Maria Grazia Sindoni

Through the looking glass: a social semiotic and linguistic perspective on the study of video chats

Abstract: This paper provides a theoretical framework for the study of the video chat, a spontaneous web-based synchronous text that allows forms of communication in which social semiotic resources come into play and produce a new terrain of investigation for researchers in the field of linguistics, multimodality, communication and media studies, visual ethnography, and digital literacy. In particular, the paper singles out some aspects for analysis, such as the alternation of speech and writing, new proxemic and kinetic patterns, gaze management, and impossibility of eye contact and discusses some examples from digital field work on multiparty video-based interactions. Speech and writing are technologically integrated, allowing participants to mode-switch, i.e., to alternate between spoken and written discourse. New arrangements of verbal and nonverbal resources attempt to simulate face-to-face conversations. However, the illusion of a face-to-face conversation dissolves as soon as video chat-specific resources are unpacked. Despite growing research into nonverbal behavior, video chat data challenge visual analysts and researchers for a number of reasons. A transcription model, developed for the purpose of analyzing these specific texts, will be sketched to give a brief account of significant data that need to be incorporated into multimodal transcription and annotation. Reflections and conclusions are drawn according to the contribution that intersemiotic studies can potentially provide for web text analysis, given the constant expansion of web-based texts and the challenge it brings as regards new notions of textuality.

Keywords: video chat, mode switching, digital speech, digital writing, multimodal transcription, synchronous video-based interaction

Maria Grazia Sindoni: Dipartimento di Civiltà Antiche e Moderne, University of Messina, Italy.
E-mail: mgsindoni@unime.it
1 The challenges of analyzing video-based interactions

The digital domain is the ideal terrain of investigation for multimodal analysts, as the repertoire of socio-semiotic modes and resources is orchestrated in texts and genres in new or emergent configurations (Kress 2010; Herring 2013; Tannen and Trester 2013). The challenge that new and emergent digital texts bring about consists in the reshaping of social semiotic resources and in the new configurations of meanings that can be mapped across web texts and genres.

Innovation is instantiated in how the repertoire of semiotic resources, including language, is orchestrated, and continuity is ingrained in the history of oral traditions and multimodal performances. This paper investigates how the traditional categories of speech and writing are altered in the multimodal web domain, also singling out other significant categories in multimodal interaction. Much of the literature on multimodality has it that speech and writing are intertwined in digital texts (Kress 2003, 2010; Kress and Van Leeuwen 2001, 2006; Gee and Hayes 2011; Thurlow and Mroczec 2011). Such categories seem to have lost part of their appeal, as their theoretical conceptualization appears to be ingrained in frayed Western logocentric traditions, but they are nonetheless useful to examine and make sense of new and emergent digital texts. If language is seen through the looking glass of multimodality, other significant resources appear in their fully communicative affordances.

This paper thus sets out to explore a digital text, the video chat, that can be defined for several reasons as a new form of textuality, as it has no observable counterpart in the non-digital world. The video chat imitates face-to-face conversations to some extent, but it also overrides the typical features of in praesentia interactions. It is a slippery text that makes full use of a wealth of resources: spoken, written and visual language, body movement and gaze, and staging of the surrounding environment through webcam positioning, to name but the most prominent. It is, as such, a text where an unusual number of different resources collide and converge. In the next sections, these resources will be described and discussed with regard to how the video chat challenges previous notions of online video-based textuality.

2 Speech and writing in video chats

A video chat is a system of computer-mediated communication (CMC henceforth, cf. Herring 1996) made possible by software programs that allow users to communicate in real time from disparate places, also seeing each other online, with a
minimum, often imperceptible, time-lag. Different platforms, e.g., Skype or MSN, are typically used for either private or public events, whereas client systems like Camfrog allow web-based multiparty interaction, where users meet and talk.

Video chats simulate face-to-face conversations, giving the illusory perception of sharing the same context of situation, a visual and psychological perception created virtually through a range of different strategies (Sallnäs 2005; Weisz et al. 2007). A fundamental factor adding to the illusory perception of a face-to-face conversation is that verbal exchanges occur in real time. Moreover, users see each other in real time, so that conversations reproduce long or short verbal exchanges in praesentia, with turn taking, pauses, hesitations, etc. However, this system of interaction is a rather crude imitation of what happens in face-to-face conversations.

With regard to the channel of communication, participants have the chance to use both live conversation and written comments during a web-based video event. I have characterized this specific option as mode switching, that is, the alternation of speech and writing in the same communicative event and in synchronous mode (Sindoni 2012). Medium affordances allow the use of both modes in a video chat, but the notion of mode switching is rather restricted and needs to be placed in specific contexts of use. Mode switching paraphrases the notion of code switching (Heller 1988), but also draws on Halliday’s notions of written and oral channels of communication (Halliday 1978).

Mode switching raises a number of research questions, for example, the reasons why users decide to switch from speech to writing and vice versa, the range and extent of mode switches, and the reasons why users may be high or low mode-switchers. Evidence shows that such switching may be occasioned by both internal and external factors (Sindoni 2013). Camfrog, or other client systems that allow multiparty web-based video chat, provide a virtual space, i.e., a room, where each user is allotted a slot where one’s own video can be projected. Akin to a real room, all participants find their own place in the virtual room, where users can also write and post written comments. In face-to-face contexts, many contextualization cues serve the purpose of regulating conversation, for example, effectively managing turn taking (Sacks et al. 1974; Sacks 1992). In virtual environments, contextualization cues (Gumperz 1982) are absent and managing turn taking is much more complex. When it comes to making sure that spoken turns do not overlap, medium-constrained systems usually guarantee the floor only to one participant at a time. For example, clicking a microphone icon allows a participant to hold the floor, thus de-activating the icon and hence the possibility to speak for all other users.

However, when multiple users are involved in the conversation and the floor is held only by one speaker, the written chat is widely used to hold parallel floors.
In other words, when the speaking floor is held by one participant, the others have two options: (i) to listen to the one speaking, or (ii) to hold parallel floors by using writing. Since overlapping among turns is particularly frequent in these environments, it is very unlikely that all users listen to one speaker. The preferential use detected so far is the second option, thus holding parallel floors via writing.

In the platform selected for study, i.e., Camfrog (versions 6.2, 6.3, and 6.4), several hundred hours of recording (ca. 800 to date) have been made to explore mode switching and other relevant phenomena. Preliminary concerns involve ethical issues, as video recordings are sensitive data, even though technically any user may be recording a video session (saving conversations is an integrated function in Camfrog). However, users were informed that I was collecting data for academic purposes and administrators posted relevant information about my research during video-recording sessions.

To understand the amount, rate and occurrences of mode switching, Camfrog affordances must be preliminarily taken into account. For example, to avoid spoken turns overlapping, only one participant at a time is allowed to speak by clicking the “Talk” icon. Conversely, participants can freely use the written channel, thus reducing possible troubles due to turn taking management. The reason for this state of affairs is that the overlap of spoken turns creates conversational chaos, also because users do not share the same context of interaction, whereas the overlap of written turns is solved by their internal linearization, i.e., they are published in chronological order and appear on screen one after the other. A screenshot of the Camfrog environment is provided in Figure 1.

![Fig. 1: Camfrog multiparty conversation screenshot](image-url)
With regard to mode-switching description, some concepts from conversation analysis have been adapted to analyze video chats. Mode switching can be described as being either self-initiated or other-initiated. The former refers to those switches performed on one’s own initiative; the latter refers to those performed following other participants’ prompts, e.g., when participant A addresses participant B in a channel other than that used by participant B. Following theoretical models from conversation analysis, mode switching can also perform the function of repairing trouble in conversation.

During a multiparty conversation involving five users, the more active user, nicknamed Essence, mode switches on a number of different occasions. See the following example, where mode switch is indicated under square brackets in turns 4 and 5 and spoken turns are in italics:

(1) 1 Essence: *do you see it? How it’s different/ like you were on the (...) version is (...) /is on my mic is like going up// like in a circle/ or whatever/ that’s what I was asking if it was too loud or not//

2 Diamond: ah ok

3 Diamond: who here is married?

4 Essence: Tinker ☺ [mode switch]

5 Essence: *no/it’s not too loud/because it’s not on the red// [mode switch]

Essence, who was holding the spoken floor in turn 1, mode-switches in turn 4, answering Diamond’s question. He uses writing as he is directly addressed via writing. This is an example of other-initiated mode switching, as the participant follows another participant’s prompt and changes the mode or channel of interaction. The second case of mode switching that is illustrated above is performed again by Essence, who mode-switches in turn 5, answering Sensei’s question by using the spoken mode. In both cases, Essence decides to adapt to the other participant’s mode, answering with spoken or written turns according to the other interlocutor’s use. In this last instance, mode switching is also used to hold multiple floors of conversation, as Essence carries forward two parallel conversations, one about the software program’s version in spoken mode, and the other about Tinker’s marital status in written mode. An overall analysis of Essence’s preferential use of the channel of communication in the whole conversation (ca. 01:23:00) indicates that he prefers the spoken mode and mode-switches from speech to writing only when specific circumstances occur.

A very common topic of discussion in Camfrog concerns the room’s rules. Talking about one’s own illegal trading activities is apparently much more tolerated than breaking the room’s rules, for example, using the wrong font or color, or writing in capital letters. When the debate becomes hot, a higher incidence of
mode switches is very likely to occur, as only one participant can hold the floor at a time. In the following example, the conversation is about the font size that a participant, Tod3344, is using. He is reprimanded by the room’s moderator, Diamond. In the following exchange, another mode switch occurs:

(2) 1 Tod3344: ARE MY FONTS RIGHT NOW?
   2 Essence: yes tod/ tod you can make it small/ you know/ make it bigger/ like on the 16/ 14 something like that/ used for it like this/ y’know//
   3 Diamond: Tod3344, Please Respect all users.
   4 Essence: they<small . . . but u changed on red [mode switch]

This is an example of self-initiated mode switching performed by Essence. He is not directly prompted by anyone, but decides to mode-switch on his own initiative. Diamond, the moderator, chooses the written mode to admonish Tod3344 and tries to use rules of “good writing” with care, starting his comment with the addressee’s name to avoid ambiguity, capitalizing some items, and using punctuation that is very infrequent in these environments (turn 3).

Repair can be also performed by mode switching in video-based interactions. In the following example, Tinker mode-switches for self-correction:

(3) 1 Tinker: Sensei your video’s fine
   2 Tinker: no I mean/ Sensei your audio’s fine// [mode switch]

This is an example of self-initiated mode switching for repair, or self-correction, and is instantiated in the spoken mode as Tinker wants to make sure Sensei receives her message. Other examples of repair can be found when participants use mode switching to bypass trouble in conversation or correct oneself or others. Another use of mode switching is holding multiple floors for conversation, as mentioned above. In the following example, Akadam and Tuga27 are interacting:

(4) 1 Akadam: yes great
   2 Tuga27: yes u can call
   3 Akadam: yes
   4 Leim: where is the button Tuga?
   5 Tuga27: down below// [mode switch]
   6 Tuga27: aka call right now [mode switch]

This example is similar to example (1), where Essence was occupying multiple floors, one in the spoken mode and the other in the written mode. Here Tuga27 is replying to Leim using the same mode, thus performing an other-initiated
mode switching, but he is, at the same time, holding two different floors, as he is engaged in two conversational threads with two different interlocutors, i.e., Akadam and Leim. In written turn 6, Tuga27 is getting back to Akadam, so he starts his entry writing his interlocutor’s nickname, a very common strategy to avoid ambiguity in digital conversations. When participants want to make sure that their addressee has their attention, they start their written entry with the addressee’s nickname, as also shown in example (2).

In other instances of self-initiated mode switching, this option is used to claim one’s own authority within the platform, as in the case of the room’s administrators, who are allowed to reproach or ban users from the community. In the example below, the room’s moderator mode-switches, from written to spoken mode, to reinforce her message:

(5) 1 Lemon_drop: You can’t say that in here.
    2 SuperShineSingleBot: $36 + 8 - 8 - 42 \times 49 = ?$
    3 Lemon_drop: MAATT!
    4 Lemon_drop: You can’t offend anyone in this room.
    5 SuperShineSingleBot: 25 seconds remaining . . .
    6 kanee21: ooii
    7 SuperShineSingleBot: 15 seconds remaining . . .
    8 Lemon_drop: Matt// read the room rules/ before you enter// [mode switch]
    9 SuperShineSingleBot: 15 seconds remaining . . . last letter is U

This exchange features some common traits in multiparty video-based interactions, in particular the thick superimposition of conversational threads. The only spoken turn is number 8 and is performed by Lemon_drop, who is addressing a participant, Matt, who is not answering Lemon_drop’s queries, either in spoken or written mode, but who was showing a confrontational attitude several turns before the reported example. In the meantime, the other participant, SuperShineSingleBot, is playing a game (also shown in Figure 1), thus having a parallel conversation and not interacting with either Matt or Lemon_drop. The latter is trying to get her message across, but receives no reply, hence performs a self-initiated mode switching to add extra force to her message and emphasize her role in the room. The fact that only one user can hold the spoken floor implies that all connected users can hear her with no distracting factors, contrarily to what happens with the written floor that is perpetually refreshed with new comments. Incidentally, we may add that Lemon_drop, in her role as a moderator, is correctly using punctuation (turns 1 and 4) to increase her authority with the
sanctioning of “good writing.” This is consistent with other findings, as has also been reported in example (2) above.

Mode switching is neither casual nor erratic in these texts. This does not mean that it has been possible to determine specific and recurrent patterns of mode switching so far, but this notion has proved useful in explorations of the spoken/written continuum in digital habitats. Mode switching is different in several ways from other possible uses of speech and writing in synchronous mode as it is an unprecedented use of language modes during the same communicative event.

Despite the fact that no overall generalization is possible, functional groupings of mode switches and typical behavior of the highest mode-switchers have been attempted after digital field work and interviews with Camfrog and other client system users, who were invited to talk about their use and preference of both channels during video conversations. Seventy-eight percent of the informants agreed on the more “authoritative” value attached to the spoken mode (Sindoni 2011). However, this language use is specifically linked to multiparty video interactions, as one-to-one conversations do not display similar features in language use. A Skype conversation between two friends does not need such strong turn claims, as more common conversation facilitators can be used in one-to-one conversations to the same purpose.

From the analysis of ca. 200 spontaneous Camfrog conversations, it has emerged that the most active participants are higher mode-switchers than less involved users, who seem more at ease when adhering to one channel of communication, either spoken or written. However, lower mode-switchers use preferentially the written mode. Overall, the favorite option in these environments is the use of the written floor, mainly because it is easier to hold.

Screencasting and issues of multimodal transcription and annotation will be briefly discussed in Section 6. From these observations, a transcription model has been developed, allowing a systematic study of mode switching within video conversation events in synchronous mode. The rationale of this study is grounded on the linearization of spoken and written conversational threads, following a time-based model, which permits the linearization of superimposed turns.

Much in tune with what has been argued regarding turn management in web-based online interactions, mode switching has been found to be one of the most significant facilitators for turn taking management, as a way of repairing trouble and as a strategy to maintain the flow of conversation across both spoken and written conversational threads.

From a multimodal perspective, video chats include two differently integrated frames showing live images of users. Contrary to what happens in face-to-face interactions, users have the chance to see themselves while listening/writing.
to their discourse partners. Observing oneself during a conversation produces psychological effects influencing the verbal and nonverbal features of the online exchange. In other words, casting a sidelong glance at oneself during a conversation may change, if not determine, the way one speaks, gesticulates, smiles, and so on. It entails not only adjusting to the addressee’s expectations, but to one’s own expectations as well. In the next two sections, social distance, and the management of space and body movements will be analyzed to further shed light on video chat-specific and generic structures.

3 Proxemics

Perception, use, structuring, and management of space in interaction all have the effect of framing social expectations, interactions, and ways in which actual communicative events take place. All social interactions, and face-to-face communication in particular, occur in a setting or context of situation (Malinowski 1923), which is perceived by individuals who negotiate their relationship with space and with one another in meaningful and culture-bound patterns. Hall (1966) defined this specific use of social distance as “proxemics.”

Social distance in video chats is fixed, and participants stand in a frozen space. This means that medium constraints (close to very close shot of the projected image of those involved) do not allow reciprocal negotiation of space. Distance is not established by those who interact, but between one person and one machine. This distance foregrounds the representation of distance among users. Furthermore, distance is construed in web-based video interactions via webcam positioning that is different from neutral webcam position, in that the former implies a conscious act of representation, whereas the latter may imply medium-constrained affordances or technicalities (for example, embedded webcam in laptops).

Webcam positioning is a different matter. In other words, users are encapsulated in a typically fixed space, but they have the possibility to alter this distance by changing webcam position. For example, they can be framed in a very close or close shot, but they can also move the webcam and modify the width and distance that will be captured by the webcam, and reproduced accordingly. Multimodal analysis also needs to take into account the strategies used to frame the space and represent position within the projected frame.

Screenshots in Figure 2 provide some examples of different forms of representation of distance in web-based video interactions and how webcam positioning can be used and to what effect.
Participants have chosen to represent different portions of themselves and of the environments where they are placed in Figure 2. They can be in very close shot, as in screenshot 1, where Moldovar’s face is captured in such a close shot that it is visible only to a very limited extent. Furthermore, such positioning was maintained by Moldovar for the whole recorded conversation, thus creating a sense of undue intimacy and “awkwardness of conversation,” as one informant reported about Moldovar’s typical uses of very close shots on Camfrog. Analysts need to take into account these aspects, first of all considering whether webcam positioning and framing are instantiated by users on a random basis or whether they are aware of their semiotic choices and maintain such positioning on a number of different sessions. These observations require large amount of data, as occasional framing or unintentional positions do not represent semiotic choices *stricto sensu*. Furthermore, direct contact with users (e.g., interviews) can help clarify specific uses.

In other instances, participants can be in a long shot, as in screenshot 6, where MidEvilll has positioned the webcam high above, thus framing his room from a high oblique angle, akin to what happens in screenshot 3, where dina_10 is taken obliquely. Perspective and the role played by oblique angles have been amply discussed, also within multimodal frameworks of analysis (Kress and Van Leeuwen 2006), and the impact they produce in digital video-based multiparty interactions is not lost on users, who are able to perceive the markedness of the
oblique perspective. Broadly speaking, oblique positioning conveys a sense of detachment: in face-to-face conversations – as the adjective aptly describes, we would hardly imagine having our interlocutors in a position other than face to face when they are directly talking to us.

In screenshots 2, 4, and 5 only a portion of the participants’ bodies is visible. In screenshot 2, Golldboy has positioned the webcam below his face that is, as such, invisible to viewers. Framing in screenshots 4 and 5 also makes partially invisible faces and eyes, which are fundamental facilitators in conversation, as will be discussed in more detail in Section 5. The body is thus partly dematerialized and this lack of semiotic cues may also determine estrangement and reduction of involvement. Some informants, who are used to selectively show and hide portions of their bodies/ faces, have revealed an ambivalent attitude toward digital video-based conversations during interviews on the matter.

All the selected examples, with the exception of screenshot 4, show a marked webcam positioning that is maintained constantly by these users. Space representation and framing bring to light different degrees of involvement and detachment that can be conscious but unintentional at the same time, hence generating different reactions on other users. Furthermore, studies on proxemics include postural identification (i.e., sitting, standing), distance, frontal orientation, and body positioning (Harrigan 2005), which may be difficult to detect, especially when they are partially visible or fall outside the projected frame.

Ways of managing space during multimodal interaction affect the ways in which communication is organized overall and carried forward. In video chats, personal and other’s space is delimited by the frame. The distance between users is determined by the medium affordances, therefore the interpretation of social distance is difficult to grasp. No matter what social variables are at stake, video-based interaction occurs within a fixed frame, eroding culture-bound proxemic patterns. In other words, regardless of social and cultural positions, video chat communication systems erase cultural differences and seem to be favoring a flat representation of social and cultural identities.

4 Kinesics

Just as proxemic patterns are linked to culture and individuals and any visual analysis needs to consider such variables, so the same may be said about kinesics. Very few body movements have invariant meaning within or across cultures, as body movements cannot be translated as directly as verbal behavior can (Kendon 1990, 2004). Apart from messages conveyed by movements, the additional question of intentionality needs to be taken into account, as discussed in Section 3.
In verbal communication, there is a *deliberate* attempt to convey a message to a recipient. With nonverbal behavior, the question of intentionality is less clear, because some actions may well be defined as intentional (e.g., deictic gestures, such as pointing to objects and people in the physical world; Norris 2004: 28). Others are halfway between intentional and unintentional (e.g., iconic or metaphorical gestures, possessing a pictorial content and mimicking what is conveyed or hinted at verbally; 2004: 28). Others are unintentional (e.g., postural behavior; 2004: 24–27).

In the screenshots presented in Figure 3, some examples illustrate actions that participants can perform during digital video-based interactions. These actions can be analyzed along a cline of intentionality, from more to less intentional.

Screenshots 7 and 8 show two participants involved in actions that have nothing to do with the online interaction they are engaged in. Their daily routine is captured by the webcam, but they do not interrupt the interaction by a momentary pause or by logging out, as is fully illustrated in Figure 4.

Camfrog conversations can last for days, weeks, even months. Daily chores are not part of the digital interactions, but are nonetheless captured by the webcam. A consistent number of users drink, smoke, eat, and even sleep without logging out from Camfrog.

Kinesics has traditionally focused on hand and head movements (Kendon 2004; Martinec 2004), which have been studied as “action” behavior and which

---

*Fig. 3: Screenshots 7–12 from Camfrog conversations*
can be studied as discrete units of analysis, including “onset” and “offset” points (i.e., beginning and ending of actions, also shown in Figure 4). Some studies focus on verbal and nonverbal congruence, analyzing speech units, pauses, and timing of body movements (Boomer and Dittman 1964), whereas social psychology studies inner states shown by nonverbal behavior (Ekman and Friesen 1969). However, multimodal interaction does not deal with what people are thinking, but with what people are communicating in interactions (Norris 2004).

In the context of video chats, body movements cannot be completely ascertained by analysts, since hand movements, for example, may be semi-visible and, as mentioned above, partly fall outside the video frame. Other body movements, such as postural shifts, may even totally fall outside the video frame and, as such, be missed both by the other discourse partners and subsequent observers (i.e., multimodal analysts). Video frames do not usually include the entire body. They typically frame the face, and optionally a part of the torso, creating a frozen yet living image of users. The frozen yet living definition attempts to capture the contradiction in the video frame, which includes a frozen, fixed, and partial representation of the body and at the same time presents a living representation of participant’s behavior, e.g., speech, kinetic action, and postural shifts. The latter are typically instantiated by hand and head movements and torso postural shifts.

For example, screenshots 9 and 11 in Figure 3 display two intentional actions performed by Strawberry and 0_boxer_guy. They can be defined as unambigu-
ously intentional, as they are both showing something to their interlocutors, performing an action that is very common in both one-to-one and one-to-many digital environments. However, the action of showing something online is directly connected to a secondary set of changes in the semiotic environment, that is, the momentarily partial or total disappearance of the participants doing the action, that can be viewed and analyzed along a cline from more to less visible in screenshots 11, 9, and 10 & 12.

Another interesting phenomenon, which can be defined as *metamessage* (Tannen 2013), is featured in screenshot 10, where Damon’s webcam has been used to capture what his screen is displaying, i.e., the online Camfrog conversation. This operation can be analyzed from at least two complementary viewpoints: (i) it is part of webcam-positioning strategies, as discussed in Section 3, and (ii) it is an intentional and metareflexive action that points to the “looking glass” properties of such texts suggested in the introductory remarks.

These observations of digital field work do not purport to represent stable or permanent relationships between intentionality and behavior. Any specific interaction may present different configurations of intentionality, as any interaction is culturally, socially, and individually determined. However, the question of intentionality is essential in multimodal analysis of body movements, as it can orient researchers’ expectations, but also bias their understanding of body movements in interaction.

The discourse partner’s face is thus a metonym for the body realized by another face: in fact, one’s own. Body fragmentation and dematerialization are a typical feature of web-based online interactions. Few occasions feature the complete bodily wholeness of users. Just as telephone calls have the voice as the sole material (or semi-material) *presence*, standing for the whole person, video chats add a portion of “virtual materiality” to interaction. However, in the study of kinesics, multimodal analysts need to be aware of the typical incompleteness of such web-based live events.

## 5 Gaze and eye contact

The materiality of body is lost on the web. However, in web-based online interactions, eyes play a pivotal role. Gaze is one of the most effective resources in interpreting and making sense of a discourse partner’s attitude, stance, and behavior. Early research on visual behavior was concerned with a series of questions on different levels of communication. As a dependent variable, gaze has been used to measure stable individual and group differences, the regulation of the flow of conversation, and the search for feedback in interaction. As an independent
variable, it has been shown to influence emotional response and cognitive attributions (Ellsworth and Ludwig 1972). In the context of video chats, significant questions may be identified in how gaze direction influences the receiver and the meanings attributed to the other’s visual behavior.

Studies in psychology, interaction analysis, and the ethnography of communication have focused on gaze and its role in interactions and have provided insights into the different levels of analysis of gaze functions. The present analysis started following Argyle et al.’s paper (1973) that illustrated the different functions of gaze, identifying the following: (i) information seeking (i.e., to obtain immediate feedback); (ii) signaling either interpersonal attitude or prosodic accompaniment of speech; (iii) controlling the synchronization of speech; (iv) the relationship between mutual gaze and intimacy; and (v) the inhibition of gaze (i.e., avoiding undue intimacy).

For example, in an interactional sequence, a user, Cuddy, looks directly at the webcam when she says: “I’m sorry about that but I remember everything, I also remember some of the guys that were very rude to me”. Her two direct gazes correspond to the two verbal italicized items, namely remember and me at the end of the utterance, and precisely where the highest peaks of intimacy are conveyed.

Sacks, Schegloff, and Jefferson (1974) identified Transitional Relevance Places (TRP) as the fundamental points at which floor changes are negotiated. Goodwin (1980, 1981) and Kendon (1967, 1990) demonstrate that turn-taking behavior in conversation is mainly regulated through gaze. Kendon shows that TRP are anticipated via gaze, while Goodwin argues that having listener gaze is so important for the turn claim process that turn claimants will restart their utterances until the speaker’s gaze assures them they do have the floor. Telephone and computer-mediated communication presents problems in turn taking, which are similar to those described in blind and sighted interaction, since gaze is a crucial facilitator (Everts 2004). In Figure 5, it looks like AKAMat is looking directly into the interlocutor’s eyes: what he is actually doing, however, is looking at the webcam.

**Fig. 5: Gaze sequence**
The conversation was about the age of Lemon_drop, who had claimed to be born in 1992. AKAMat is counting years on screen and looks at the webcam as if directly looking at Lemon_drop, thus challenging and maybe checking the validity of her claim.

Management of gaze is determined by many factors, such as context, culture, media, and so on, and, as such, reflections on its use cannot claim universal interpretations. However, some uses of gaze may be intuitively and successfully interpreted, especially in conversations.

This property is complex in video chats and multimodal analysts may have serious trouble in studying gaze in spontaneous CMC. First and foremost, users may choose to hide their own projected image, thus excluding the possibility of using this important discourse facilitator, as shown in Figures 2 and 3. Furthermore, direct eye contact is impossible in video chats, because participants either look at the frame where the discourse partner’s image is projected or at the webcam, thus giving the other the impression of eye contact, as AKAMat in Figure 5. However, virtual gaze is never reciprocal in video chats. If participant A looks at the webcam, participant B perceives something similar to a direct gaze, but this perception cannot be reciprocal, since participant B cannot reciprocate gaze at the same time. Giving the other discourse partner the impression of direct gaze paradoxically excludes the possibility of seeing the other tout court. The problem is solved only if the webcam is embedded in the centre of the screen, but perfect eye contact simulation is very complex in CMC.

Vector analysis is thus complex, since the impossibility of eye contact alters traditional parameters of analysis. Gaze is a facilitator in CMC, but its incidence and role are not easily gauged by analysts. Should gaze be absent, as in the case of telephone conversations, other facilitators, such as prosodic elements, take over, but the presence of nondirect and nonreciprocal gaze requires rethinking of well-established models of visual analysis to be undertaken (Van Leeuwen and Jewitt 2001).

6 Multimodal transcription and annotation

A major problem when analyzing video chat conversations is in transcribing and annotating them, taking into account all the components that we have discussed so far. Researchers into visual data have taken different lines of thinking when addressing different research questions in multimodal analysis (Banks 2001; Van Leeuwen and Jewitt 2001; Norris 2004; O’Halloran 2004). Web-based video chats are less easily classified than other emerging web-based texts. For example, phases, subphases, and transitions (Baldry 2004) are not easy to identify, in that
editing is for obvious reasons absent and a video chat is like a long, uninterrupted single take or tracking shot.

Conventional analysis of gaze vector is also impossible for the reasons discussed in the previous section, but an analysis complemented by a careful reading of conversational threads can be helpful. As regards visual frames, the chief invariants are location and its features, which convey some information to a somewhat smaller extent, in that the image is always delimited and typically fixed. What is in the foreground is a very close or close shot of participants.

However, visual kinesis takes on new connotations in a video chat, because this process is reciprocal and individuals find their visual kinesis realized in the other’s “progressive picture” (i.e., image in movement, Gibson 1986 [1979]). The point of observation is, in fact, fixed and provides a static perspective. However, movements in video chats can be analyzed and transcribed following criteria, which describe configurations of semiotic relationships and define indexical orientation. For the purpose of transcribing and annotating video chat in different forms (e.g., one-to-one or one-to-many), I have developed a transcription model, grounded on multimodal theories of analysis (Sindoni 2013).

Data have been collected through screencasting, i.e., video screen capture or digital recording of computer screen output, also containing audio narration. A screencast is a film reproducing all events that appear on the monitor and, as such, seems to be an effective and flexible strategy to capturing dynamics (e.g., turns and moves) that come into play during a video chat, as has been illustrated in previous sections. This system has the advantage of encapsulating both events pertaining to video chats per se, and other parallel moves the users are likely to perform, for example opening and closing other windows and surfing the web. Consequently, multitasking in CMC lowers levels of attention during web-based and synchronous conversations. This is particularly frequent in the case of multiparty video conversations, where participants often lose the floor or have problems in projecting TRP.

Screencasting also allows the exploration of other issues, such as the general low level of attention during a video chat and especially so when the context is informal and the other user has the (speaking) floor. More specifically, multitasks performed during a video conversation (e.g., checking e-mails or web pages) seem to demonstrate a parallel high level of reciprocal tolerance with regard to the low levels of the other participant’s attention. This view may be partially explained by the consideration that a low level of monitoring of the discourse partner’s gaze direction and impossibility of eye contact weaken reciprocal control.

Special difficulties in handling video data sets are the superimposed written and spoken turns in multiparty video interactions. Users, in both one-to-one and multiparty video interactions, have the chance to use both conversation and writ-
ten comments, as discussed in Section 2. Turns overlap to a considerable extent and subsequent analyses are particularly complex to follow, let alone transcribe.

The transcription model developed for analyzing spontaneous and synchronous video-based interactions is based on transcribing the notions discussed in this paper. It is a time-based model (i.e., unit of seconds), transcribing and annotating spoken turns, written turns, mode switching, kinesics, staged proxemics (see the difference between webcam position and webcam positioning), gaze, and representation of participants (either iconic with drawings, or mimetic with screenshots, see Figures 6 and 7).

This transcription model is centered around the seminal notion of mode switching, which has been identified as the central strand of investigation for these communicative events. However, other approaches can be fully compatible with this model, as it includes also other multimodal features of communication, such as detailed descriptions of kinetic action, postural shifts, gaze management, and use of other semiotic resources, for example the use of visuals, emoticons,
etc. However, as shown in Figure 7, a simplified version of this model that features screenshots of participants can be used. A full-length description of this model goes beyond the scope of this paper, which outlines the main components of analysis for any spontaneous and synchronous video-based interaction.

7 Conclusions

This paper has explored a range of semiotic resources that come into play in the realization of web-based spontaneous and synchronous video interactions. Video chats are texts that challenge traditional notions of textuality in several ways. They include a complex interaction between spoken and written modes of language. Speech and writing alternate in the same communicative event and spoken and written conversational threads overlap or are superimposed in the case of multiparty video interactions.
Other similar examples can be found on the web in asynchronous mode, as is the case of media-sharing platforms, where multimodal texts are created through the assembling of multiple socio-semiotic resources, including speech (e.g., videos) and writing (e.g., comments, posts). However, examples of texts that display alternations of speech and writing in the same communicative event and in synchronous mode are much less frequent. The notion of mode switching is limited to the alternation between spoken and written modes of communication, but other forms of alternation have been found in other forms of textuality on the web, including patterns of alternation with regard to resources other than speech and writing. Such patterns of alternation are less restricted than mode switching and have been defined as resource switching and discussed elsewhere (Sindoni 2013).

The other resources discussed in this paper, i.e., proxemic patterns, kinetic action, and gaze management, indicate that these texts need to be viewed in broad terms, drawing tools of analysis from a range of different fields of studies. However, multimodality is the umbrella theory and practice that can help revise some traditional and well-established notions within the field of mainstream linguistics.

The transcription model briefly mentioned in the previous section has addressed a research question that is relevant within multimodal approaches to communication in general, and in CMC in particular, i.e., the linearization of superimposed spoken and written turns and the incorporation of other nonverbal resources. Transcriptions need to address specific research questions and the model introduced for the study of video chats foregrounds all the resources that contribute to the shaping of the text. However, analysts need to be alerted to the fact that video chats are slippery, deceitful, and contradictory texts. They are live events, but people involved do not share the same context of situation. CMC significantly alters traditional patterns of in praesentia communicative exchanges. Users do not ultimately interact with each other in video chats, but with the medium. The context of situation is both shared and not shared at the same time. Furthermore, the meaning of “context” in virtual environments and the real contribution of contextualization cues in mediated discourse is controversial.

Each participant is, as such, an agent of illusions for others, in orienting their gaze and in co-deploying verbal and nonverbal strategies to create a virtual environment where communication takes place. Ultimately, the backbone of reality is constituted by interaction with a machine. Verbal and nonverbal strategies add to a sense of reality while dissolving it at the same time.
References


Bionote

Maria Grazia Sindoni, PhD, is tenured Assistant Professor in English Linguistics and Translation at the Università di Messina (Italy). She has published several books, including Creolizing Culture (Atlantic: New Delhi, 2006), Systemic-Functional Grammar and Multimodal Studies (Como-Pavia: Ibis, 2011), and Spoken and Written Discourse in Online Interactions: A Multimodal Approach (London & New York: Routledge, 2013). She is the author of articles published in national and international journals and has edited two books. Her research interests include systemic-functional linguistics, multimodality, corpus linguistics, theories of semiosis of communication and computer-mediated interaction.