact moment at which a screen shot has been taken
#         is indicated with a specific sign showing its position within turn at talk.
```

Figure 10.2. Initial description: *No initial description.*

Understanding Mondada’s transcription method requires dedicating time to understanding her new conventions, but it does go a long way to providing a thorough transcription method. Both gaze and gesture actions can have initiation, length and retraction encoded. Although it could become very complex with more participants, Mondada makes a strong case for using play-script style transcription to encode all the information required.

Mondada, Lorenza – ‘Challenges of multimodality: Language and the body in social interaction’



```
1  VIV    #+.hh   #+↑oui. parce que:# i'*m' #sem+**ble: eh i- **ici# c'étai::t
           .hh yes. because it seems to me uh he- here     it was
          +raises H+....comes in w pen.......+points w pen---------->>
   lau                                *opens folder-------*holds folder->
   lau    >>writes---------------------------**stops w----**
   fig    #fig1.1 #fig1.2           #fig1.3 #fig1.4                #fig1.5
```

VIVIANE
PIERRE-ALAIN
1.1
1.2
LAURENCE
1.3
1.4
1.5

```
2         qui- c'que ça voulait représen[*ter,# c'était
          w- what this was meant to represen[t, [it was
3  LAU                                  [*c'est des am*andes ça?#
                                        [are these almonds?
   lau                                   *...........*points w finger-->
   fig                                       #fig1.6          #fig1.7
```

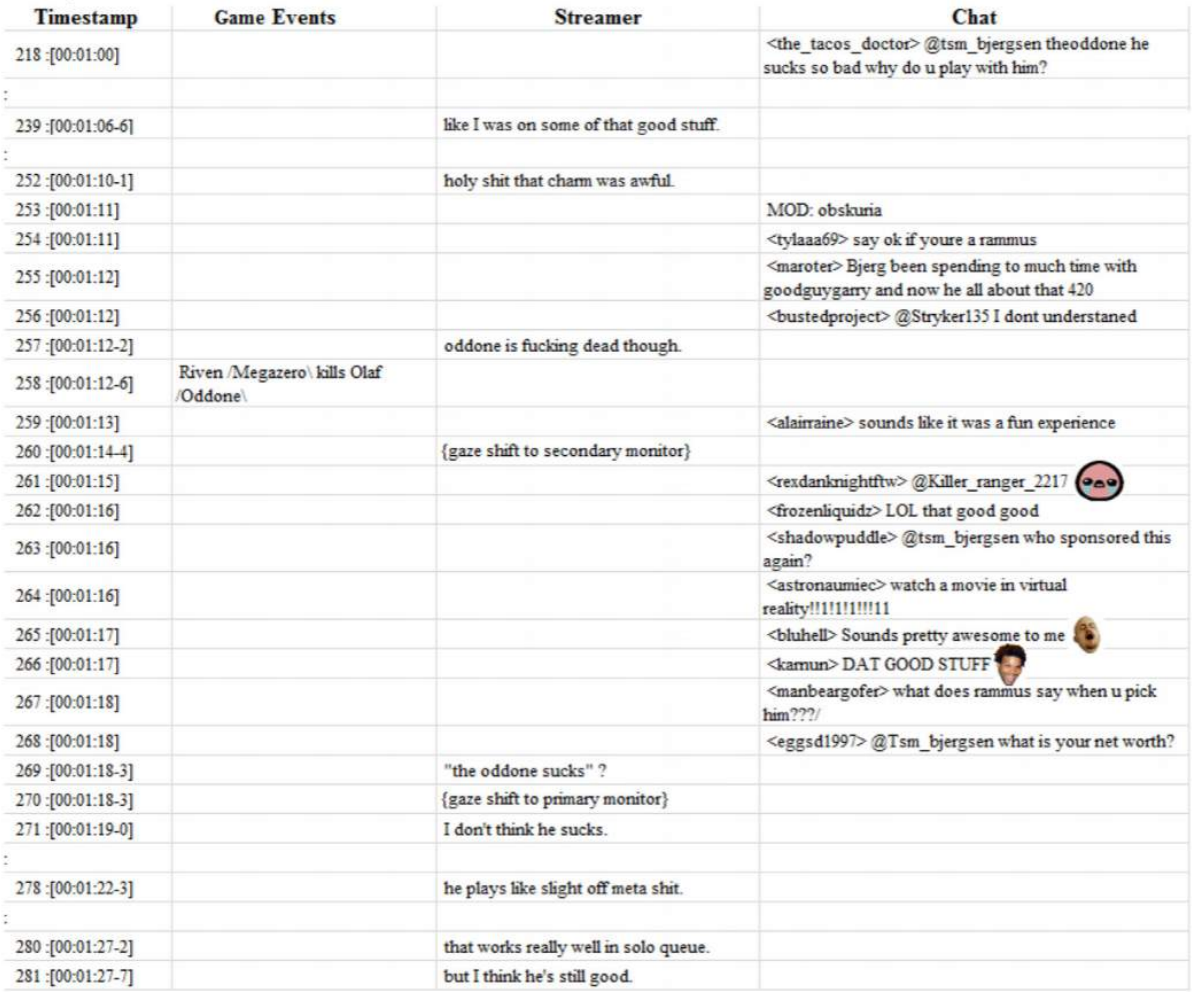
1.6
1.7

```
4  VIV    oh ça c'était des: aman*des,* [c'était aussi l'idée que ((cont.))
          oh these were almonds, [it was also the idea that ((cont.))
5  LAU                                  [ouais
                                        [yeah
   lau                         -->*,,,,*
```

Figure 11.1. Initial description: *No initial description*.

In this more recent paper, Mondada demonstrates her conventionalised multimodal transcription method with the addition of still-frame images. This is only a small adaption from Figure 10.1, using captured images to further exemplify the gestures and gazes she described linguistically within the transcript. This is an improvement insofar as it removes some of the ambiguity in describing gaze and gesture linguistically. However, this transcription method retains the same downsides as its initial demonstration: the conventions are complex, and it could definitely become difficult to handle with more participants involved.

Recktenwald, David – ‘Toward a transcription and analysis of live streaming on *Twitch*’

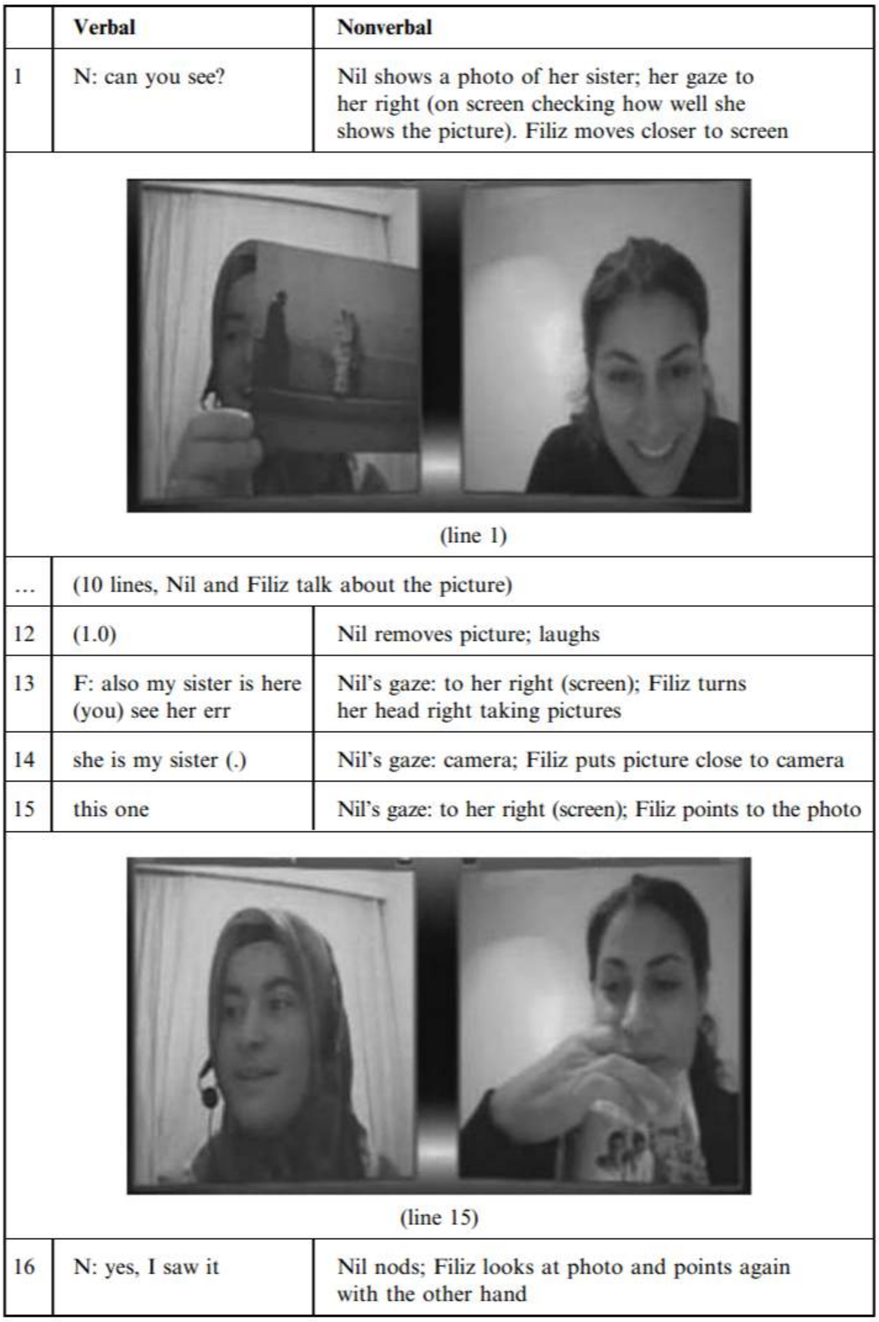


Timestamp	Game Events	Streamer	Chat
218 :[00:01:00]			<the_tacos_doctor> @tsm_bjergsen theoddone he sucks so bad why do u play with him?
:			
239 :[00:01:06-6]		like I was on some of that good stuff.	
:			
252 :[00:01:10-1]		holy shit that charm was awful.	
253 :[00:01:11]			MOD: obskuria
254 :[00:01:11]			<tylaaa69> say ok if youre a rammus
255 :[00:01:12]			<maroter> Bjerg been spending to much time with goodguygarry and now he all about that 420
256 :[00:01:12]			<bustedproject> @Stryker135 I dont understaned
257 :[00:01:12-2]		oddone is fucking dead though.	
258 :[00:01:12-6]	Riven /Megazero\ kills Olaf /Oddone\		
259 :[00:01:13]			<alairraine> sounds like it was a fun experience
260 :[00:01:14-4]		{gaze shift to secondary monitor}	
261 :[00:01:15]			<rexdanknightftw> @Killer_ranger_2217
262 :[00:01:16]			<frozenliquidz> LOL that good good
263 :[00:01:16]			<shadowpuddle> @tsm_bjergsen who sponsored this again?
264 :[00:01:16]			<astronaumiec> watch a movie in virtual reality!!1!1!1!!!11
265 :[00:01:17]			<bluhell> Sounds pretty awesome to me
266 :[00:01:17]			<kamun> DAT GOOD STUFF
267 :[00:01:18]			<manbeargofer> what does rammus say when u pick him???/
268 :[00:01:18]			<eggsd1997> @Tsm_bjergsen what is your net worth?
269 :[00:01:18-3]		"the oddone sucks" ?	
270 :[00:01:18-3]		{gaze shift to primary monitor}	
271 :[00:01:19-0]		I don't think he sucks.	
:			
278 :[00:01:22-3]		he plays like slight off meta shit.	
:			
280 :[00:01:27-2]		that works really well in solo queue.	
281 :[00:01:27-7]		but I think he's still good.	

Figure 12.1. Initial description: *The Oddone sucks?*

Recktenwald’s transcription method seeks to transcribe multimodal data from the online streaming platform twitch, which shows a game being played, a webcam on the streamer (playing the game) including their speech, and a text chat box for people watching the stream. Recktenwald’s method, therefore, is surprisingly minimalistic. Events are split into turns each with a timestamp, although the transcription is only thirty seconds long. The nature of many people interacting through the text chat box means many turns can occur within a very short time. Gaze and gesture are obviously not ranked as particularly important within this transcript, as they are placed in a column named ‘Streamer’ which includes all of the streamer’s verbal communication, gaze and gesture behaviour. This transcription serves to reinforce the conventions of the tabular transcription style, but further than this it offers little ingenuity in terms of multimodal transcription.

Satar, Müge – ‘Multimodal language learner interactions via desktop videoconferencing within a framework of social presence: Gaze



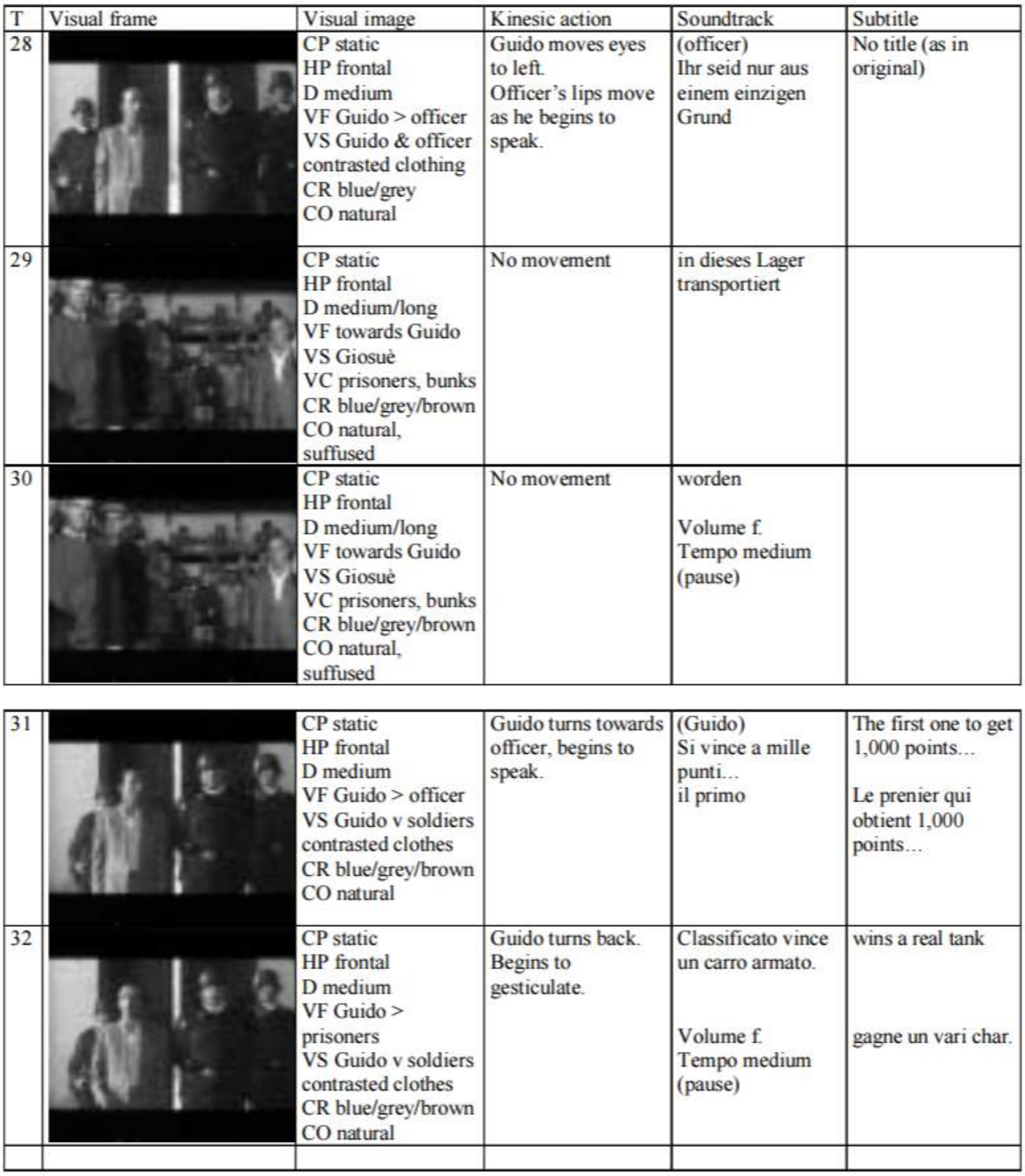
	Verbal	Nonverbal
1	N: can you see?	Nil shows a photo of her sister; her gaze to her right (on screen checking how well she shows the picture). Filiz moves closer to screen
	(line 1)	
…	(10 lines, Nil and Filiz talk about the picture)	
12	(1.0)	Nil removes picture; laughs
13	F: also my sister is here (you) see her err	Nil's gaze: to her right (screen); Filiz turns her head right taking pictures
14	she is my sister (.)	Nil's gaze: camera; Filiz puts picture close to camera
15	this one	Nil's gaze: to her right (screen); Filiz points to the photo
	(line 15)	
16	N: yes, I saw it	Nil nods; Filiz looks at photo and points again with the other hand

Figure 13.1. Initial description:

Can you see?

Satar’s paper focuses expressly on gaze in videoconferencing environments. Whilst this leads to an incomplete transcription method in terms of gesture, and the finer qualities of verbalisation like intonation, very few papers consider gaze in such detail. Thus, Satar’s paper contains a great deal of information regarding gaze that may prove invaluable for transcription within a more comprehensive method. The transcript sets verbal language alongside a collective ‘Nonverbal’ column, combined with still-frame images to exemplify the gaze movements. Whilst this example of transcription is none too extensive, information elsewhere in the paper regarding gaze times will prove valuable in future transcription.

Taylor, Christopher – 'Multimodal Transcription in the Analysis, Translation and Subtitling of Italian Films'

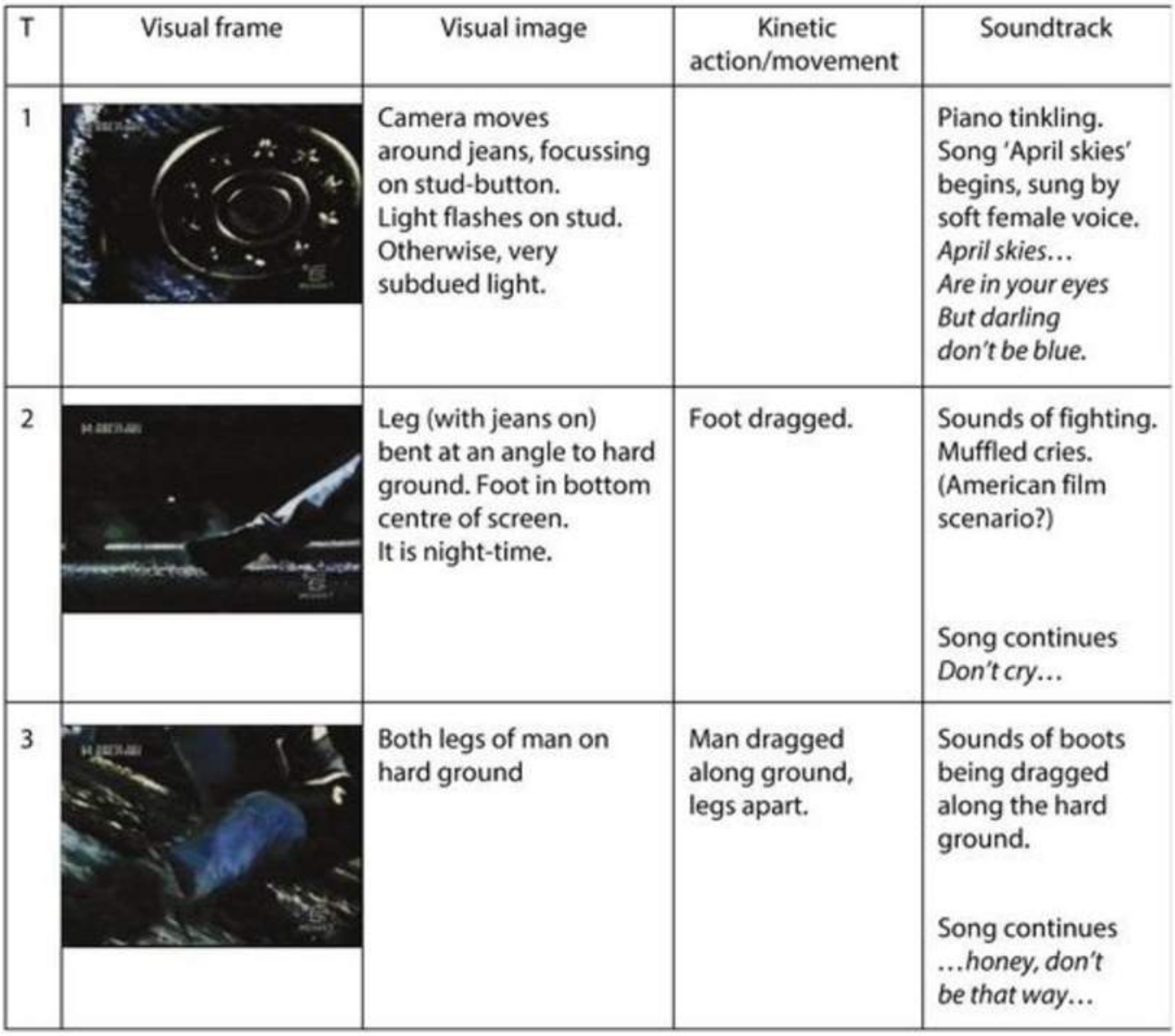


T	Visual frame	Visual image	Kinesic action	Soundtrack	Subtitle
28		CP static HP frontal D medium VF Guido > officer VS Guido & officer contrasted clothing CR blue/grey CO natural	Guido moves eyes to left. Officer's lips move as he begins to speak.	(officer) Ihr seid nur aus einem einzigen Grund	No title (as in original)
29		CP static HP frontal D medium/long VF towards Guido VS Giosuè VC prisoners, bunks CR blue/grey/brown CO natural, suffused	No movement	in dieses Lager transportiert	
30		CP static HP frontal D medium/long VF towards Guido VS Giosuè VC prisoners, bunks CR blue/grey/brown CO natural, suffused	No movement	worden Volume f. Tempo medium (pause)	
31		CP static HP frontal D medium VF Guido > officer VS Guido v soldiers contrasted clothes CR blue/grey/brown CO natural	Guido turns towards officer, begins to speak.	(Guido) Si vince a mille punti... il primo	The first one to get 1,000 points... Le prenier qui obtient 1,000 points...
32		CP static HP frontal D medium VF Guido > prisoners VS Guido v soldiers contrasted clothes CR blue/grey/brown CO natural	Guido turns back. Begins to gesticulate.	Classificato vince un carro armato. Volume f. Tempo medium (pause)	wins a real tank gagne un vari char.

Figure 14.1. Initial description: *Multimodal transcription of scene from La vita è bella*

In this transcription, Taylor seeks to transcribe from an Italian film, including subtitles. The visual image column of this transcription is particularly interesting, as it provides a great deal of detail regarding the composition of the camera shot. This is something that could be useful in future multimodal transcription. Other than this, however, there is little unique about this particular example of tabular transcription.

Taylor, Christopher – 'Multimodality and audiovisual translation'

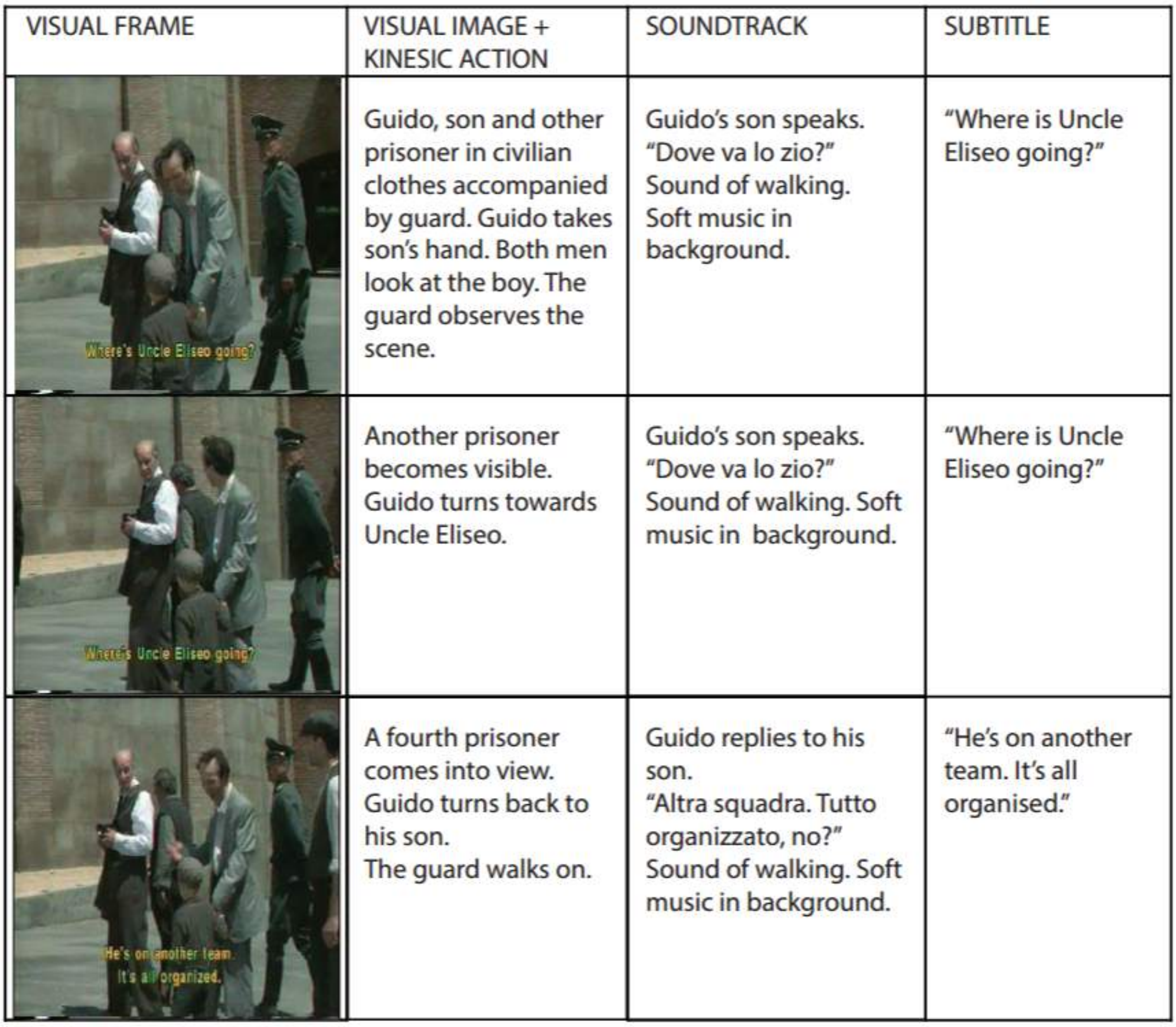


T	Visual frame	Visual image	Kinetic action/movement	Soundtrack
1		Camera moves around jeans, focussing on stud-button. Light flashes on stud. Otherwise, very subdued light.		Piano tinkling. Song 'April skies' begins, sung by soft female voice. *April skies…* *Are in your eyes* *But darling* *don't be blue.*
2		Leg (with jeans on) bent at an angle to hard ground. Foot in bottom centre of screen. It is night-time.	Foot dragged.	Sounds of fighting. Muffled cries. (American film scenario?) Song continues *Don't cry…*
3		Both legs of man on hard ground	Man dragged along ground, legs apart.	Sounds of boots being dragged along the hard ground. Song continues *…honey, don't be that way…*

Figure 15.1. Initial description: *Mutimodal transcription*

Taylor offers a simplistic tabular multimodal transcription method in this chapter regarding multimodal transcription. Taylor's chapter within the book is something of an introduction to multimodality and the issues with its transcription, so the method he presents is somewhat elementary. It is a fairly conventional tabular layout with the still-frame image being given prominence on the leftmost column (save for the turn number). Other than this, information is described linguistically, considering the visual image, action within the video and backing sounds. This transcription method is simple, but effective. A more extensive tabular layout would be required for a comprehensive multimodal transcription method, however.

Taylor, Christopher – 'The multimodal approach in audiovisual translation'



VISUAL FRAME	VISUAL IMAGE + KINESIC ACTION	SOUNDTRACK	SUBTITLE
	Guido, son and other prisoner in civilian clothes accompanied by guard. Guido takes son's hand. Both men look at the boy. The guard observes the scene.	Guido's son speaks. "Dove va lo zio?" Sound of walking. Soft music in background.	"Where is Uncle Eliseo going?"
	Another prisoner becomes visible. Guido turns towards Uncle Eliseo.	Guido's son speaks. "Dove va lo zio?" Sound of walking. Soft music in background.	"Where is Uncle Eliseo going?"
	A fourth prisoner comes into view. Guido turns back to his son. The guard walks on.	Guido replies to his son. "Altra squadra. Tutto organizzato, no?" Sound of walking. Soft music in background.	"He's on another team. It's all organised."

Figure 16.1. Initial description: *Multimodal transcription from La vita è bella*

In this more recent text, Taylor's tabular multimodal transcription layout is seen to have undergone little change. In this instance, Taylor is also transcribing subtitles from the still-frame images. Otherwise, the description of the visual image and action therein, and soundtrack columns remain. This transcription method supports Taylor's purposes, but will still require expansion to support more dynamic multimodal data. Thus, this transcription method may be useful as a reference point within a larger collection of tabular layouts, but is not sufficient alone.