MIND, CULTURE, AND ACTIVITY, 7(1&2), 105-123

Seeing What One Sees: Perception, Emotion, and Activity

Aug Nishizaka

Department of Sociology
Meiji Gakuin University

In this article, it is demonstrated (a) how seeing is organized in the spatiotemporal arrangement of bodies and conduct within which the participants display and manage their orientations to the ongoing activity, and (b) how seeing and emotion are mutually constituted in the precise coordination of conduct and how they, can constitute resources for organizing the ongoing activity. The view advanced in this article sharply contradicts the traditional conception of visual perception, according to which the verb "see" names a discrete process, event, or state hidden under the individual's skin. Seeing is rather an organizational feature of an embodied, visible activity.

INTRODUCTION

According to the typical cognitivist conception of seeing, when one sees something, first light strikes one's eyes or retinas, then the information it carries is processed in the brain so as to produce a seeing of the thing. For example, Nakayama, He, and Shimojo (1995) begin their chapter contributed to an introductory volume on visual cognition by mentioning "[o]ne of the most striking things":

> Retinal images are formed on the back of our eyeballs, upside down; they are very unstable, abruptly shifting two to four times a second according to the movements of the eyes. … Yet, the visual scene appears to us as upright, stable, and homogeneous. Our perception is closely tied to surfaces and objects in the real world; it does not seem tightly related to our retinal images. (p. 1)

From here, they seek "a critical intermediate stage of vision poised between the earliest pickup of image information and later stages, such as object recognition." It is, they say, "the first stage of neural information processing, the results of which are available to us as conscious perceivers." I do not have anything to say about any particular hypotheses presented by either psychologists or physiologists. However, in what follows, I focus on some essentially important aspects of visual perception that are neglected by the basic conception underlying those hypotheses. The conception is that seeing is a process starting from retinal images or an event resulting from (neural or other) information processing, which occurs under the individual's skin. What I want to show in

Requests for reprints should be sent to Aug Nishizaka, Department of Sociology, Meiji Gakuin University, Minato-ku, Tokyo 108-8636, JAPAN. E-mail: augnish@soc.meijigakuin.ac.jp

this article is that, contrary to this conception, seeing belongs within the public and normative order of activity, rather than taking place under an individual's skin.

In a lecture in 1967, Harvey Sacks indicated that we can see what others think as easily as we see others "eating lunch"; people who have had some university training are specially wont to say, "You don't really know about somebody until you … etc."

> But our language is not built in such a way. Persons use psychological terms with the same freedom and "lack of knowledge of other persons" as they do any other terms. And persons perfectly well figure they know what somebody is thinking; they know why people are doing things, etc., on the same basis that they know that they are "white" or "Jewish," etc., and on the same basis that they know what they're doing. (Sacks, 1964–1972/1992, Vol. 1, p. 558)

We can tell quite easily that someone in front of us is angry and what he or she is angry at. We can tell these things "via what they are doing" (Sacks, 1964–1972/1992, Vol. 1, p. 559). In the same way, we can tell quite easily, too, that someone in front of us sees something or some state of affairs and what he or she sees. Seeing is, in this sense, a public phenomenon, rather than a genuine "mental" one that is basically inaccessible to others. We use vision not only to identify objects, for example, to navigate around rocks located in a rapidly moving stream, and so on. Seeing is also an interactional resource for coordinating actions to complete a given task in a distinct activity; the participants must and do see not only the proper thing exactly at a recognizable specific moment, but also see it in such a way that they can see what and how each other sees.

In the analysis of audio-visual recordings of distinct activities where more than one person is involved that follows, I demonstrate that—and how—seeing is jointly achieved in and through the actual course of an activity.[1] What one sees and how one sees it are interactionally organized both through the temporally unfolding course of interaction and the spatial arrangement of bodies. Then I discuss some implications of the demonstration.

Note that I do not intend to *prove* any hypotheses with empirical data, but rather elucidate, with the help of concrete examples, some aspects of the knowledge of seeing we as members of society already have. In these terms, here are chosen for analysis just those materials that seem to demonstrate well the public and normative character of seeing.

THE INTERACTIVE ACCOMPLISHMENT OF SEEING

Seeing Within an Embodied Activity

The first data to be examined in this demonstration are excerpted from an audio-visual recording of three teenagers jointly playing a computer game. The game, "AlgoBlock," was designed by educational engineers to support collaborative learning.[2] It is unique in that all the commands in the programming language used to move a submarine on a computer screen are given to the players as blocks approximately 10 cm × 10 cm × 10 cm, which are linked to the computer. Players "write"

[1] This viewpoint has already been developed by Charles Goodwin and Marjorie Harness Goodwin (Goodwin, 1994. 1995, 1996; Goodwin & Goodwin, 1996). See also Lynch (1988) and Lynch and MacBeth (1998) for the organization of vision in distinct activities.

[2] For the designers' own description of the system, see Suzuki and Kato (1995).

programs by laying out the blocks on a table in the order they want. Once they have done this, they press the play button connected to the blocks to see if their instructions actually produce the turns and other movements of the submarine on the computer screen as they want. While moving on the screen, the submarine produces a series of beeps; these cease momentarily each time the submarine makes one of its programmed turns. In Fragment 1, the players try to bring the submarine back to its home position through some points on the screen.

Fragment 1 starts when one of the participants presses the play button after they have finished laying out blocks in their second attempt. They have failed in their first attempt, that is, the submarine did not make a turn at the second turning point as they had expected, and, therefore, they had to program the movement of the submarine once again. Whereas on the first attempt they expected the submarine to go straight toward its home position after the second turn, they have decided to let the submarine make another turn, a third turn, to its home position by only adding another few blocks, instead of replacing the "wrong" ones:[3]

#1 (AB: 0:14:58)

Transcript 1 (The Original Transcript and its Phrase-by-Phrase Translation)

1 C: ((Presses the play button to let the submarine start))

2 A: *Yossha.* *Korede* *ikeru* *kana*?=
All right this way go well I hope

3 B: = *Daijoobu,* *daijoobu*
okay okay

4 C: (*Nopposan //doko*)
((untranslatable))

5 A: *Kokoni* *sashikon* *de* …
here insert and

[3] Symbols used in transcripts are:

// A double oblique marker indicates the point at which the next utterance starts.
(1.6) A number in parentheses indicates in seconds and tenths of a second the length of a time interval within an utterance or between utterances.
(.) A dot in parentheses indicates an untimed brief interval (more or less than a tenth of a second).
( ) Empty single parentheses indicate no hearing.
= An equal sign indicates that an item latches on the preceding one.
. A period indicates a stopping fall in tone.
::: Colons indicate that the prior sound is prolonged.
? A question mark indicates a rising intonation at the end of a phrase.

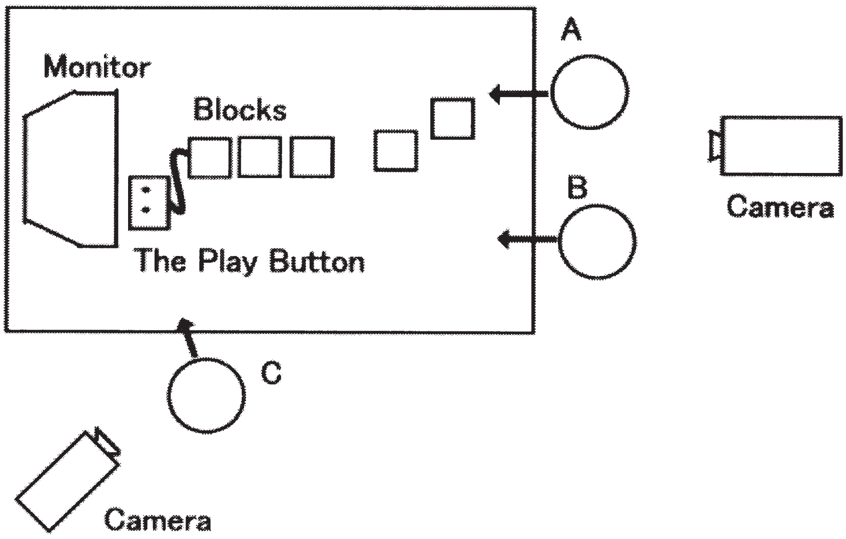


FIGURE 1 The participants, tools, and the location where the participants look at the beginning of Fragment 1, as indicated by the arrows.

Before going into the details of their interaction, it should be noted that all the three participants' bodies are arranged in such a formation that they can appreciate where each other is positioned. Sitting or standing very close to each other, they surround a table on which the items relevant to their activity are placed (see Figure 1). That is to say, they are in a position to be *mutually* oriented to each other's orientations.[4] Now, their faces (or gazes) are directed toward the computer monitor when one of them, C, who is nearest to the button, presses it (at line 1). They can see then that each other is oriented toward the monitor. In and through creating and preserving the very formation in which their mutual orientation is secured to the computer monitor, exactly what to see now (i.e., that they should now see something on the computer screen) is collaboratively made clear to the participants. A "shared" visual field is established on the computer screen in this way. Indeed, they are now running the program they have written with the blocks to see if the submarine is actually moving on the monitor screen as they had programmed.

However, as soon as the monitor is switched on, A and C look away from the computer monitor to those blocks laid out on the table, which is another visual field possibly relevant to their activity in progress (see Figures 2A & 2B):

[4] This kind of formation is called "F-formation" by Kendon (1990). The point here is, however, the establishment and maintenance, not just of a *common* "transactional segment" (each individual's transactional segment being "the space from which [the individual] immediately and readily reaches for whatever objects his current project may require he manipulate," that is, "the place immediately in front of him that the individual projects forward and keeps clear if he is moving" [p. 248]), but rather of a bodily arrangement where the individuals' bodies with their own transactional segments are put within each other's transactional segment. See also Goodwin's (1981) and Heath's (1986) arguments on participation.

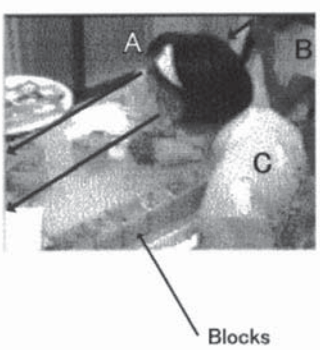


FIGURE 2A They look at the monitor screen.

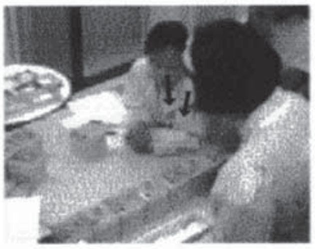
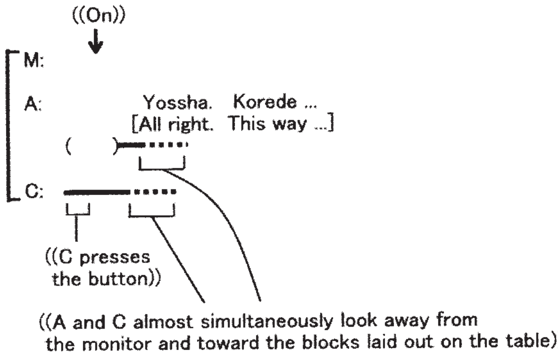


FIGURE 2B They look away from the monitor to the blocks.

#1 (AB: 0:14:58)

Transcript 2[5]



((On))

M:

A: Yossha. Korede ...
[All right. This way ...]
( )

C:

((C presses
the button))

((A and C almost simultaneously look away from
the monitor and toward the blocks laid out on the table))

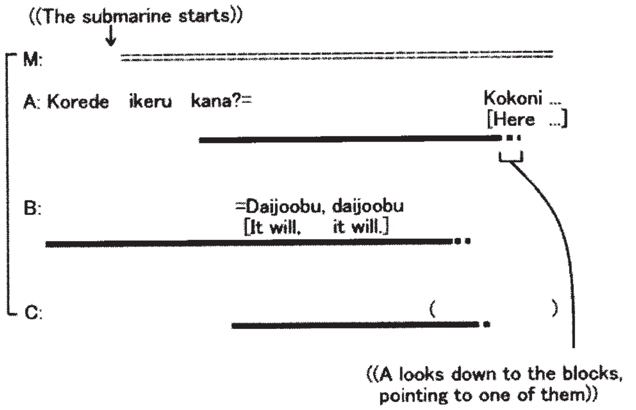
Certainly, all of the participants once look at the monitor just after the submarine starts (i.e., starts to emit beeps), but soon they look away again; A looks toward the blocks again, commenting on their layout, while C, this time, toward A, utters something inaudible:

[5] The following conventions are used in this and the following transcripts:

- All the lines advance in time simultaneously from left to right.
- At lines designated as A, B and C, the participants' utterances are written, immediately under which are their translations in parentheses.
- == at the line designated as M indicates the sound of the submarine.
- Solid lines under each line designated as A, B or C indicate that each participant's face is directed to the computer monitor during the time period corresponding to the length of the lines.

#1 (AB: 0:14:58)

Transcript 3



Why do they not keep their eyes on the visual field that has been established as "shared"? In what follows, I focus on one utterance in Fragment 1, that is, A's utterance made immediately after the button was pressed, "*Yossha. Korede ikeru kana*? [All right. This way everything will go well?]" (at Line 2). It brings into prominence *this* attempt as contrasted with the preceding one in which they failed, especially by starting the substantial part of her utterance with "*korede* [this way]." Immediately after this utterance of A's, B assures A that they will succeed this time, by saying "*Daijoobu* [It will]" (at Line 3). This small exchange between A and B specifies what to see, by contrasting this attempt with the preceding one. Moving her gaze down to those blocks on the table, A refers to the "program" in front of them that they have just written. What they have done this time to the program, that is, the layout of blocks, is only to add those blocks related to the third turn and the subsequent course of the submarine. They did this because their task at hand this time is to let the submarine make another turn after they failed in the preceding attempt to get it to turn toward its home position at the second turn (see Figures 3A & 3B). In view of all this, it is not simply accidental that once they looked at the monitor screen after the submarine started, it never happened that all of them returned their faces to the screen again until it came close to the third turn (although, of course, they turned their faces to the monitor separately during that time). This conduct serves to bring into prominence where what to see lies against the background of other places that are not so important for their purpose at hand.

As noted previously, while moving, the submarine emits a unique sound that stops briefly while the submarine is in the process of making a turn. Therefore, without looking at the monitor, each of the three could know approximately where the submarine was at each moment from the sound, along with the positions of the submarine on the screen they saw the previous time.[6]

[6] It is suggested here that not only vision but all the sensory perceptions are embedded in the current activity. I am indebted to Chuck Goodwin for his drawing my attention to this point. In a personal communication, he indicated in connection with the case being discussed here that "a full 'multi-sensory' environment" is being organized by, and being used as a resource for organizing, the actual course of action. Indeed, not only sensory perceptions but also emotion are sequentially organized within, and sequentially organizes, an activity, as we see shortly.

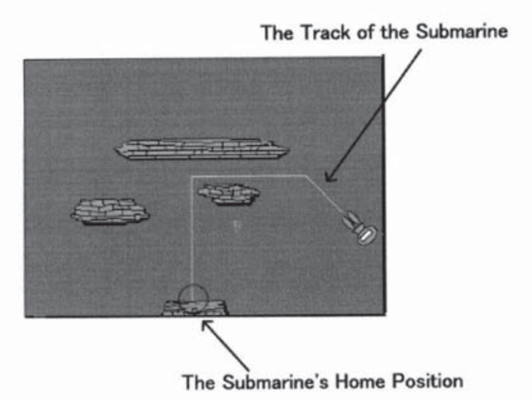


FIGURE 3A An image on the monitor screen. The actual track of the submarine at the first attempt.

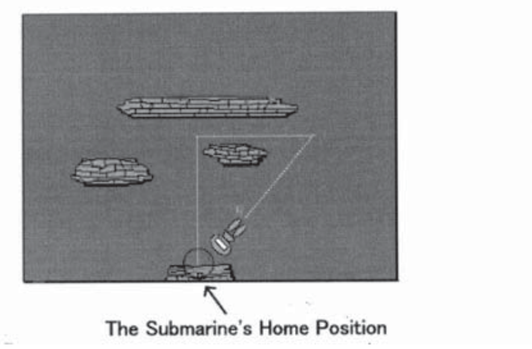


FIGURE 3B The route they wanted at the first attempt.

What to see is thus embedded in the local history of the activity in progress. It depends on what the participants have achieved so far and where they are right now in the course of the ongoing activity. However, on the other hand, it should be kept in mind that this local "history" is only activated through specifically designed talk such as A's utterance at Line 2, which in turn is just part of the activity encompassing its whole history. History and talk (and other relevant conduct, perhaps) elaborate on each other such that a distinct activity is organized as a distinct and unique one.

Now when participants returned their faces to the monitor after the second turning point, all three jointly built up the same formation as when the play button was pressed (see Transcript 2 of Fragment 2):

#2 (AB: 0:15:11)

Transcript 1

1 B: ((Coughs twice))

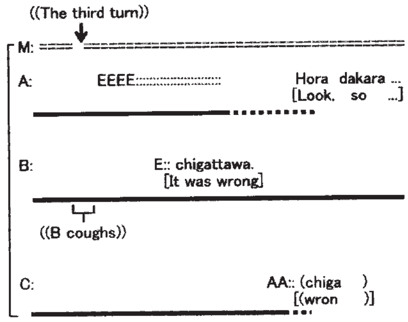
2 A: *EE::::::::::::: //:::::::::*

3 B: *E:: chigattawa.*
it was wrong

4 C: *AA:: //(chiga*)

5 A: *Hora dakara 45 do tteitta jan*
look so degree I said you know

Transcript 2



By building the formation up, the participants can display to each other their understanding that they have now come close to what to see. Moreover, by coughing at almost the same time as when the submarine makes the third turn (at Line 1), B may also possibly indicate to others that here and now is precisely the point where *the* thing lies. All of this corresponds to that exchange of utterances at Lines 2 and 3 in Fragment 1, in which what to see (i.e., the submarine's third turn and its subsequent course) is specified jointly by the participants. Now, it can be said, the fact that what they should see is exactly here and now is achieved through the spatiotemporal arrangement of their bodies and conduct.

As seen from Fragment 2, the participants express their surprise or even shock one after another when the submarine makes the third turn ("EEEE," "EE," "AA" at Lines 2 to 4). The point I want to make here is that not only do they see what to see, but also they see it *in an appropriate way*. In the preceding segment of their interaction reproduced as Fragment 1, when they specified what to see and where to see it, they also specified *how* to see it, that is, how to react when they see what to see. It should be noted that what they should see is not only the submarine's third turn (and its subsequent course) but whether this attempt succeeds or fails, especially at the submarine's third turn. In A's uttering "*Yossha* [All right]" (which indicates the speaker's confidence in success) at the beginning (at Line 2) in Fragment 1 and B's assuring A that it will succeed by saying "*Daijoobu* [It will]" (at Line 3 in Fragment 1), they observably do expect it to succeed. That is to say, if they "see" that it fails, being surprised should be the appropriate reaction; they *should* see its failure with surprise. In fact, they did so, and by doing the "right thing," they display to each other that they now see what to see (see Figure 4). By being surprised, they make it mutually visible that they now see what runs counter to their expectation.[7]

[7] We may be reminded here of Goffman's (1981) observation that emotional expressions (or "response cries") are often used interactionally; although they are not addressed to others, they are very often intended to be heard by copresent others. The most relevant here is Goodwin's (1996) discussion. He remarked that emotional expressions are "organized as social phenomena that provide very powerful resources for shaping the perception and action of others" (p. 393). For the organization of emotional displays as social phenomena, see also Coulter (1979).

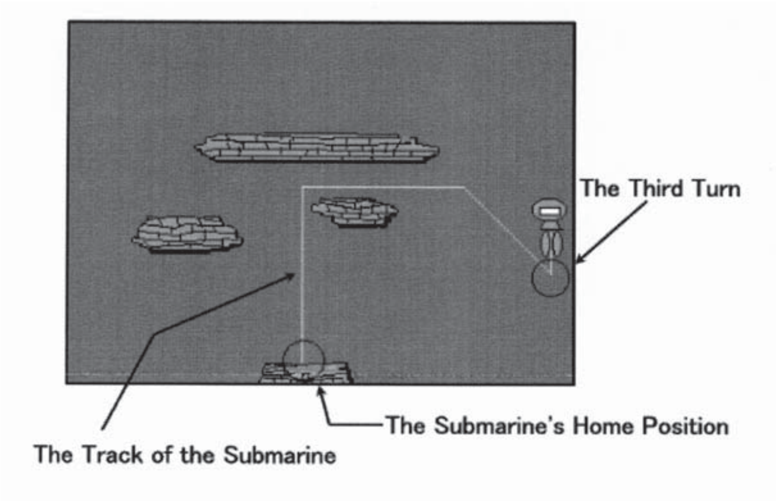


FIGURE 4 An image on the screen right after the submarine took the third turn at the second attempt. The submarine went in an unexpected direction.

Furthermore, it is remarkable that after expressing her surprise, A criticizes B, who took the initiative in designing the rearrangement of blocks at the second attempt, by talking as if she had not expected that the second attempt would succeed ("*Hora dakara ... tte ittajan* [Look, so I told you that ...]" [at Line 5]). Apparently, this contradicts her other conduct, that is, saying "All right" (with falling intonation) before the start and expressing surprise at seeing the failure. Note also that C's expression of her surprise ("A:::" [at Line 4]) is somewhat delayed. It looks as if they are surprised because they should.

Their being surprised together is an interactional resource for displaying and shaping their seeing in ways relevant to their activity in progress, rather than simply being "caused" spontaneously by what happens before their eyes. Being mutually surprised organizes the factual nature and sense of what they see within the historical context of the ongoing activity they are jointly involved in.

The game players saw on the computer monitor that they failed, that is, that the submarine did not go as they had programmed. This natural fact of seeing belongs within the normative order of activity.[8] They saw what they should see in the way they should see it. What to see and how to see it depend on what they are doing. Their seeing can only be meaningful within the actual context of the ongoing activity; it is embedded within the activity they are currently engaged in. Indeed, during the interaction reproduced in Fragments 1 and 2, they turned their faces in various directions, and various patches of light and color must have struck their eyes, but they cannot be said to have *seen* all these things. Seeing must be relevant to the development of the current activity and oriented to by the participants as a part of their activity in progress; their seeing that they failed is rel-

[8] I mean by "normative" what Sacks (1964–1972/1992) calls "programmatically relevant" or "protected against induction." In so far as one sees is what one should see, not just what one see as a matter of fact, the absence of one's seeing becomes observable and accountable. Indeed, if a participant of the computer game had not seen what she saw, that is, that they failed again in the second attempt, she might have been blamed for not seeing it (e.g., "You should've paid attention to the monitor") and some remedial practices might have been in order.

evant both to the failure of their first attempt and to their prospective next attempt; their seeing is retrospectively sensitive to the previous steps of the development of their activity and also prospectively gives them directions as to what and how to do in the next steps. That is to say, what one sees and how one sees it must be appropriate to the actual development of the ongoing activity.

Instructions for Seeing and the Sequential Organization of Emotion

Here we see that seeing is accomplished interactively and sequentially. The exchange between A and B at Lines 2 and 3 in Fragment 1 functions as instructions for seeing. It projects what to see, when to see it, and how.[9] These instructions for seeing make the sequence of the participants' emotional expressions an appropriate response to what they see then. Instructions for seeing generally organize the subsequent course of interaction in such a normative way that even less conventional emotional expressions are provided with specific meaning and accountability in reference to those instructions. In fact, vision and other perceptions, emotion, (verbal and nonverbal) conduct, and other phenomena are mutually organized in and through the actual development of interaction to build an "activity system." How emotion and vision are mutually organized is clearer in another fragment, which is excerpted from an audio-visual recording of a private Japanese word-processing lesson.

In this lesson, the instructor gives tasks to the learner, and the learner attempts to complete them. The result of each attempt by the learner is commented on by the instructor, that is, the proper way of operation is described, why she has failed is explained, and so on. The task in the fragment being reproduced in the following is to input the string of letters "IBM" into a computer in a special way. The instructor was a part-time instructor at the Information Center at a university when the session was recorded.

Fragment 3 starts when the instructor (A) gives the learner (B) a new task: Input a "half-sized IBM." The task given just prior to this one, which the learner has just completed, was to input a "full-sized IBM." All the Japanese word-processing devices have two types of Roman letters, full-sized and half-sized. What we call half-sized letters are ones we use when writing in English or other European languages; full-sized ones are specially used for Japanese kana and kanji characters. ("Full-sized" means "of the same size as normal Japanese characters.") Therefore, as long as writing in Japanese with a computer, we usually use full-sized letters, and inputting half-sized ones requires one extra operation. In these terms, it is quite reasonable to give the task of inputting a half-sized IBM only after that of inputting a full-sized one in a Japanese word-processing lesson. The interactional relevance of this point is mentioned shortly.

#3 (WP: 0:42:50)

Transcript 1-a (A Free Translation of the Japanese Transcript)

1 A: Then now la(rge [?])- This is (.) a full-sized "IBM," right?
2 B: Yes
3 A: Then, try to input a half-sized "IBM." An application.

[9] See Goodwin's (1996) discussion on what he called "prospective indexicals."

4 : (7.4)
5 A: Yeah. (.) O?
6 B: O?
7 : (0.2)
8 B: (*ne*)
9 A: *Aaaa:::.* That's the way. So because pressing key number 8 first doesn't work, …

Transcript 1-b (The Original Transcript and a Phrase-by-Phrase Translation; see Figure 5)

1 A: *Jaa ima oo- Korettesaa,* (.) *zenkakuno "IBM" desu yone:::.*
then now this full-sized is
2 B: *Hai*
yes
3 A: *Hankakuno "IBM" jaa irete mite. Ooyoo*
half-sized then input try application
4 (7.4)
5 A: *Un* (.) *Are?*
yeah o
6 B: *Are?*
7 (0.2)
8 B: (*ne*)
9 A: *Aaaa:::. Soodesne. Dakara, ikinari* (.) *hachi banwo* (.)
that's the way so first eight number
oshi temo ( ) *kara::*
press though because

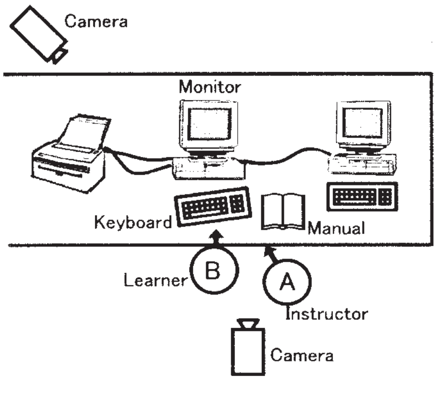


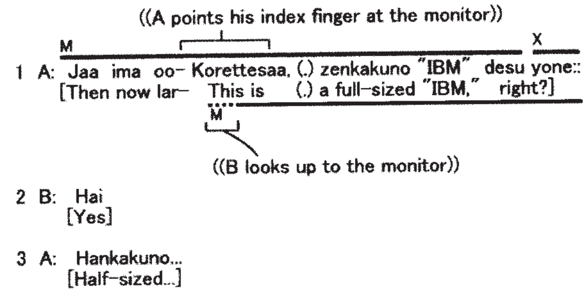
FIGURE 5 The participants' positioning of their bodies and tools on the table.

At Line 3, A, the instructor, gives B, the learner, an exercise, which functions also as instructions as to what to see and how to see it.[10] At Line 9, A reacts (to what he sees) in an appropriate way with an emotional expression "*Aaaa:::*," which has a unique sound contour, that is, starting with a high pitch and ending with the pitch falling down sharply. This expression achieves a special interactional job here precisely through being produced at a particular sequential position, and precisely in doing this job it is provided with special accountability.

Before going on to a closer examination of the sequential organization of emotion, it should be noted that the force of those instructions for seeing is interactively and sequentially achieved as such. In the case of Fragment 1, as we saw, the exchange of talk between two participants constituted instructions for seeing. Moreover, the exchange was positioned in the actual context of interaction; that is, the exchange was made right after a shared visual field where the thing they should see was established and the play button was pressed. The exchange is also embedded within (while at the same time activating) the local history of the joint activity in progress. In Fragment 3, too, the instructions for seeing are positioned in the actual context of interaction in some way:

#3 (WP: 0:42:50)

Transcript 2[11]



((A points his index finger at the monitor))

M X

1 A: Jaa ima oo– Korettesaa, (.) zenkakuno "IBM" desu yone::
[Then now lar– This is (.) a full–sized "IBM," right?]

M

((B looks up to the monitor))

2 B: Hai
[Yes]

3 A: Hankakuno...
[Half–sized...]

When A starts to give B an exercise, B looks down at the table or the keyboard. Then A interrupts his own utterance and points his index finger at the computer monitor in front of B (see Figure 5) while restarting his utterance with the demonstrative "*kore* [this]." This self-interruption and pointing at the monitor induces B's gaze to the monitor, and B gives A an explicit verbal response ("*Hai*") to A's utterance. It is only in this interactional context that A moves on to giving B the exercise ("*Hankakuno* ... [Half-sized ...]"). Thus, the instructions for seeing in Fragment 3 are also produced in the arrangement of bodies in which both of the participants display to one another their orientation to the monitor.

It should be noted, too, that here again the exercise A gives is contrasted to what has already been attained. In A's mention of the full-sized "IBM," the task of inputting a half-sized "IBM" is

[10] The words "*Hankakuno* [half-sized] IBM" function as what Goodwin (1996) called a "prospective indexical." As seen later, their "referent" will have to be fixed on the monitor in a specific way, that is, as a successful outcome of the learner's operation.

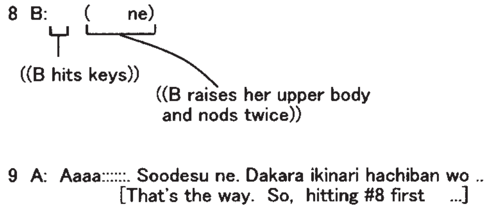
[11] Solid lines led by M's and X's just above or under each utterance indicate that A's or B's face is directed to the monitor and their coparticipant respectively.

made prominent as a next step against the full-sized one already on the monitor screen. Furthermore, I guess that the self-interrupted phrase should be "*ookii* IBM" or "*ookina* IBM," which means "large IBM" or the like; as you see now, full-sized letters are larger-sized compared with half-sized ones. Taken together, A's and B's gazes are jointly focused not just on one visual field, but on the visual field specifically relevant for the next step of their activity. Again, instructions for seeing are provided in a relevant fashion and at a particular place and time in reference to the historical context of the ongoing activity the participants are jointly engaged in.

Now, let us turn to the responses with emotional expressions. As for Fragment 2 (from the AlgoBlock material), I indicated previously that all the participants expressed their surprise in the special bodily arrangement (where they can display to each other their orientations to the monitor) in an orderly way, that is, that their being surprised was finely attuned to each other's conduct rather than a spontaneous, natural expression of a mental state. A's production of an emotional expression in Fragment 3 (from the word-processing lesson) is also very precisely coordinated with B's conduct:

#3 (WP: 0:42:50)

Transcript 3



8 B: (ne)

((B hits keys))

((B raises her upper body
and nods twice))

9 A: Aaaa::::::. Soodesu ne. Dakara ikinari hachiban wo ...
[That's the way. So, hitting #8 first ...]

After hitting some keys, B, the learner, raises her upper body, looking at the monitor, and nods twice, while saying something (inaudible from the recording). This conduct, it looks, marks out the completion of her operation. (In doing so, the learner shows that she recognizes those letters as the ones she was assigned to input and ties them back to the task assigned by the instructor.) It is only immediately after the completion of the current task has thus been marked out that A, the instructor, produces the emotional expression "*Aaaa"*. This precise coordination of conduct marks out exactly what they see.

Certainly, A must have seen the "IBM" independently of B's recognition of it, because he had been looking at the monitor while she was operating on the keyboard. However, the string of letters "IBM" is not all they see. Insofar as their seeing is linked up to B's performance of the exercise, they do not see a mere string of letters "IBM" any more than those game players in Fragment 2 saw just a mechanical movement of the submarine. They see that B has succeeded in inputting a half-sized IBM and completing the assigned task, just as did those game players see that they had failed in letting the submarine come back to its home base again. Both a mere string of letters in this case and a mechanical movement of the submarine in Fragment 2 would be rather an artificial abstraction detached from our lived world.

In this connection, attention should be drawn to those remarks A made subsequently to his emotional expression ("*Aaaa*"). He starts his remarks with "*Soo desune. Dakara* ... [That's the way. So …].” Those words, referring by a kind of demonstrative (“*soo* [that]”) to what has been in their perceptual field of mutual orientations, mark out that A is now going to give a review of what A and B have jointly perceived (with what can be called a “summing-up” token, i.e., “*dakara* [so]”). Then A goes on to comment on, in a general way, what was wrong and what was right in B’s operation (i.e., that pressing key number 8 first does not work and that key number 9 should be pressed first, etc.). In this way, what he and B have just (visually) perceived on the monitor screen is accounted for and formulated as a result of B’s operation on the keyboard.

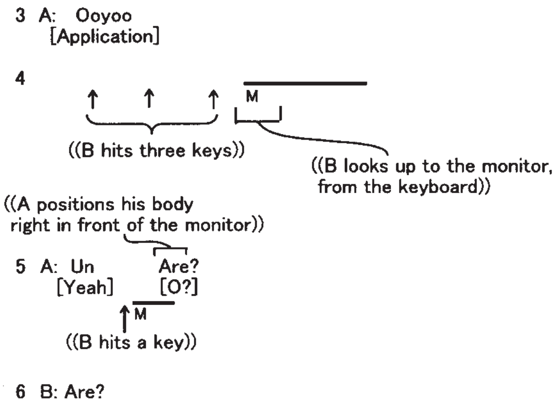
A’s production of an emotional expression achieves an interactional job of marking out the visibility of B’s success in inputting a half-sized IBM, not just the mere string of letters “IBM,” and precisely in doing this job, the sound with that unique contour (“*Aaaa:::*”) can be an expression of being impressed by B’s performance.

Seeing Through Being Embarrassed Together

Of course, it is quite rare for the instructor to express his being impressed in such a marked way in response to the learner’s success in completing a task. There is a reason why he did so there. The following is a detailed transcript of the middle part of Fragment 3:

#3 (WP: 0:42:50)

Transcript 4



3 A: Ooyoo
[Application]

4

((B hits three keys))

((B looks up to the monitor, from the keyboard))

((A positions his body right in front of the monitor))

5 A: Un Are?
[Yeah] [O?]

((B hits a key))

6 B: Are?

Here, too, producing an emotional expression plays an interactional role in the organization of seeing. After hitting three keys in response to A’s assignment, that is, making three letters appear on the monitor, the learner (B) shifts her gaze from the keyboard to the monitor screen. Here some delay in hitting function keys has become noticeable. A’s “*Un* [Yeah]” marks it out, encouraging B to go ahead while claiming there is no problem so far.[12] In this interactional context, not just some string of three letters, but an “incomplete state of operation” becomes visible on the monitor.

[12] See Schegloff’s (1982) discussion on continuers.

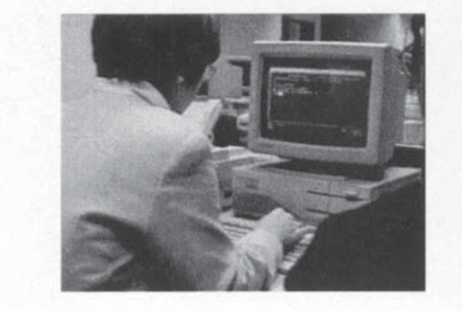


FIGURE 6A The learner has just hit a key and looked up to the monitor (at Line 5).

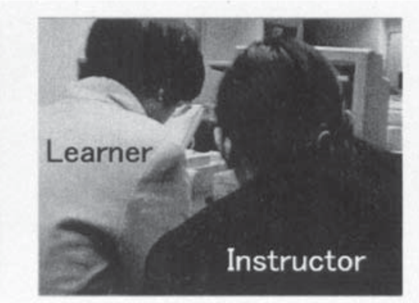


FIGURE 6B The instructor has brought his upper body in.

Then B hits a key. Immediately A and B say “*Are*?” one after another in response to what has been made on the monitor by B’s hitting the key. “*Are*?,” a rather conventional expression of surprise, especially used when something unexpected has happened, indicates that something is wrong. A and B are, as it were, jointly being embarrassed at what is going on, which is, in reference to the given instructions for seeing, an appropriate thing to do when a half-sized “IBM” did not appear.

Their production of the expression “*Are*?”, too, is interactively and sequentially finely coordinated. A, while saying “*Are*?,” brings his upper body in front of the monitor, as if he were inspecting what is going on (see Figure 6A & 6B). The real reason for his bodily movement is not important. What is important is that he is searching for the problem publicly and observably in front of B. B produces “*Are*?,” it seems, in response to this whole conduct of A’s. By doing this, B shows that she understands A has seen something wrong on the monitor screen at the same time as she shows she recognizes something wrong for herself (and in so doing, displays that, despite having made an error, she knows what counts as competent performance in this task). A’s and B’s being embarrassed together now makes the visibility of B’s failure in her operation available and embodied on the monitor. This visibility of B’s failure encourages her to make another attempt on the keyboard.

Back to A’s expression of being impressed (at Line 9): It apparently responds to this publicly visible failure. In reference to this visible failure, his expression of being impressed is organized such that B’s success is now visible against that failure that was visible a moment ago.[13] Both the visibility of success and that of failure are relevant and useful in unfolding the ongoing activity A

[13] In this way, emotional displays “are embedded in local contexts of social action” and in those contexts, “achieve meaning and import” (Whalen & Zimmerman, 1998, p. 158). The demonstration in this section is another example of “the integration of the study of emotional display with the sequential organization of talk-in-interaction.”

and B are now jointly engaged in, and emotion is organized in the same sequential order of the ongoing activity in which vision is organized. Vision and emotion are mutually organized within the actual context of an activity.

SEEING AS A VISIBLE PHENOMENON

Seeing Relevant to the Ongoing Activity

We have seen that seeing is a public and normative phenomenon, which is achieved in and through the actual course of a distinct activity. I do not deny that when one sees something, some physiological processes or events take place under one's skin and that these processes or events are a very important area of the study of visual perception. I do deny that those processes or events *are* seeing or visual perception. What puzzles me about the orthodox conception of vision is that it seems to fail to take into consideration the fact that we can see the lack or absence of something; indeed, the game players in Fragment 2 *saw* that the submarine had *not* gone the way that they expected it to go, and the instructor and the learner in Fragment 3 (at Lines 5–6) *saw* that a half-sized "IBM" did *not* appear on the monitor. It is implausible that there are physiological conditions *specific* to the lack of something or some fact, although there might occur some physiological changes.

Probably, Harvey Sacks is the first sociologist who was seriously surprised at the fact that we often use such expressions as "is not," "do not," "none," "nothing," and the like. Generally speaking, there are an infinite or indefinite number of things that someone does not or did not do. Therefore, when one says "He did not do …" or "She is not doing …," one does not say this just because it is true (Sacks, 1964–1972/1992). For example, we sometimes say something like this: "She did not greet anyone." It is also true that I have not greeted anyone for the past couple of hours; I have been by myself at my office. Do we say, however, "I did not greet anyone"? It is when the statement is embedded in the specific context of an activity, such as the distinct activity of opening up an encounter, for example, when some people have just greeted her, that we can say she did not greet anyone.

The same holds true also for other kinds of statements. In one lecture held in 1966, Sacks (1964–1972/1992) cited the following example. In midst of a group therapy session, one boy uttered the words "We were in an automobile discussion." It is true that they had been discussing automobiles. Obviously, however, he did not produce this statement precisely at that time just because it was true; indeed, there must have been an indefinite number of things the participants had been doing prior to his statement. The statement was produced just after a newcomer had been introduced by the therapist to those present. According to Sacks, automobiles are generally (or normatively) expected to be accessible as a common topic for teenagers and, by uttering those words precisely at that time, the boy invited the newcomer, who was also a teenager, to join their discussion, showing him that they had been talking about very ordinary things any teenager could be interested in even if that one may not be interested in automobiles at all actually. Thus, the fact that he said what he said right there and then is embedded in the distinct activity of inviting a newcomer to the ongoing interaction.

In the same way, although it is possible to say that the participants in the word-processing lesson saw that a string of words did not appear, it does not make any sense to say that they saw that

Mozart did not appear, even though it is true that Mozart did not. What one sees is embedded in the activity one is engaged in, for example, performing an exercise or whatever. I saw my colleagues at a faculty meeting today, but did I see their eyelids? When voting last time, I saw a quadrangular space surrounded by black lines on a slip of paper, inside which I wrote down a candidate's name; but did I see those lines were slightly crooked at several places if they were actually so? A young baby saw a ball rolling up to her and tried to hold it, but did she see also a lot of stains on the surface of the ball?

This said, however, an activity is not something like a container for seeing. Not only is seeing lodged in an encompassing activity. What I have demonstrated is that seeing is organized through the precise and fine coordination of the participants' conduct. It is not that the participants' current activity, for example, playing a computer game jointly, lies somewhere above and over their actual conduct and constrains it from the outside; playing a game is accomplished jointly in, through, and as the spatiotemporal arrangement of their bodies and conduct. An activity is organized as a distinct one through the mutual organization of (visual and other) perceptions, emotions, and other various kinds of things in which the participants display and manage their orientations to that very ongoing activity. Although seeing is accomplished within the actual arrangement of bodies and conduct that constitutes the ongoing activity, seeing is also an interactional resource for (re)organizing the actual arrangement of bodies and conduct.[14]

The Fallacy of Reification

The orthodox conception considers the verb "see" to refer to some process or event or some state inside an individual. However, the word is not the name of a process or occurrence, a state, or even an activity. The orthodox conception is caught up in what Coulter (1989) called "the fallacy of reification." In Ryle's (1949/1963) terms, the verb "see" is an achievement word: "'see', 'descry', and 'find' are not process words, experience words, or activity words. They do not stand for perplexingly undetectable actions or reactions, any more than 'win' stands for a perplexingly undetectable bit of running, or 'unlock' for an unreported bit of key-turning" (pp. 145–146). One criterion for being an achievement verb is that "in applying an achievement verb we are asserting that some state of affairs obtains over and above that which consists in the performance, if any, of the

[14] In this connection, it is worth mentioning Coulter and Parsons' (1991) criticism of J. J. Gibson's ecological approach to visual perception. Gibson's (1979/1986) criticism of the orthodox psychology theory of visual perception was so radical that he went so far as to argue "perception of the environment is direct" and that "it is not mediated by *retinal* pictures, *neural* pictures or *mental* pictures" (p. 147). However, his approach is still so orthodox that he does not take into consideration the "embedded-in-activity" (i.e., normative) character of vision. At least in the human case, verbs covering visual perception vary very widely, including "see," "observe," "notice," "read," "examine," and so on. According to Coulter and Parsons, Gibson does not pay enough attention to this variety in modalities of visual perception. They remarked, "Regardless of which modality is invoked, displayed, or presupposed in members' activities, to stipulate a continuity in our visual orientations is to violate logical grammar" (p. 263). The various modalities of our visual orientations constitute what Wittgenstein (1958) called "family resemblance"; they do not have any characteristics in common; there are no general properties of visual orientations. Then, how should we be able to speak of visual perception in general? Any attempts to construct a general theory of visual perception seems doomed to failure. All that is left to us is to examine in detail how vision is organized in members' activities of various kinds. For a criticism of the Gibsonian approach, see also Ueno (1996). Ueno indicated a direction in which the Gibsonian ecological approach could develop in a productive way, emphasizing the social character of human vision.

subservient task activity" (Ryle, 1949/1963, pp. 143–144). The fallacy consists in searching inside the individual for the referent of the verb "see" because obviously the verb does not refer to any observable process or state. However, not only does it not refer to any *observable* process or state, but it does not refer to *any* process or state. As Coulter remarks, achievement words presuppose the demonstrable facticity or accuracy of what it is one is claiming to have seen, found, and so on. They "do not name discrete states, events or 'phenomena' susceptible to first-person 'revelation,' 'observation' or 'internal monitoring' of any kind. Claims to facticity are dismissible, defeasible, by public recourse to convincingly established counter-evidences of all sorts" (Coulter, 1989, p. 121). On the other hand, as I said previously, neural conditions for visual perception are not visual perception per se, but just its conditions. The verb "see" does not name any process or activity that takes place under an individual's skin.

Incidentally, it is very misleading to describe physiological processes accompanying seeing as a kind of information processing. It is well known that Kenny (1984) pointed out an absurdity resulting from applying to a part of a person or an organism (e.g., the nervous system) those predicates that can be only applied to a person or an organism as a whole in its ordinary use. He called this confusion "the homunculus fallacy." The trouble is this: If for us to do some activity, information has to be processed by the nervous system, that is, a homunculus, then for the latter to do this distinct activity of information processing, another homunculus has to process information inside the nervous system. It is easy to see here the absurdity of regress ad infinitum.

I will not go into philosophical arguments here. The point I want to make here is that we are often engaged in a distinct activity of information processing, the performance of which requires one to see various things. The game players in Fragments 1 and 2 saw the failure of their second attempt, that is, they extracted visual information from the environment organized through their coordinated conduct and processed and transformed it into information as to how to lay out the blocks. Information processing is a publicly *seeable* activity that the participants are mutually oriented to. Visual perception is not simply a discrete state resulting from information processing, but rather a resource for the activity of information processing. As Sharrock and Coulter (1998) emphasized, "to gather/obtain information of many kinds *presupposes* (and thus cannot *explain*) seeing" (p. 156). To call even metaphorically those physiological processes accompanying visual perception "information processing" seems to have caused unnecessary confusions and misconceptions. Particularly, it induces us to conceive the word "seeing" as a label of some hidden activity or its resulting state.

A Consequence to Human–Machine Interaction Studies

One should not suppose in advance of analyses that there are two different kinds of entities, that is, human beings on the one hand and tools and other objects on the other, and attempt to explore the interaction of these entities. As has been shown, even such a simple object as a half-sized "IBM" on the computer monitor has its visibility embodied in the actual arrangement of bodies and conduct in which the participants' visual and other orientations are displayed and managed. It is, as it were, the participants' "extended body." It does not stand by itself out in the world but constitutes, together with human bodies and other artifacts, talk and other conduct, and so on, an activity system. Seeing is not a processing of that information that comes from objects in the outer world into the human body, but a structural feature of an activity system.

ACKNOWLEDGMENTS

Many thanks to Dom Berducci, Jeff Coulter, Chuck Goodwin, John Hindmarsh, Mike Lynch, Lucy Suchman, and Naoki Ueno for their encouragement and helpful comments on earlier drafts. I am very grateful to Hiroshi Kato and Hideyuki Suzuki for their permission to use the AlgoBlock material, and to Satoyuki Morita for his corporation in collecting the word-processing lesson data. I wish to thank Chuck Goodwin also for his very detailed editorial comments, which were of great help.

REFERENCES

Coulter, J. (1979). *The social construction of mind.* London: Macmillan.

Coulter, J. (1989). *Mind in action.* Cambridge, England: Polity Press.

Coulter, J., & Parsons, E. D. (1991). The praxiology of perception: Visual orientations and practical action. *Inquiry, 33,* 251–272.

Gibson, J. J. (1986). *The ecological approach to visual perception.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc. (Original work published 1979)

Goffman, E. (1981). *Forms of talk.* Philadelphia: University of Pennsylvania Press.

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

Goodwin, C. (1994). Professional vision. *American Anthropologist, 96,* 606–633.

Goodwin, C. (1995). Seeing in depth: Space, technology and interaction on a scientific research vessel. *Social Studies of Science, 25,* 237–274.

Goodwin, C. (1996). Transparent vision. In E. Ochs, E. A. Schegloff & S. A. Thompson (Eds.), *Interaction and grammar* (pp. 370–404). Cambridge, England: Cambridge University Press.

Goodwin, C., & Goodwin, M. H. (1996). Seeing as situated activity: Formulating planes. In Y. Engeström & D. Middleton (Eds.), *Cognition and communication at work* (pp. 61– 95). Cambridge, England: Cambridge University Press.

Heath, C. (1986). *Body movement and speech in medical interaction.* Cambridge, England: Cambridge University Press

Kendon, A. (1990). *Conducting interaction.* Cambridge, England: Cambridge University Press.

Kenny, A. (1984). *The legacy of Wittgenstein.* Oxford, England: Basil Blackwell.

Lynch, M. (1988). The externalized retina: Selection and mathematization in the visual documentation of objects in the life sciences. *Human Studies, 11,* 201–234.

Lynch, M., & MacBeth, D. (1998). Demonstrating physics lessons. In J. G. Greeno & S. V. Goldman (Eds.), *Thinking practices in mathmatics and science learning* (pp. 269–298). Mahwah, NJ: Lawrence Erlbaum Assocites, Inc.

Nakayama, K., He, J. Z., & Shimojo, S. (1995). Visual surface representation: A critical link between lower-level and higher-level vision. In S. M. Kosslyn & D. N. Osherson (Eds.), *Visual cognition* (pp. 1–70). Cambridge, MA: MIT Press.

Ryle, G. (1963). *The concept of mind.* Middlesex, England: Peregrine Books. (Original work published 1949)

Sacks, H. (1964–1972/1992). *Lectures on conversation* (Vol. 1 & 2). Oxford, England: Basil Blackwell.

Schegloff, E. A. (1982). Discourse as interactional achievements: Some uses of “uh huh” and other things that come between sentences. In D. Tannen (Ed.), *Georgetown University roundtable on languages and linguistics* (pp. 71–93). Washington DC: Georgetown University Press.

Sharrock, W., & Coulter, J. (1998). On what we can see. *Theory & Psychology, 8,* 147–164.

Suzuki, H., & Kato, H. (1995). Interaction-level support for collaborative learning: AlgoBlock—an open programming language. *Proceedings of CSCL ‘95,* 349–355.

Ueno, N. (1996). Jokyo ninchi to Gibson [The situated cognition and Gibson]. *Gengo, 25*(1–6).

Whalen, J., & Zimmerman, D. H. (1998). Observations on the display and management of emotion in naturally occurring activities: The case of “hysteria” in calls to 9-1-1. *Social Psychology Quarterly, 61,* 141–159.

Wittgenstein, L. (1958). *Philosophical investigations.* Oxford, England: Basil Blackwell.