Multimodal Transcription as Academic Practice: A Social Semiotic Perspective

Jeff Bezemer (Imperial College London, j.bezemer@imperial.ac.uk) & Diane Mavers (Institute of Education, University of London, d.mavers@ioe.ac.uk)

Abstract
With the increasing use of video recording in social research methodological questions about multimodal transcription are more timely than ever before. How do researchers transcribe gesture, for instance, or gaze, and how can they show to readers of their transcripts how such modes operate in social interaction alongside speech? Should researchers bother transcribing these modes of communication at all? How do they define a ‘good’ transcript? In this paper we begin to develop a social semiotic framework to account for transcripts as artefacts, treating them as empirical material through which transcription as a social, meaning making practice can be reconstructed. We look at some multimodal transcripts produced in conversation analysis, discourse analysis, social semiotics and micro-ethnography, drawing attention to the meaning-making principles applied by the transcribers. We argue that there are significant representational differences between multimodal transcripts, reflecting differences in the professional practices and the rhetorical and analytical purposes of their makers.

Introduction
Transcription is a common academic practice. In social research investigating language and communication, a ‘transcript’ usually refers to a distinctive genre associated with turning a strip of ‘naturally’ occurring talk – e.g. a job interview, a conversation at the dinner table – into writing. This genre has analytical as well as rhetorical purposes: to develop insights into the moment-by-moment and in situ construction of social reality and to provide evidence in developing an argument for an academic audience. With the increasing use of video recording in social research, methodological questions about multimodal transcription are more timely than ever before. How do researchers transcribe gesture, for instance, or gaze, and how can they show to readers of their transcripts how such modes operate alongside speech? Should researchers bother transcribing these modes of communication at all? What are the epistemological implications of choices of inclusion and exclusion? What does one gain from inclusion of modes other than speech if the aim of transcription is to focus on a selection of the vast amount of data collected? In this paper we investigate how some researchers, including ourselves, have dealt with these issues. We discuss the emergence of multimodal transcripts in social research, review the theoretical perspectives on transcription adopted in conversation analysis and linguistics and develop our own, social semiotic take. We then use this framework to analyse and compare a selection of multimodal transcripts which have appeared in recent academic publications.

The emergence of multimodal transcripts
Transcripts are widely produced and used in a range of disciplines, including linguistics, sociolinguistics, pragmatics, linguistic anthropology, conversation analysis, and discourse analysis. In some of these disciplines the transcript has been a relatively stable and fixed genre for several decades. In conversation analysis, Gail Jefferson’s transcription conventions, first published in 1974 (Sacks, Schegloff & Jefferson, 1974), have since been
widely adopted, and similar, standardized conventions can be found in other disciplines. For instance, McWhinney’s (2000) transcription system is widely used in applied linguistics. Researchers have subsequently introduced new forms of representation to the genre, such as line drawings (McDermott, Gospodinoff & Aron, 1978), and stills from video footage (Heath, Hindmarsh & Luff, 2010), variously set out on the page, integrated with writing and overlaid with other graphic features such as arrows (Norris, 2004) and musical notation (Erickson, 2004), and invariably accompanied by a written commentary explaining what is made visible. No longer necessarily ‘scripts’, nor exclusively ‘visuals’, these ‘transcripts’ (or ‘transvisuals’) are diverse and flexible. There is multiplicity in the practices not only between researchers, but also in the work of individual researchers themselves as they make transcripts for different audiences and analytical purposes. These changes towards visualisation and variability of the transcript coincide with a growing interest in the ‘multimodality’ of communication and a felt need to account for social practices through a lens that attends to the multiplicity of how people communicate; and with a weakening of disciplinary boundaries, notably between ethnography and the discourse-oriented disciplines.

Audio recordings have been widely used for many decades and, whilst video recorders have been available almost throughout the same period, it is only now that their use is becoming more standard practice (see Erickson, this issue, for an historical overview of the use of video in research). With increasing availability and use of digital video technology, the issue of how to re-present multimodal interaction has become a key issue not only for disciplines concerned with language and communication, but for video-based social research more generally (e.g. Pink, 2001; Flewitt, 2006; Dicks et al., 2006; Kissmann, 2009). Where formerly observational researchers would produce and analyse field notes, they are now often faced with the task of producing and analysing an audiovisual record. This has far-reaching consequences, one of which is the construction and use of the multimodal transcript, and this significantly reshapes the presentation of academic discourse as well as its accounts of social interaction. In the following sections we aim to begin to identify, document and analyse these methodological and epistemological implications in terms of the shift from predominantly written accounts of social interaction to accounts in which the visual features equally prominently.

Transcription as theory
Whilst (multimodal) transcripts are widely and increasingly employed, the study of their construction and use as a socially and culturally organized practice of knowledge production is limited. Linguists, discourse analysts and conversation analysts have reflected on their own transcription practices, and they have produced a small body of literature on the subject, focusing largely on the transcription of speech (for a review of this literature see Davidson, 2009). However these reflections are typically only articulated outside the publications in which the transcripts first appeared, in separate publications ‘on transcription’, such as the present paper. Rarely do researchers systematically analyse a range of different transcripts, i.e., a corpus of transcripts, comparing them in methodological and theoretical terms.

While transcription is still under-researched in studies on language and communication, it is not even acknowledged as a topic of interest or concern in many qualitative research handbooks (Davidson, 2009). This is related to epistemological differences. Researchers adopting a positivist stance believe that a transcript can objectively represent speech (and by implication other forms of embodied action). Those adopting a constructionist stance have aimed to demonstrate that transcription, like any form of representation, is shaped by ‘theory’ (Ochs, 1979) and ‘politics’ (Bucholtz, 2000). It is also related to differences in ‘professional
vision’ (Goodwin, 1994); scholars who define themselves in terms of an interest in ‘talk’ (such as conversation analysts) are likely to see the differences between speech and writing more sharply, and develop resources that help them make sense of those differences; and those with in an interest in social inequality (such as sociolinguists) are more likely to see the ideological implications of the representation of, say, accent. Their interests shape their reflection on the ‘aptness’ of their representations of speech and on their own role and responsibility in the production and reproduction of social and cultural differences. Thus, perhaps unsurprisingly, transcription has been problematized by scholars embracing constructionist and critical approaches to doing research, whilst is has been treated as a relatively ‘straightforward’ intermediary step in other approaches to qualitative research, for instance approaches derived from grounded theory. For instance, a common approach to analysing interviews is to have them transcribed by ‘transcribers’, and subsequently identify themes and categories across the different transcripts.

In the remainder of this paper we aim to contribute to the methodological debate on transcription in a number of ways. First, we focus on multimodal transcription, not only the transcription of speech. Second, building on constructionist notions of transcription we begin to develop a social semiotic framework to account for transcripts as artefacts, treating them as empirical material through which transcription as a social, meaning making practice (and changes therein) can be reconstructed. Social semiotics is particularly apt to inform such an account as it is concerned with changes in the multimodal landscape of representation and communication, whether in interaction mediated by the body and in physical co-presence, or graphically in film or books (Kress, 2010). A social semiotic perspective on transcription draws attention to meaning-making principles, and the potentials and constraints of modes of transcription with the purpose of gaining analytical insights and developing theoretical arguments.

**Transcription as semiotic work**

In this section we begin to theorize transcription from a social semiotic perspective. We start from the assumption that transcription is ‘semiotic work’ (Kress, 2010) and that that work is guided by certain interests and principles that are socially regularized and individually instantiated in response to the particular representational need – here, gaining analytical insights and persuading an audience. Our perspective foregrounds the agency of transcribers in that they make significant representational choices, whilst acknowledging that they are constrained by the social context. These are related to questions such as: How do I frame the transcript? What do I select for transcription? What do I highlight in the transcript? We assume that these representational choices shape the social relations between transcribers and readers, between transcribers and the participants represented in the transcript, between the represented participants and readers, and between the represented participants themselves.

**Framing**

Transcripts never operate in isolation; they are part of academic practice, such as the writing of a journal article, a paper presentation at a conference, a data session or a course on teaching methods. These contexts of use – particular social environments – ‘frame’ (Goffman, 1974) the transcript. In one frame, a transcript may constitute ‘evidence’ used by the presenter to persuade an audience of ‘peer researchers’; in another frame, the same transcript may constitute an ‘example’ of a teaching style presented by a teacher educator to student teachers. These social relations can be realized, for instance, by the placement of the transcript in the text. In, say, a journal article, the transcript may be the starting point of a discussion and therefore appears first, whilst the elaboration of the transcript appears
subsequently. In another text the transcript may be the end point, illustrating what was first described, and may even appear outside the main body of the piece in an appendix.

Video extracts that are selected for transcription are both framed by the communicational aims of the original interaction and by the purpose for which the graphic version is being made. For instance the observed activity may be framed as a ‘demonstration’ in a classroom, defining social relations in terms of a teacher who models and learners who need to know how to do what is being demonstrated. If this is transcribed, a new frame is added by the researcher in addition to the interaction that was constructed by the participants as it is examined with certain research objectives and from a particular theoretical perspective. This is a process of re-framing in terms of entextualisation (Silverstein, 1992). An activity is placed in a new social context where it is made to correspond to that new – in our case – academic framing. The transcript brings out the categories that are legitimate in this academic context; it views the ‘original’ observed activity through a professional lens which is, inevitably, different from the lens through which the participants in the ‘original’ activity constructed it. This is not always acknowledged by the makers of the transcript (Blommaert, 2005).

Selecting

Transcripts, like any representation, are partial – and this includes ‘multimodal’ transcripts. Processes of sampling from available audiovisual recordings, including principles of and criteria for such selection, are discussed in most introductions to video-based social research. Among common criteria, especially in classroom research, are ‘telling’, ‘critical’, or ‘key’ clips in which social norms – ways of saying and doing things which are normally taken for granted – become subject to (re)negotiation, for instance when new members of a community, unaware of certain practices, get corrected if they deviate from them, or when members who are well aware of those norms challenge them.

Of central methodological concern are the principles of selection and omission that are in play as the researcher transcribes the selected strip of interaction. In a more or less articulated way, researchers choose to include an extract that involves certain participants – e.g. they may choose to represent ‘ratified’ and ‘side participants’, whereas ‘eavesdroppers’ and ‘bystanders’ (Goffman, 1981) might be excluded – and features of the site of interaction such as artefacts or representations, whilst excluding others. What is selected is typically guided by analytical and rhetorical purposes, not by the principle that the more one includes in the transcript the more ‘accurate’ it becomes.

Once a portion of the video footage has been selected for particular attention, the researcher engages with recorded materials in an incremental process of refinement; the transcript becomes increasingly honed in responding to the issue for which the extract was selected.

Highlighting

Salience refers to what is highlighted in the transcript, or which of the re-made features are given prominence. Researchers draw the attention of their readership to elements of the focal interaction by highlighting them in the transcript, for instance through size and positioning. The selected strip of interaction is reconstructed, foregrounding certain features, which may even have appeared in the background of the original interaction. In writing, bold or enlarged typography and positioning to the left (in English) can give prominence. In image, a forward-facing orientation can be used to construct ‘demand’ in the relationship with the viewer (see Kress and van Leeuwen, 2006). That which is not considered central can be backgrounded,
and features not deemed relevant to the analysis can be excluded. In Transcript 1 (below), for instance, the focus is on what the students were expected to attend to: the visual (what could be seen on the screen) and the aural (the teacher’s voice-over). This omits other co-present features such as the teacher’s gaze, facial expression, position, clothing, and so on, as well as the 29 children in the class and the environment of the classroom – even though the reader knows they were there – because the current rhetorical focus is the framing of the task.

[ Insert Figure 1 about here as Transcript 1 (Excerpt from Mavers, 2009, p.146) ]

Principles of selecting, highlighting and framing are applied throughout the research process, and the transcript fixes a particular theoretical and rhetorical moment in that process. At other moments, for instance, during data collection, choices are made along similar lines. For instance the focus of the video camera is selective, because it highlights certain things and frames what goes on in particular ways.

**Transcription as ‘Transduction’: Shifting Modes**

In representing social interaction in transcripts, ‘translations’ are constantly made between (ensembles of) modes. By ‘mode’, we mean a socially and culturally shaped set of resources for making meaning, such as speech, gesture or image. Modes have different materialities and, shaped by the histories of cultural work, offer certain possibilities for re-presentation. Given that material and cultural difference, there can never be a perfect ‘translation’ from one mode to another: image does not have words, just as writing does not have depiction; relations which in speech or writing are expressed in clauses and verbs are realized through ‘vectors’ in image; forms of arrangement (‘syntax’) differ in modes which are temporally or spatially instantiated (Bezem er and Kress, 2008; Mavers, 2011). The change from one mode to another characteristic of the ‘transduction’ of transcription brings with it a change of entities and changes in the structural organization of those entities.

It is through re-making video as a multimodal transcript that researchers come to see differently. The modifications brought about by transduction are not only necessary, but it is precisely the re-making of observed activities in a transcript that can lead to fresh insights. It is for this reason that a social semiotic perspective on transcription takes issue with distinctions between ‘description’, ‘analysis’ and ‘interpretation’. Video data which are turned into multimodal transcripts are not merely descriptive, nor are they mere ‘translations’. They are transducted and edited representations through which analytical insights can be gained and certain details are lost. Thus the transcript is a crucial analytical tool, facilitating and articulating a particular ‘professional vision’ (Goodwin, 1994), rendering visible the socially and culturally shaped categories through which the researcher sees and reconstructs the world. From this perspective, the ‘accuracy’ of a transcript is dependent not on the degree to which it is a ‘replica’ of reality, but how it facilitates a particular professional vision. For instance, if one aims to facilitate the ‘vision’ of a conversation analyst the transcript may be seen as ‘inaccurate’ if it does not represent that which the participants can be seen to ‘orient to’ on the video record, such as their bodies and the tools they use in the activity that they engage in as the question of what participants orient to is key in conversation analysis. If one aims to facilitate the view of the variational sociolinguist, the transcript may be seen as ‘inaccurate’ if it does not represent phonologically distinctive features of what participants say. Thus our view is not that all transcripts are equally ‘valid’, but that they can only be assessed within their context of production. Like any artefact, when moving transcripts across different disciplines/discourses its representational claims may no longer be recognized (cf.
This withholding of recognition is one way in which scholars construct boundaries between disciplines and contest one another’s production of knowledge.

**Analysing transcripts**

In this section we reconstruct some of the transductions through which multimodal transcripts were made. We look at an admittedly small selection of transcripts which is by no means representative of the range of disciplines that the makers align themselves with. Rather we have chosen them to exemplify the types of transduction that transcribers engage in when transcribing. The transcripts appeared in monographs and journals. Had we looked at ‘working transcripts’ or ‘drafts’ then we would probably have found that the researchers made different choices at different points in the research process and in view of different audiences. Even so, the semiotic principles applied would have been the same. In the biographies and theoretical frameworks of these publications the authors have aligned themselves with Conversation Analysis (Christian Heath, Jonathan Hindmarsh and Paul Luff); Social Semiotics (Diane Mavers); Ethnography, Anthropology and Conversation Analysis (Frederick Erickson); Conversation Analysis, Interactional Sociolinguistics and Discourse Analysis (Sigrid Norris); and Linguistic Ethnography and Social Semiotics (Jeff Bezemer).

**Writing in transcripts**

Beyond lexis and grammar, certain resources of writing are used in transcripts to re-present the sounds and rhythms that can be heard on the video recording. Strings of phonemes are usually transcribed in a standard spelling. Non-standard accents are rarely detailed, or only selectively, and instead are often transducted into standard orthography, which serves ‘the best speakers of the standard dialect’ (Duranti, 1997, p.139). Detailed sets of conventions have been developed for making the most of the affordances of alphabetic script for showing the cadences and turn-taking of speech, capitalizing on such resources as typography (e.g. emphasis, underlining), punctuation and layout, especially in linguistics and conversation analysis. For example, features of speech, such as the tone at the end of a breath unit, may be represented as a question mark (e.g. line 60 in Transcript 2), or additional, non-standard typographic symbols may be used to mark the tone as ‘rising’.

[ Insert Figure 2 about here as Transcript 2 (Excerpt from Erickson, 2004, p.58) ]

Transduction in writing not only involves re-presenting the world, but also attaching a ‘reality status’ to those representations. Academic writers ‘hedge’ their statements, for example ‘It seems plausible that the interviewee misunderstood the job interviewer’s question.’ Researchers suggest a ‘reality status’ in transcription too. Representing hesitations and false starts suggests a form of realism that may be preferred in one context (e.g. in a journal for conversation analysts), whereas indirect speech implies greater abstraction that may be favoured in other contexts (e.g. an ethnography). This awareness is made explicit with regard to Transcript 2, which ‘should [...] be considered as illustrative sketches, as selective renderings with particular heuristic purposes’ (Erickson, 2004, p.29).

Transcript 3 shows a different use of writing in a transcript. Rising and falling intonation patterns are expressed in curves, vocal stress is expressed in the size and boldness of letters, participants are allocated distinctive assorted shades of font colour, and the English translation of the German talk is set in a separate box. The timing of speech is not delineated in such a detailed manner as Transcript 2, but it is clearly related to visual aspects of the interaction.

6
Writing can also be used to transcribe vocal articulations such as singing, laughter, sniffing and throat clearing (Lancaster, 2001), voice quality or volume (e.g. line 63 in Transcript 2). Even so, lack of specificity can render meanings ambiguous. For example, the aurality of laughter transcribed as ‘ha ha’ can take many different forms: giggling, nasal emission of air, suppressed chortling and guffawing (Mavers, 2011, p.116). Certain features of orality are difficult to represent in transcription.

The lexis of writing provides a range of choices for transducting – ‘describing’, ‘annotating’ – modes of communication other than speech: in Transcript 2, gaze (‘looks at...’), head movements (‘shakes head’), upper body movements (‘turns to other child’) and facial expressions (‘slight frown’). Which terminology is selected to describe movements is significant for how the original interaction is re-presented. Punctuation can also be used to transcribe gaze. For instance, following Goodwin (1981), Bezemer (2008) underlines parts of the transcribed utterances of a teacher to indicate when a student looks up from reading a handout to the teacher. A similar system is used by Heath, Hindmarsh & Luff (see Transcript 6), where ‘a continuous line indicates that the participant is looking at the co-participant, a series of dots (‘.....’) that one party is turning towards another, and a series of commas (‘,,,,,’) indicates that one party is turning away from the other’ (p.71).

When writing is used to represent a range of modes, these are not always transcribed in parallel and over the full length of the activity. In Transcript 2, speech appears to have been transcribed throughout, whereas meaning made in other modes appears fragmentarily. For example, the head movement of student B is picked out (line 53) and elsewhere facial expression (lines 56 and 58).

**Image in transcripts**

Different types of images are used in transcripts, including video stills (e.g. Transcripts 2 and 6), drawings and computer-generated images (e.g. Transcripts 4-5 and 7). Although speech and its characteristics can be shown in image (e.g. mouth shape, writing-like squiggles), it is usually transcribed as writing. Image is largely used by researchers to depict, at a glance, the visual characteristics of people, objects and places, and relationships between them, as well as sequences of action. Unlike some written transcripts where modes such as gesture and speech are separated out (e.g. Transcript 1), with image it is often left to the reader to discern different modes, or to refer to a written commentary.

Photographic stills such as those of Transcript 3 can appear inclusive. Unless deliberately edited out, images captured with digital devices include whatever is within their focal range. Irrespective of participation in the interaction, whoever is physically co-present and within shot of the camcorder is included. These stills have the potential to bring out visual characteristics and appearance such as skin colour, hairstyle, clothing, facial expression, posture, gesture and spatial proximity, supplying a certain specificity that must be described, or omitted, in writing. The transcript ‘reader’ can see those communicational features which fall within the frame of the camera at that particular moment in time with relative exactitude, and can – and will – make inferences, for example with regard to identity, social relations, activity, mood, and so on. This also raises ethical issues (Flewitt, 2006), particularly in relation to the anonymity of the research participants. Identifying features which enable
recognition are not usually transcribed in writing, whereas, depending on the camera angle, they are ‘just there’ in video stills.

Just like writing, image-in-transcription involves processes of selection. It may be actually only suggestive of visual inclusivity both because recording is itself a selection and as participants and items are omitted, removed, blurred or blocked out either in handmade representation or using readily-available photo-editing software. By transforming video stills into line drawings, attention was drawn to a student who remained – and wanted to remain – largely on the periphery of a joint reading exercise in an English classroom (Transcript 4). As the manner of her embodied engagement became the focus of analysis, the social actor dominating the spoken interaction – the teacher – was excluded, whereas she did feature in a written transcript of the same activity, presented elsewhere in the article (Bezemer, 2008).

[ Insert Figure 4-5 about here as Transcript 4 (Excerpt from Bezemer, 2008, p.173) and Transcript 5 (Excerpt from Bezemer, 2008, p.173)]

Image, like writing, can be used to suggest how ‘real’ the representation of a statement is: whether it should be read as an abstraction, or as a concrete instance, as imagined or real objects and processes. This ‘modality’ – an indication of the ‘reality status’ – of an image (Hodge & Kress, 1988; Kress & van Leeuwen, 2006) is suggested using a range of different resources, including: 1) Spatial detail: image may accurately represent the spatial proportions of people and objects or adjust their size and positioning; 2) Pictorial detail: items may be given more or less detail; 3) Depth: a continuum can be suggested by variation in size, overlap or shading; 4) Colour: represented items may or may not be given (their—conventionally—typical) colour, and may be varied in terms of saturation, differentiation and modulation (see Kress & van Leeuwen, 2002, p. 5) Background: may or may not be included in a more or less recognizable way. Transcripts 4 and 5 show how these resources have been used to create different forms of realism in transcripts. The computer generated line drawing (Transcript 5) provides more details of space, depth and background than the hand-made line drawing (Transcript 4), thereby suggesting a more ‘concrete’ and ‘complete’ form of realism.

Layout in transcripts
Transcription is not just a case of choosing image or writing, but also making decisions about how these will be set out on the page or screen. Whether a transcript consists entirely of writing or entirely of images, or a multimodal mixture of the two, researchers use spatial organization to construct separation and cohesion, to disconnect certain parts of the writing and images and to show which parts belong together, as well as to suggest an order of attendance. By separating out modes to show how they occurred in temporal succession, and by presenting them in such a way as to reconstruct their combination, the complexities of meaning making can be demonstrated (e.g. Figure 7). Layout also plays a particularly important function in representing how the observed activity unfolded in time. Single video stills are ‘snapshots’ frozen at a particular instant in time; an exact moment is captured. For example posture, expression and gaze are re-presented as they are at that split second. Video footage may also be edited or remade in a drawing in order to conflate different moments in time into one image, so that participants are not represented according to the simultaneity of the original interaction. Image can show states highly specifically (e.g. a frown, or a key point in a gesture or a movement), but moment-by-moment shifts in unfolding motion as change over time must be inferred from a sequence of stills.
Laid out as a ‘plate’ that is intended to be read from left to right and top to bottom, Transcript 3 presents a sequential series of video stills. Rather than selecting frames at an even time interval, from a period of about 30 seconds the researcher has chosen to highlight the moments in an unfolding activity to which she wishes to draw attention. Norris wants to discern ‘simultaneously performed higher level actions’ (2004: 101), and so frames the moments where new higher level actions are added to what is going on, for instance when one of the participants starts a phone conversation whilst also sustaining the conversations in which she was already engaged. Compared to Transcript 2, which is made in writing only, this transcript represents how modes operated simultaneously, but only at selected moments – split seconds – in time. Superimposition of writing onto the images suggests how speech coincided with other modes such as gesture. Even so, as video footage consists of a number of frames per second, this correspondence is an approximation. Within this ambiguity, the synchrony –or a-synchrony– of different modes in the transitoriness of interaction is subject to the interpretation of the transcript ‘reader’.

Transcript 1 shows another common layout for transcription: the tabular format. This layout too has particular ontological implications. It constructs temporality on a vertical axis and modal separation horizontally. This provides an impression of how the meanings made unfolded synchronously and diachronically, and how they map onto each other. A tabular layout restructures communicational organization by placing selected modes in columns. In Transcript 1, modes are split into gesture and speech respectively. Other transcripts may be broken down in terms of other modes. This arrangement may suggest certain pathways of ‘reading’. Positioning action or gesture in the left-hand column may imply that these modes should be attended to first, and that speech (located to the right) should be understood against the backdrop of these movements. Even so, ‘readers’ retain agency, and may choose to ignore the implied directionality and design a course of their own.

A third type of layout is shown in Transcript 6. In contrast with Transcript 1, temporality is arranged horizontally, whilst the participants and modes are separated out on a vertical axis. Placed at the centre of the transcript, the timeline constitutes a ‘base line’. Speech is mapped onto this in rows immediately below, with the first speaker assigned the upper line and the second speaker that below. Other modes are represented in rows elsewhere; meanings made by the first speaker in modes other than speech appear at the top, and meanings made by the second speaker in modes other than speech appear at the bottom. Thus the transcript places speech at the centre, and attaches other modes to it. This indicates precisely where actions sit on the timeline: a single dash represents pauses and silences of up to one tenth of a second (for a very short time span). On the other hand, the three stills appearing at the top are mapped onto the timeline approximately.

[ Insert Figure 6 about here as Transcript 6 (Heath, Hindmarsh & Luff, 2010, p.70) ]

**Concluding remarks**

In this paper we have analysed and discussed a small number of ‘multimodal’ transcripts from a social semiotic perspective with the aim of gaining a better understanding of the methodological implications of this changing academic genre. We have suggested directions along which to investigate transcripts: how common principles (framing, selecting and highlighting) and modes of transcription (writing, image and layout) are differently used in video-based social research. Whilst these principles operate in all modes, each mode provides distinctive potential for re-constructing video data, and these choices shape the account of social interaction in significant ways. Such reconstructions are inevitable and essential.
outcomes of any video analysis, and it is through reconfiguring video data that researchers and their audiences can see the observed interaction in the categories appropriate to their discipline(s) and position themselves in relation to that discipline(s).

We have identified some significant methodological differences between the various transcripts. We can see, for instance, that their makers have chosen to represent strips of interaction ranging from a few seconds (Heath et al.) to a couple of minutes (Erickson). Their transcripts also point to different units of analysis, such as: ‘turns’ (Erickson), ‘actions’ (Heath et al.), ‘higher level actions’ (Norris), and ‘modes’ (Mavers). These differences reflect the professional interests of the makers, and their analytical and rhetorical concerns. As conversation analysts Heath et al. are particularly concerned with the temporal unfolding of action, and they tend to look for more detail in shorter clips than for less detail in longer clips. They want to show that interaction is sequentially organized, and that this interaction unfolds in different forms of action. As social semiotician, Mavers calls these forms of action ‘modes’, and highlights the modal organization of interaction. Norris wants to discern ‘simultaneously performed higher level actions’ (2004: 101) in a 30-second clip, and so frames the moments where new ‘higher level actions’ are added to what is going on. Bezemer identifies ‘types’ of bodily configurations in a 5-minute clip, thus like Norris he moves beyond the ‘micro’ actions which are in focus in Heath et al.’s transcript. Another difference between the transcripts is their treatment as a ‘source of evidence’. In conversation analysis evidence for an argument ought to reside in the transcript, and other, ethnographic sources of evidence may be considered as ‘supplementary’. In the ‘micro-ethnographic’ approach adopted by Erickson such sources of evidence are given more equal weight. Thus the transcripts have different positions in the originating analysis and rhetoric.

In this paper we have looked at transcripts as finished products, as professional artefacts. We treated these artefacts as mediating social interaction between the ‘makers’, the represented materials, and the (imagined) ‘readers’. The framework set out here is not only useful from the perspective of the sociology of science, but can also be applied by individual researchers undertaking video-based social research to reflect on the methodological and theoretical implications of choices around transcription. Such reflection, we believe, should not focus on representational ‘accuracy’. Rather, transcripts should be judged in terms of the ‘gains and losses’ involved in re-making video data. It is crucial to make those gains and losses transparent, for example, which modes of communication used in the observed activity have been excluded from the transcript and why, and what the effect is of that exclusion on the analysis and subsequent reader interpretation? It also promotes reflection on the effects of transduction: how use of the mode of transcription shapes what is re-presented. Transcription conventions accommodate such transparency and consistency, but are currently utilized only in transcribing speech to writing. Contemporary practices in multimodal transcription may require information, for example, on how images were constructed. Such conventions cannot and need not be standardized beyond the study/project/publication for which they are used, but they need to be made transparent to readers.

We have discussed only a small number of transcripts in this paper, and so our conclusions are provisional. We are in the process of expanding our data set, so that we will be able to systematically analyse a broader range of multimodal transcripts. This corpus will increase the variety of transcripts we have worked with so far, to include, for instance, musical notation as a mode of transcription showing the relative temporal locations of vocally stressed syllables in talk (e.g. Erickson, 2004), ‘laban’ script as a mode of transcription for movement (Duranti, 1997, p.149) and geographical maps detailing the direction of gaze (e.g.
Haviland, 2003). Further research will also need to involve interviewing the makers of transcripts, as well as observations of transcription activities (c.f. Vigouroux, 2007) to gain more insight into transcription as a situated practices (Mondada, 2007) and ‘effects’ on readers. Transcripts, like any form of representation, are not only socially and culturally shaped, they are also situated (Mondada, 2007), that is, they are produced in a local, social, physical context in which certain representational resources are available and others not. Increasingly, computer software technologies are part of those resources, and they shape ‘transcription’ (Plowman & Stephen, 2008). Transcription may also be a collaborative activity, involving a number of participants who are engaged in a ‘data session’. It would be particularly interesting to compare different versions of a transcript, made in different situations, involving different participants, and the processes of review and reformatting. The interactions between transcripters, and between transcripter and computer, and the local ecologies within which these interactions unfold need to be researched ethnographically. So does the ‘reading’ of transcripts: to our knowledge, no one has observed how people engage with transcripts. We do not know what they attend to, in what order, or indeed if they read the transcript at all. In short, our paper is only a very modest theoretical and empirical contribution to important methodological questions.

Acknowledgements
We are grateful to our colleagues at the Centre for Multimodal Research, Institute of Education, University of London, with whom we have had many fruitful discussions about some of the issues raised in this paper. We would particularly like to thank Gunther Kress, Charalampia Sidiropoulou and Carey Jewitt for their helpful feedback on earlier versions of this paper and their encouragement to proceed with this line of inquiry. We would also like to thank the anonymous reviewers, who raised many important issues and provided very useful suggestions for developing our understanding of transcription; and the authors of the transcripts discussed here, who kindly gave us permission to reprint (excerpts) of their transcripts in this paper.

References