Chapter for: J.M. Carroll, "Toward a multidisciplinary science of human-computer interaction"

"Common ground in electronically mediated communication: Clark's theory of language use"

Andrew Monk, University of York, UK
Department of Psychology
University of York
York, Y010 5DD, UK
A.Monk@psych.york.ac.uk

HCI has come to encompass technologies that mediate human-human communication such as text-based chat or desk-top video conferencing. The designers of equipment to electronically mediate communication need answers to questions that depend on a knowledge of how we use language. What communication tasks will benefit from a shared whiteboard? When are text messages better than speech? Thus the theory that informs the design of these artefacts is a theory of human-human interaction.

Previous theories of language use divide into the cognitive and the social. Most psycholinguistic accounts of language production and comprehension are very cognitive. They are solely concerned with the information processing going on in an individual's head. Ethnomethodological and other sociological accounts of language use are, in contrast, social. They concentrate on the structure that is observable in the behaviour of groups. Herbert Clark has developed a theory of language use that bridges these two camps and that can make practically relevant predictions for the design of facilities to electronically mediate communication.

The key concept in Clark's theory is that of common ground. Language is viewed as a collaborative activity that uses existing common ground to develop further common ground and hence to communicate efficiently. The theory: (i) defines different kinds of common ground; (ii) formalises the notion of collaborative activity as a "joint action", and (iii) describes the processes by which common ground is developed through joint action.

The next section in this chapter explains why a purely cognitive model of communication is not enough and what is meant by the phrase "collaborative activity". Section 2 introduces the idea of common ground and how it is used in language through an example of two people communicating over a video link. Section three indicates where the interested reader can find out about the antecedents to Clark's theory. Section 4 sets out the fundamental concepts in Clark's theory. Section 5 uses three published case studies of mediated communication to illustrate the value of the theory.

1. Motivation

Previous chapters have been concerned with understanding how humans interact with computers, as would seem quite proper in a book on theory in human-computer interaction. However, the discipline of HCI has come to include the study of electronic devices for the purpose of communication, e.g., video conferencing systems, text-based chat and email. Some of the questions designers need to answer about these systems are to do with human-computer interaction, e.g., how to use the limited display on a mobile phone, but others are to do with the way that we use language, e.g., what communication tasks will benefit from a shared whiteboard. The theory that answers these questions is a theory of human-human communication.

A common view of human-human communication conceptualises language as a sender producing some utterance that is then comprehended by a receiver. While this has value it is not the whole story.

1.1 Production + comprehension ≠ communication

The upper part of Figure 1 depicts a much simplified model of how two computers communicate with one another. Computer A sends the sequence of characters forming an email message by looking up a digital code for each letter. Each digital code is then translated into a pattern of voltage changes on a wire. Computer B reverses this process. It registers the pattern of voltage changes, converts this into a digital code and looks up the letter. When enough letters have been accumulated it can display the email. This conception of information transmission was used by Shannon & Weaver, (1949) to formulate a mathematical theory of communication that has been used by communication engineers for many years.

- - - - - -

Figure 1 about here

- - - - - -

The lower part of Figure 1 takes the information transmission model as an analogy for human-human communication. Some representation of the meaning of a word in Person A's head is looked up to find its phonemic representation and that is then converted to sound pressure changes in the air by Person A's vocal apparatus. Person B's ear registers these pressure changes and auditory processing in B's brain converts them to a phonemic representation and then to a representation of the meaning of the word.

This information processing model allows one to decompose the process of communication into two parts, speech production and speech comprehension. Speech production is the process of converting meaning to sound pressure changes and speech comprehension is the process of converting speech pressure changes back into meaning. Figure 1 is a very simplified version of current understanding. The linguists, psycholinguists and speech scientists who study what goes on within each of these two processes have developed sophisticated

models hypothesising many different representations that may be generated along the way (see for example Altmann, 1997).

The models developed have resulted in many practical advances. Research on speech comprehension has led to improvements in digital hearing aids and speech recognition software. The research on speech production has led to speech synthesis software and speech therapy programmes for stroke victims. This approach to languages use has however proved less useful in providing guidelines for the design and configuration of electronic communication systems. For example, if one is designing a video conferencing configuration, should one use the camera to convey as much information as possible about detailed facial expression and lip movements of the person currently talking, or would it be more valuable to provide a wide angle view of what everyone at the other end is doing? When does text have advantages over speech?

The problem is that models of speech production and speech comprehension are cognitive models. They are models of what goes on in an individual's head. It turns out that to answer the questions posed above we need a social model, i.e., a model of how a pair or group of individuals use language as an ensemble. It is not intuitively obvious why this should be so. Common sense says that if we have a model of how a speaker produces speech and another of how a listener comprehends it then it should be possible to simply put them together to form a model of language use. The next section explains why we need something more.

Roger:	Did you have oil in it	
Al:	Yeah, I-I mean I changed the oil, put new oil filters, r-completely redid the oil system, had to put new gaskets on the oil pan to stop-stop the leak, and then I put -and then-	
Roger:	That was a gas leak	
Al:	It was an oil leak buddy	
Roger:	It's a gas leak	
Al:	It's an oil	leak!
Roger:		on the number one jug
Al:	It's an oil leak!	
Roger:	Outta where, the pan?	
Al:	Yeah	
Roger:	Oh you put a new gasket on it stopped leaking	
Al:	Uh huh	

Table 1. A snippet of real conversation (Jefferson, 1987, p. 90.)

1.2 Language use as a collaborative activity

Consider the conversation recorded in Table 1. Al has been mending Roger's car. Roger comes to the conversation thinking that the problem involved a petrol ("gas") leak. Al has just fixed an oil leak. What follows is a process of re-

alignment. This language process is described as "repair". It starts when Roger senses "trouble" in the conversation because Al is talking about fixing the oil system. He signals this to Al with the utterance "that was a gas leak". They then collaborate until conversational repair is achieved. Eventually Roger signals that he now sees there was an oil leak by saying "outta where, the pan?", the pan being the oil sump. Al then signals that he understands that Roger now understands this with his utterance "Yeah".

This is very different to the picture of communication presented in Figure 1. First of all notice how ill formed and imprecise the utterances are with repetitions and re-starts (e.g., "r- completely re-did"). There is also overlapping speech. The tabulation in Table 1 shows that "leak!" and "on the number one jug" were overlapping in time. Al and Roger get away with this imprecision because communication is a collaborative activity not just a matter of using a well defined code to replicate the contents of one person's head in another's.

Al and Roger come to the conversation with different assumptions and priorities. They go away with different assumptions and priorities but they have developed sufficient common ground to serve each of their separate purposes. The conversation is a collaborative process in which they each endeavour to communicate sufficiently for their own purposes. At the same time they monitor the conversation for evidence that the other person is or is not communicating sufficiently well for their purposes. Thus each of them has an obligation to signal to the other when they sense communication is failing. Each of them has an obligation to monitor the conversation for such signals and to take appropriate actions to repair the situation until the other signals all is now well. This mutual pact is the basis of every conversation.

We can now see what the information transfer model depicted in Figure 1 is lacking. Communicating computers have a common code. This is possible because the code is well defined and can be programmed into both computers by engineers. In contrast everyday spoken language is very ambiguous and only works because the parties actively collaborate to make it do so. Experience may have programmed you and I with the same rules for converting sounds into phonemes and for combining phonemes into words. However, when it comes to communicating intent or history I cannot just look up a recipe that will copy what is in my head into yours. Nor would I want to. Spoken language use is efficient precisely because only the information relevant to each individual's separate needs is communicated.

The above is the starting point for the collaborative model of conversation assumed by Herbert Clark. The remainder of this chapter is to describe his theory in more detail and to illustrate how it can be used to explain various observations about electronically mediated human-human communication.

2. Overview

This section introduces the notion of common ground and how it is used in language. Section 4 contains a more detailed treatment of the other fundamental concepts in Clark's theory.

Clark's theory is based around the concept of common ground, that is the things we know about what is known by the person we are talking to. If this seems rather recursive that is because it is. Clark's definition[1] of common ground implies that:

> a proposition *p* is only common ground if: all the people conversing know *p*; and they all know that they all know *p*.

This definition of common ground allows one to move between a view of language as an activity carried out by an ensemble of people (the social viewpoint) and a view of language as an activity carried out by individuals (the cognitive approach). The social viewpoint is developed by providing a detailed description of the activity by which the ensemble of conversants use and increase common ground. The cognitive viewpoint is developed by describing how an individual comes to know what is known by the others.

The nature of common ground is best explained by an example. This example will also illustrate how Clark's theory can help us understand the way that technology may affect the process of communication. Consider two people using a desk-top video conferencing package to discuss an architectural plan. They are wearing headphones with a boom microphone and can hear what each other says without difficulty. They can view changes in the other person's facial expression via a head-and shoulders view in a small video window. The remainder of the screen is taken up with a shared view of the architectural plan. Let us say that they have never met before. Even so they can make some assumptions about common ground. First there will be some common task defined by the work context. Let us say that Anne is an architect and Ben is someone who has hired Anne to design a house for him. The common task, negotiated in their previous correspondence is to agree what small changes need to be made to make the plan final. They also know they have the common ground that comes from living in the same town.

They can assume certain conventions with respect to the communication process. They will speak English. They will both try to use language that the other will understand and to monitor the conversation for potential misunderstandings. When they feel that they do not understand something sufficiently for their current purposes they will signal this to the other person.

From the video images they can make assumptions about their respective ages and genders that may have a bearing on how they express themselves. Also Anne will assume that Ben will not have the same detailed knowledge of

---

[1] Clark's formal definition of common ground is as follows:

> *p* is common ground for members of C if and only if:
> *i.* the members of C have information that *p* and that *i*.

This implies:

> everyone in C knows *p*,
> everyone in C knows everyone in C knows *p*,
> everyone in C knows everyone in C knows everyone in C knows *p*,
> and so on.

architectural terms that she has. As the conversation develops she modifies this opinion. Ben uses the term "architrave" correctly so she tries more technical (and hence more concise) terms in her utterances. These do not cause trouble in the conversation so she continues to use them. Later, however, Ben does not understand the term "lintel". Anne picks this up from his facial expression and explains it to him. During this explanation Ben demonstrates his understanding and they now both assume that this is common ground.

Ben describes how he would like the door of one bedroom moving, the one that faces south. The architectural drawing is larger than the screen and so this bedroom can only be seen by scrolling from the initial view. In their discussion of a previous detail Ben has scrolled to this view but Ann has not. He has no way of knowing this. Everyday experience leads him to assume the general principle that what he can see she can see also. This false assumption of common ground causes problems when he uses the phrase "up there on the left". After some time they realise they are talking at cross purposes and go about repairing their common ground.

At the end of the meeting they check their common ground regarding the original work objective and agree that the drawing can be sent to the builder. As this has legal implications, Anne suggests that she sends Ben a paper copy of the modified plan and Ben agrees to formally accept the plan in a letter. This change of communication medium permits re-reading so that each party can ensure that they really have achieved common ground.

The scenario sketched above illustrates the way common ground is used and how technology can effect the process of developing it. Table 2 summarises some of the common ground exemplified there under three categories: conversational conventions, communal common ground and personal common ground.

Conversational conventions

We will each try to be concise as possible but take account of the background of the other person

We will each make it clear to the other person when we cannot understand something sufficiently for our (individual) current purpose.

Communal common ground

We will speak in English

We are both professional people

We both live in the same town

Personal common ground acheived before the conversation

Our joint purpose is to sign off the plan

Personal common ground developed during the conversation

The door on the bedroom that faces south has to be moved

When we use the term "lintel" we mean the horizontal supporting beam above a door or window

We can both (now) see the bedroom that faces south on the plan

The plan can go to the builder

Table 2. Some of the common ground used and developed, see text for explanation.

Conversational Conventions are the assumptions Clark states we must make in order to converse at all. The two examples given here are not meant to be exhaustive or well defined, Clark takes a whole book to do this! Knowing what communities a person belongs to allows us to make certain assumptions about existing common ground. Communal Common Ground is common ground that can be assumed from our experience of these different communities. Personal Common Ground is the common ground personal to the particular conversants under consideration, that is the common ground assumed from our experience with the other individual.

By describing language use in this way we can begin to understand how the technology impinges on the conversation in the way that it does. If Ann had not been able to detect Ben's puzzlement because there was no video image of his face then Ben would have had to have signalled it in what he said. In some circumstances Ben might have been loath to do this and a serious conversational breakdown could have occurred. The false assumption of common ground made by Ben could have been avoided if scrolling on his machine automatically resulted in scrolling on Ann's (so called "linked scrolling"). We can also see why some media are better than others in certain circumstances.

This section has explained what common ground is as an introduction to Clark's theory. Clark's theory explains the process by which common ground is used

and developed in conversation. This, the main part of the theory, is outlined in section 4.

3. Scientific foundations

Questions concerning the interpretation of language are not new and have been explored by philosophers of language for centuries. In the late sixteen hundreds John Locke, for instance, attempted to conceptualise at an abstract level how simple and complex words are used and interpreted. But it is only relatively recently that social scientists have conducted empirical studies of language use. Technological developments such as audio and video recorders meant that talk as opposed to text could be documented and analysed at a level of detail not before possible.

In the late nineteen seventies sociologists such as Garfinkel, Sacks and Goffman turned their attention to the everyday and the taken for granted. As techniques such as discourse analysis developed it became possible to identify ethnomethods the taken-for-granted means of accomplishing interaction. In depth qualitative analyses uncovered previously overlooked phenomena such as turn taking, the process by which we signal that we are about to respond or we wish our interlocutor to respond.

The view of language use as simple information transfer corresponds to many people's common sense view of what is going on and so it has taken many years for this alternative notion of language use as a collaborative activity to gain popularity. As indicated above, the prime movers in this shift have been social scientists. Ethnomethodologists such as Goffman (1976), Sacks, Schegloff & Jefferson (1974) have been very influential, as have philosophers such as Grice (1957). As social scientists these authors take an approach that is at odds with the cognitive approach that is more commonly adopted by psychologists. For example, sociological accounts generally avoid attributing intentions to individuals, whereas intention is the basis of more cognitive accounts (Monk, 1998). What Clark has achieved is a marrying of these two approaches through his concept of a "joint action" (see below).

Readers with an interest in the building blocks of his approach can consult the following. McCarthy & Monk (1994) is a longer tutorial paper along the lines of section 1. Clark's book (1996) is a coherent statement of his whole theory that cites many references to the social science it is based on. There are also the original papers cited in these two sources.

4. Detailed description

Section 2 defined different kinds of common ground and informally described some of the mechanisms by which common ground is developed through an example. This section develops these ideas through some more formally defined concepts. The first part of the section sets out the fundamental assumptions made by Clarke. First he argues that face-to-face communication, rather than written language, should be the basis of a theory of language. He then points out, and

defines for his own purposes, some known properties of face-to-face communication, that it: involves more than just words; is a joint action; minimises effort, and develops common ground. The second part of this section outlines some concepts that build on these fundamentals. These are: the process of grounding, levels of collaborative activity, layers and tracks.

4.1 Fundamentals

Face-to-face conversation is "basic". Much work in linguistics starts from an analysis of well formed written text. Clark argues that real spoken conversations are a better starting point, even if they are messier. Children appear to learn how to do face-to-face communication spontaneously. Learning to read and write requires formal instruction. Indeed, a large part of the population of the world only has spoken language. If face-to-face speech the basis of all our language behaviour then our understanding of other ways of communication should build on our understanding of face-to face communication, not the other way around.

Face-to-face conversation involves more than just words. One of the major contributions of ethnomethodologists such as the Conversational Analysts (see for example Sacks, et al., 1974) has been to describe in detail how we use: hands, face, eyes and body in combination with the world we are in to facilitate the conversation. As well as the various cues used to manage turn taking these "instruments" can be used to signal meaning to someone else. Table 3 is adapted from Clark (1996) and lists examples of how we do this. Normally we think of language just as a process of describing things using words, i.e., the table cell in italics, but we sometimes describe things with our hands. We might describe the shape of something by making our hands into that shape. Pointing is another important signal in language use. Pointing saves a lot of words and can be done by voice (e.g., "that there"), with a finger or even with the eyes and face. Clark's final category of signal is demonstrating. We can demonstrate a gesture or tone of voice by imitating it. Clark suggests that a smile is best thought of as a signal to demonstrates one's happiness to someone else.

Instrument	Describing-as	Indicating	Demonstrating
Voice	*words, sentences*	"I", "here"	tone of voice
Hands, arms	emblems	pointing	iconic gestures
Face	facial emblems	pointing	smiles
Eyes	winking	eye gaze	widened eyes
Body	junctions	pointing	iconic gestures

Table 3. Methods of signalling. The voice is not the only instrument for communication in a face-to-face conversation. Adapted from Clark (1996, p.188.)

Face-to-face conversation is a joint action. As explained above, it does not make sense to think of language use except as a joint action involving two or more people. As such it presents the same problems as any other joint action such as playing a duet or shaking hands. In particular there is a need for "coordinating devices" such as conventions or jointly salient perceptual events that are part of

common ground. Clark uses this observation to explain many of the more detailed characteristics of language use described in the book. The key characteristics of a joint action are that both people involved intend to do their part and believe that the joint action includes their part and the other's. He uses a recursive definition of joint action[2].

Ensemble A-and-B is doing joint action *k* if and only if:
0. the action *k* includes 1. and 2.
1. A intends to be doing A's part of *k* and believes 0.
2. B. intends to be doing B's part of *k* and believes 0.

This definition, that applies to all joint actions including language, implies:
A believes *k* includes A's part plus B's part,
A intends to do A's part,
B believes A intends to do A's part,
A believes B believes A intends to do A's part,
and so on.

Face-to-face conversation uses common ground to minimise the effort required to communicate. As should be apparent by now, the key concept in Clark's theory is common ground.

> "Everything we do is rooted in information we have about our surroundings, activities, perceptions, emotions, plans, interests. Everything we do jointly with others is also rooted in this information, but only in that part we think they share with us." Clark (1996) p. 92.

As was pointed out in section 2, we make our assumptions about common ground on various bases. Some are to do with the groups we belong to. Very soon after meeting you, I will be able to make assumptions about the extent and detail of our common ground coming from our languages, nationalities, genders, ages and occupations. Other bases for making assumptions about common ground depend on our history together.

By making assumptions of common ground face-to-face conversation becomes extremely efficient. Even a grunt can communicate meaning in a context that is well understood by both conversants. This extreme efficiency is only possible because the joint action of language includes an intention to communicate efficiently. I must be able to assume that you are intending that I should understand what you are saying. Further, I must be able to assume that you are intending to do this in the most efficient way possible, otherwise ambiguities will arise. This notion of efficiency was reformulated by Clark and Brennan (1991) as a matter of minimising communication costs and then used to predict the effects of different ways of mediating communication (see Section 5.1).

Face-to-face conversation develops common ground. The effect of conversation is to test, reformulate and add to our common ground and so the most important source of common ground is our history of joint actions together.

---

[2] I am aware that some readers of this chapter may not find these quasi-mathematical formalisms as useful as I do. If you are such a reader you should be able to follow the argument from the text surrounding them alone.

One example of this personal common ground is the private lexicons of words that lovers develop together. Another more mundane example is the use of whiteboards or flip charts in meetings to form easily accessed references to previously established common ground. Thus someone can point at a somewhat cryptic heading on a whiteboard and in a single gesture refer to the common ground that may have taken several minutes to establish in the first place. So economical and effective is this form of common ground that people talking in the corridor have been known to construct imaginary "air whiteboards" that they can point to later in the conversation.

4.2 Grounding, levels, layers and tracks

The previous section presented the concepts that Clark's theory is based on. Before going on to describe how these concepts relate to studies of electronically mediated communication four further constructs need to be explained. They are the process of grounding, levels of joint action, layers and tracks.

Figure 2 depict the micro structure of the process that Clark describes as "grounding", i.e., the process of developing common ground.

> (a) Anne presents an utterance *u* for Ben to consider. Anne takes account of the common ground that already exists between them in order to present *u* in a form she believes Ben will understand. Ben attempts to infer the import of *u*, interpreting it as *u'*.
>
> (b) Ben provides some evidence *e* that, from his point of view, all is well with the conversation. This might be simply to continue with the next turn in a sensible way. Alternatively, Ben might rephrase the utterance and play it back to Anne. Anne interprets *e* as *e'*. On the basis of *e'* and the common ground they have already developed, Anne then has to make a judgement whether or not Ben has understood *u* "sufficient for current purposes".
>
> (c) Finally, Anne signals to Ben that she understands that he has an understanding sufficient for current purposes. Again, this is most commonly done by simply continuing with some relevant next utterance. If necessary words like "yeah" or "uhuh" can also serve this purpose. If she is not satisfied that *e'* meets the grounding criteria she can query *e* or re-present *u*.

This notion of a closely coupled grounding process is used in section 5.2 to explain the problems observed with a CSCW system.

- - - - - - - -

Figure 2 about here

- - - - - - - -

The process of grounding described above elaborates the sequence in which common ground is observed in the structure of face-to-face conversation. The notion of levels of shared action further elaborates the process by describing the joint actions that all have to be in place for this process to work.

Table 4 lists the four levels of shared action that Clark suggests are necessary for effective conversation. They can be thought of as an "action ladder" to be read from the bottom. So the first requisite is that A and B have joint action 1. Refer back to the definition of a joint action in section 4.1. Joint action 1 has two parts

one for A (behaving for B) and one for B (attending to A). The definition of a joint action implies that they are both intending to take these parts and believe that the other is doing likewise. Joint action 2 is for A to present signals to B and B to identify them. Joint action 2 depends on joint action 1 happening simultaneously. If B is not attending then she cannot identify the signal. Clark describes this as the principle of upward completion. Joint action 3 which depends on and happens simultaneously with joint actions 1 and 2 is where A signals some proposition and B recognises that A means that proposition. Finally, joint action 4 is where A proposes a joint project and B considers it.

	Speaker A's part	Addressee B's part
4	A is proposing a joint project *w* to B	B is considering A's proposal of *w*
3	A is signalling that *p* for B	B is recognising that *p* from A
2	A is presenting signal *s* to B	B is identifying signal *s* from A
1	A is executing behaviour *t* for B	B is attending to behaviour *t* from A

Table 4. The action ladder. Levels of simultaneous joint action needed to converse.

The example Clark uses to illustrate this is an occasion when he bought something in a drug store. Clark walks up to the counter where the assistant is busy checking stock. The assistant says "I'll be there". At level 1 Clark and the assistant have engaged on a joint action where the assistant says something and knows that Clark will listen. At level 2 they are similarly engaged in a joint action where the assistant utters the words "I'll", "be" and "there" knowing that Clark will identify them. At level 3 the assistant knows that Clark is engaged in recognising this signal as a proposition. Of course, what the assistant was really doing was the level 4 joint action of proposing a joint project. Clark's part in this joint proposal is to wait, the assisitant's part is to finish what he or she is doing. The notion of levels of joint action is used in section 5.3 to predict the effects of media on conversations where there is a "peripheral party".

The concept of tracks is a way of distinguishing between "the official business" of a conversation and talk about the communicative acts by which that business is conducted. When Al says "uh huh" in the conversation described in Table 1 he is not making a contribution to track 1, the business of discussing the repair of the car. He is instead contributing to track 2, talk about the communicative acts that achieve track 1. When Al says "uh huh" he is commenting on Roger's signal that the conversational repair had been successful.

The concept of layers is used to cope with the problem of pretence in fiction, irony teasing and so on. When I say "There were an Englishman, a Scotsman and an Irishman standing in a field" you know I am telling a joke. Layer 1 is to pretend layer 2, layer 2 is me proposing the proposition that there were an Englishman... The Clark's concepts of tracks and layers have not to my knowledge been used to discuss mediated technology. They are included here for completeness.

5. Case studies - applying the theory to the design of technology for communication

This chapter takes as case studies three published papers that have applied Clark's theory to the design of technology to mediate communication. The theory was developed to explain unmediated face-to-face conversation. As explained in section 4, Clark sees this as the logical starting point for a theory of any kind of language use, indeed his book's title is "Using Language". However, it is unreasonable to assume that the theory should be able to explain or predict the effects of mediating technology without further elaboration and each of these case studies has to extend the theory accordingly. Part of the interest in developing these examples is to examine how much has to be added to make the theory useful in design.

5.1 The costs of grounding (Clark & Brennan)

A basic principle in Clark's theory, explained in Section 4.1, is that conversants seek to minimise the effort required to communicate and that this is in a sense the purpose of developing common ground. Different communication media present different costs to different parts of the grounding process. For example, typing a text message will take more effort than speaking on the phone. However, reading complex instructions from the screen may be easier than having them read to you over the phone. Clark and Brennan (1991) elaborate the theory by analysing these costs as they apply to different communication media. The extended theory can then be used to explain some of the problems people have with media in particular contexts.

Clark and Brennan (1991) characterise the differences between different communication media in terms of which "constraints on grounding" they do and don't provide. In everyday life "constraints" may be thought to be bad, in this context they are good as they reduce ambiguity. Take the first constraint copresence. Say we are in the same room and I can see you are looking at a vase of flowers. I can use this common ground to construct a very efficient utterance - "dead eh?" to which I might get the expected reply "OK I'll get rid of them". Had we been conversing on the phone I would have had to construct quite a long utterance to engage you in the same shared project - "I don't suppose you could possibly chuck out the flowers in the vase on the hall table please?" The phrase "dead eh?" is too ambiguous without the constraints provided by copresence. You might prefer to think of constraints on grounding as "resources for grounding". Here we will stay with Clark and Brennan's terminology.

Clark and Brennan's (1991) complete list of constraints on grounding is given in Table 5. Equipment for mediated communication that provided all these constraints would be very good. All these constraints, can be viewed as an analysis of the findings from many studies of mediated communication in terms of Clark's theory. The first six of the constraints are advantages of face-to-face conversation that may be absent in mediated communication. These come from the theory in the sense that mechanisms identified by Clark will not be possible if these constraints are absent. For example, many of the methods of signalling enumerated in Table 3 will not be available without the constraints of copresence

and visibility. The tightly coupled process of grounding, described in section 4.1, will be difficult without audibility, contemporality, simultaneity and sequentiality. The last two constraints in Table 5 are advantages of written communication identified in studies comparing written and spoken electronic communication.

Copresence: A and B share the same physical environment. If I am in the same room as you, I can see and hear what you are doing, I know what you can see and hear, and what you are looking at.

Visibility: A and B are visible to one another. If we are video conferencing I can see you but will not have all the information I would have about you if we were copresent.

Audibility: A and B communicate by speaking. If we are on the phone I can hear you but will not have all the information I would have about you if we were copresent.

Contemporality: B receives at roughly the same time as A produces. On the phone you understand what I say at the same time or very soon after I speak. If we are communicating by voicemail this is not the case.

Simultaneity: A and B can send and receive simultaneously. Face-to-face I can nod or grunt to show I understand while you are speaking. Other devices may not allow this.

Sequentiality: A's and B's turns cannot get out of sequence. Misunderstandings often arise when emails are read in a different order to which they were sent. This is unlikely to be a problem on the phone.

Reviewability: B can re-view A's messages. Written material can be re-read and re-visited. Speech fades quickly.

Revisability: A can revise message for B. Emails can be read and revised before they are sent. Voice communications have to be repaired in subsequent turns or with extra words in the same turn if trouble is anticipated.

Table 5. Clark and Brennan's (1991) constraints for grounding.

In order to predict the problems users may have with a new communication medium one simply asks which of these constraints are present or absent. The consequence of some medium lacking one or more of the constraints is to increase the costs of some part of the grounding process. For example, if the conversation between the architect Ann and the homeowner Ben developed in section 2 had taken place without the video window Ben would have had to use words to indicate that he did not understand the word "lintel". This would have been more costly in terms of effort and possible loss of face than looking puzzled. Had they been communicating by writing in a chat window the cost in effort of signalling, detecting and repairing this trouble in the conversation is potentially even larger.

People evaluate costs in ways that depend on the purpose of the conversation. Two lawyers communicating about a case may choose the medium of typed

letters because it affords the constraints of revisability and reviewability. Here the cost of an inappropriate joint project being construed by either party is considerable and so the cost of losing all the other constraints is justified. Also they already have extensive common ground as they are both lawyers who have dealt with this kind of case before. They may choose to meet their clients face-to-face. This is because they need all the constraints they can muster to create some common ground. They know that their view of the case, as a technical problem that must be formulated within a particular legal framework, is quite different to the client's view of the case as a personal problem.

Clark and Brennan's approach has the potential to make detailed predictions about the costs and benefits for using different media for different purposes. However, it has yet to be fleshed out in sufficient detail to allow someone not immersed in the theory to make predictions using mechanical rules or heuristics.

5.2 Why Cognoter did not work (Tatar, Foster and Bobrow)

Cognoter was a software tool for use in electronic meeting rooms developed in the 1980s at Xerox PARC as part of the Colab project. The Colab electronic meeting room contained networked computers arranged so that a small group of people could have a meeting together. In a conventional meeting room people use a whiteboard to coordinate the work. Cognoter was to emulate and enhance the function of a whiteboard through the networked computers and a large screen central display. The obvious advantages of such a system is that material can be prepared in advance, displayed to the others, changed by the group and saved for future use. These are all things that are much less easy to do in a conventional meeting room. In addition, Cognoter was designed to facilitate brainstorming by allowing participants to work in parallel. Participants created "items" in an edit window. Items were then displayed to the others on an item organisation window as a short catch phrase or title. Anyone could move an item in the item organisation window or open it to read and edit the content.

The experiences of users of Cognoter were mixed and so Tatar, Foster & Bobrow (1991) recruited two groups from outside of the Colab research team to study in detail. Each group consisted of three long term collaborators who were asked to brainstorm about some subject of their own choosing that would be useful in their work. It was observed that neither group were able to use the item organisation window in the way intended. Also there were numerous conversational breakdowns where Cognoter got in the way of the work they were trying to do. Tatar, et al. (1991) conclude that the designers of Cognoter had used an inappropriate model of communication, corresponding to the information transfer model depicted in Figure 1. The idea of a Cognoter item as a parcel of information that is constructed and then transmitted to the others may be good for individual brainstorming but simply does not fit in with what happens in the rest of the meeting when discussing what to do with the ideas generated. If one views language use as a closely coupled process of collaborative activity, as depicted in Figure 2, a very different picture emerges. From this perspective Cognoter items have two functions: as elements in the conversation (signals), and as elements that may be conversed about (common ground). Cognoter did not support either function very well.

When someone is writing on a whiteboard other participants in the meeting know that they are doing so and can coordinate their actions accordingly. Creating an item with the item editor was a private activity making this difficult. Also with a conventional whiteboard the other participants can see the emerging text as it is written. This allows them to propose modifications and otherwise negotiate and signal common ground as described in Clark's process of grounding. With Cognoter the author of an item had no idea whether the others had read or even seen it. They could make no assumptions about its status in terms of the level of joint action it was involved in. In terms of Table 4 they could not make any assumptions about levels 1 and 2 in the action ladder. In terms of Clark and Brennan's (1991) analysis presented in section 5.1 Cognotor did not provide the normal grounding constraints expected from copresence, even though all the participants in the meeting were in the same room.

There was a further problem when people tried to refer to items on the item organisation window as the others were likely to be looking at a different version of the display. This was partly due to network delays (an absence of Clark and Brennan's contemporality constraint) but mainly because each display could be scrolled independently. A participant might have scrolled the item organisation window so the item another was referring to was not visible. To add to the confusion the central screen could be displaying a third view onto the item organisation window. As was indicated in section 4.1, pointing is a very effective conversational resource (see Table 3). Pointing may be done with a finger, by voice, or with your eyes and is known in this literature as deixis. Deixis broke down when the person making the reference was looking at a different version of the display to the version the others were looking at. This is another breakdown in the normal grounding constraints provided by copresence. Because of our experience of face-to-face conversation we expect that what we can see everyone else can see too and so it is quite difficult to repair these breakdowns.

Tatar, et al. (1991) suggest some modifications to Cognoter. The features they suggest are now commonly accepted as advantageous with this kind of system and have been implemented in commercial systems such as Timbuktu and Netmeeting. They are: (i) fast communication and update of displays; (ii) shared editing, where everyone can see the message being composed, letter-by-letter, backspaces and all; (iii) consistent positioning of windows and if I scroll so do you. Point (ii) comes under the more general design guideline of maximising "awareness", making everyone aware what everyone else is doing. Point (iii) is an example of the design guideline What I See is What You See (WISIWYS). These now widely accepted design principles are given a sound theoretical underpinning by Clark's theory and may even have been to some extent inspired by his ideas.

5.3 Predicting the peripherality of peripheral participants (Monk)

Watts & Monk (1999) studied doctors (General Practitioners, GPs) in their treatment rooms communicating over a videophone with medical specialists in a hospital. Figure 3 presents a schematic of this arrangement. The GP was usually in the presence of a patient. There might also be other legitimate overhearers. For example, in one consultation that they observed the patient was a young girl accompanied by her mother. The consultant was talking to the girl over the video

link and asked if she "ate well" to which she replied in the affirmative. The mother disagreed with this and was eventually able to break into the conversation and make this clear.

Watts & Monk (1998) characterise the legitimate overhearers, who are not currently actively involved in the work of the conversation as peripheral participants. The people currently actively involved in the work are described as primary participants. So, in the above case the primary participants were the consultant in the hospital and the girl in the treatment room. The mother and the GP were, at that time, peripheral participants. When the mother heard the child indicate that she was a good eater she felt the need to change her participatory status.

Another example of a legitimate overhearer might be a nurse. Two of the sites that were visited had a nurse who organised the video link and who would generally be present during the consultation. The same nurse might well be involved in treating the patient after the consultation. Having heard the discussion of treatment between GP, patient and consultant, as a peripheral party, this nurse was in a better position to explain the treatment to the patient. In Clark's terms the nurse had additional personal common ground due to overhearing.

At all the sites visited the camera was positioned to give a limited view of the person sitting directly in front of the video link, hence peripheral participants in the treatment room were unlikely to be visible to the consultant in the hospital. On the basis of Clark's theory Watts and Monk (1999) formed the hypothesis that if the specialist in the hospital could not see a peripheral participant it might make them more peripheral. It might be harder for them to change their participatory status and join the conversation. Also the primary participants might make fewer allowances for them, in their use of language, for example.

- - - - - - - - - - - -

Figure 3 about here

- - - - - - - - - - - -

The challenge for Clark's theory then is to predict how a particular audio-video configuration could affect how peripheral a peripheral participant will be. Monk (1999) extends Clark's levels of joint action (Table 4) to do this. The starting point is a Participant Percept Matrix (PPM) (Watts & Monk, 1998). This shows who can see and hear what. Table 6 is a PPM for the situation described above. The GP, patient and nurse are co-present so they can all see and hear one another. However, because audio is via telephone handsets and the image is of limited scope, not all the percepts are available to all the participants.

Percepts:	*Participants:* Specialist	GP	Patient	Nurse
Specialist's face	-	yes	yes	yes
GP's face	yes	-	yes	yes
Patient's face	no	yes	-	yes
Nurse' face	no	yes	yes	-
Specialist's voice	-	yes	no	no
GP's voice	yes	-	yes	yes
Patient's voice	no	yes	-	yes
Nurse' voice	no	yes	yes	-

Table 6. Participant Percept Matrix for one instance of telemedical consultation. The specialist and GP are communicating with telephone handsets and the camera provides the specialist with a limited scope image that only shows the head and shoulders of the GP.

	Speaker A's part	Side participant C's part
4	No joint action	No joint action
3	A is signalling that *p* for B and C	C is recognising that *p* from A
2	A is presenting signal *s* to B and C	C is identifying signal *s* from A
1	A is executing behaviour *t* for B and C	C is attending to behaviour *t* from A

Table 7. Levels of joint action with a close peripheral participant (a side participant).

Table 7 extends Clark's theory as represented in Table 4 for a two person conversation to the case of a three person conversation where C is a close peripheral participant, i.e., someone who is really a part of the conversation but is not the addressee. See Monk (1999) for a full explanation of the term "side participant".

Table 8 then lists the evidence that might lead A and C to assume that the other is taking part in each level of joint action. There is no joint action at level 4 because C is only a side participant. However, C may feel able to assume they are part of lower level joint actions. Some of this evidence comes from being able to hear the other person ("H" in Table 8) some from being able to see them ("S" in Table 8). When using this table one should also recognise Clark's principle of downward evidence in the action ladder. A level 3 joint action is only possible if the corresponding level 1 and 2 joint actions are too. This means that evidence that the other person is joining you in a level 3 joint action is also evidence that they are joining you in the corresponding level 1 and 2 joint actions.

	Evidence leading speaker A to consider side participant C	Evidence leading side participant C to consider speaker A
4	No joint action	No joint action
3	C has responded appropriately to previous signals (H); A can hear verbal back channels from C (H); A can see visual back channels from C (S)	A's signal is directed at B and C (H); A's signal refers to common ground specific to C (H)
2	Only by downward evidence	Only by downward evidence
1	A can see C is attending (S)	C can see A's behaviour is directed at B and C (S)

Table 8. Evidence that the other person is taking part in the joint action, speaker and addressee. (H) = must be able to hear other; (S) = must be able to see other.

Table 8 can be used to determine what evidence is available to a primary participant, say the specialist, that would lead them to consider a peripheral participant, say the nurse, to be a side participant, and vice versa. Combining this with an analysis of the evidence available to the other primary participant, the GP, allows an assessment of the overall peripherality of the nurse, i.e., how easy it will be for them to join in the conversation.

The above account shows how Clark's model can be elaborated to make predictions about the effects of small changes to the way a video link is configured. Monk and Watts (2000) present a laboratory experiment where such predictions are made and tested with encouraging results. However, much more data is needed before we can say with confidence that the model has real predictive power.

6. Current status

The three case studies presented above demonstrate that elaborated versions of Clark's theory is capable of make useful predictions in the area of electronically mediated communication. As with any theory, the question then becomes "how realistic is it to apply the theory in a real design context?" In an ideal world a theory should be encapsulated as a set of guidelines or rules that could be used by a designer with very little background in human factors of human communication. Failing this, the theory should be formalised as principles that could be used by a human factors consultant who has had the time to get to understand the theory and the background material needed. At the earliest stages, which is where we are now with Clark's theory, the theory is only really usable by researchers with a specialist knowledge of the area.

The reason for this can be seen in the case studies. In two of the three the theory has to be added to before it can make predictions. Clark & Brennan (1991) had to add the concept of a grounding constraint to complete their analysis. Monk (1999) had to generalise Clark's levels of joint action to three participant conversations and specify the evidence leading a participant to consider they are being joined in a joint action. Only Tatar et al. were able to use the framework with little modification. As more researchers use the theory to reason about electronically mediated communication, the bounds of the theory and the additional assumptions needed will become more apparent. It is to be hoped that it will then be possible to set out principles that could be used by our hypothetical consultant. The next phase of development will be to gain sufficient practical experience of using the theory in real design contexts to make the next shift to a set of well specified guidelines for use in particular contexts. Guidelines for configuring multiparty video conferencing, guidelines for desk-top video, guidelines for asynchronous communication, and so on. We are a long way from this ideal state at the moment. I very much hope that this chapter will serve as a stimulus to taking some steps in this direction.

7. Further reading

Readers interested in the background material (scientific foundations) which Clark's theory draws on should read the tutorial review paper by McCarthy & Monk (1994).

Clark's book (1996) is an accessible and coherent statement of his whole theory. It has useful orienting summaries at the beginning and end of each chapter. Also the first and last chapters provide accessible summaries of the whole book. He also goes to some lengths to explain the scientific foundations of his work. Before this book it was hard to find a coherent statement of Clark's framework that could be described as a theory; the concepts were distributed in a number of papers. For this reason the work had not had a large impact. Readers interested in this theory are strongly recommended to buy the book and get it straight from the horse's mouth.

Other readers may wish to find out more about the research literature on electronically mediated communication. Finn, Sellen & Wilbur (1997) is a comprehensive set of papers (25 chapters, 570 pages) on video-mediated communication. The CHI, CSCW and ECSCW conferences are also good sources of papers (see for example, Kraut, Miller & Siegel, 1996; McCarthy, Miles & Monk, 1991; Neuwirth, Chandhok, Charney, Wojahn & Kim, 1994; Tang, 1991; Veinott, Olson, Olson & Fu, 1999; Vertegaal, 1999).

Of the case studies, Clark & Brennan (1991) is very accessible. Tatar, Foster & Bobrow (1991) is rather lengthy while Monk and Watts' work on peripheral participation is distributed amongst several papers. Watts & Monk (1997) is a good 2-page starting point.

8. References

Altmann, G. T. M. (1997). *The ascent of babel.* Oxford: Oxford University Press.

Clark, H. H. (1996). *Using Language.* Cambridge: CUP.

Clark, H. H. & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. Levine & S. D. Teasley (Eds.), *Perspectives on Socially Shared Cognition.* (pp. 127 - 149). Washington, DC: American Psychological Association.

Finn, K. E., Sellen, A. J. & Wilbur, S. B. (1997). *Video-mediated communication.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Goffman, E. (1976). Replies and responses. *Language in Society, 5*, 257 - 313.

Grice, H. P. (1957). Meaning. *Philosophical Review, 66*, 377-388.

Jefferson, G. (1987). On exposed and enclosed corrections in conversation. In G. Button & J. R. E. Lee (Eds.), *Talk and social organisation.* Clevedon: Multilingual Matters.

Kraut, R. E., Miller, M. D. & Siegel, J. (1996). Collaboration in performance of physical tasks: effects on outcomes and communication. In M. S. Ackerman (Eds.), *CSCW96*, Boston, MA. (pp. 57-66). City: ACM.

McCarthy, J. C., Miles, V. C. & Monk, A. F. (1991). An experimental study of common ground in text-based communication. In *The ACM CHI'91 Conference on Human Factors in Computing Systems*, (pp. 209 - 214). City: ACM Press.

McCarthy, J. C. & Monk, A. F. (1994). Channels, conversation, cooperation and relevance: All you wanted to know about communication but were afraid to ask. *Collaborative Computing, 1*, 35 - 60.

Monk, A. F. (1998). Cyclic interaction: a unitary approach to intention, action and the environment. *Cognition, 68*, 95-110.

Monk, A. F. (1999). Participatory status in electronically mediated collaborative work. In *Proceedings of the American Association for Artificial Intelligence Fall Symposium "Psychological models of communication in collaborative systems"*, North Falmouth, MA. 73-80 . City: Menlo Park, CA: AAAI Press.

Monk, A. F. & Watts, L. A. (2000). Peripheral participation in video-mediated communication. *International Journal of Human-computer Studies, 52*, 775-960.

Neuwirth, C. M., Chandhok, R., Charney, D., Wojahn, P. & Kim, L. (1994). Distributed Collaborative Writing: A Comparison of Spoken and Written Modalities for Reviewing and Revising Documents. In *Proceedings of ACM CHI'94 Conference on Human Factors in Computing Systems.* (pp. 202).

Sacks, H., Schegloff, E. A. & Jefferson, G. (1974). A simplest systematics for the organisation of turn taking in conversation. *Language, 50*, 696 - 735.

Shannon, C. E. & Weaver, N. (1949). *The mathematical theory of communication.* Urbana: University of Illinois Press.

Tang, J. C. (1991). Findings From Observational Studies of Collaborative Work. *International Journal of Man-Machine Studies, 34*, 143-160.

Tatar, D. G., Foster, G. & Bobrow, D. G. (1991). Designing for conversation: lessons from Cognoter. *International Journal of Man-machine Studies, 34*, 185-209.

Veinott, E. S., Olson, J., Olson, G. M. & Fu, X. (1999). Video helps remote work: speakers who need to negotiate common ground benefit from seeing each other. In *CHI99*, Pittsburgh. (pp. 302-9). City: ACM Press.

Vertegaal, R. (1999). The GAZE groupware system: mediating joint attention in multiparty communiation and collabroation. In M. G. Williams, M. W. Altom, K. Ehrlich & W. Newman (Eds.), *CHI'99*, Pittsburgh, PA. (pp. 294-301). City: ACM Press.

Watts, L. A. & Monk, A. F. (1997). Telemedical consultation: task characteristics. In S. Pemberton (Eds.), *CHI'97 Conference on Human Factors in Computing Systems*, Atlanta, Georgia. USA. Proceedings (pp. 534 - 535). City: ACM Press.

Watts, L. A. & Monk, A. F. (1998). Reasoning about tasks, activity and technology to support collaboration. *Ergonomics, 41*, 1583-1606.

Watts, L. A. & Monk, A. F. (1999). Telemedicine: what happens in teleconsultation. *International Journal of Technology Assessment in Health Care, 15*, 220-235.

Acknowledgements: The author was supported by the UK EPRSC (Grant GR/M86446) and the PACCIT programme (Grant L328253006) during the period of this work. I would also like to acknowledge the help of members of York Usability Research, students, and other authors in this volume for their comments on previous drafts of this chapter.

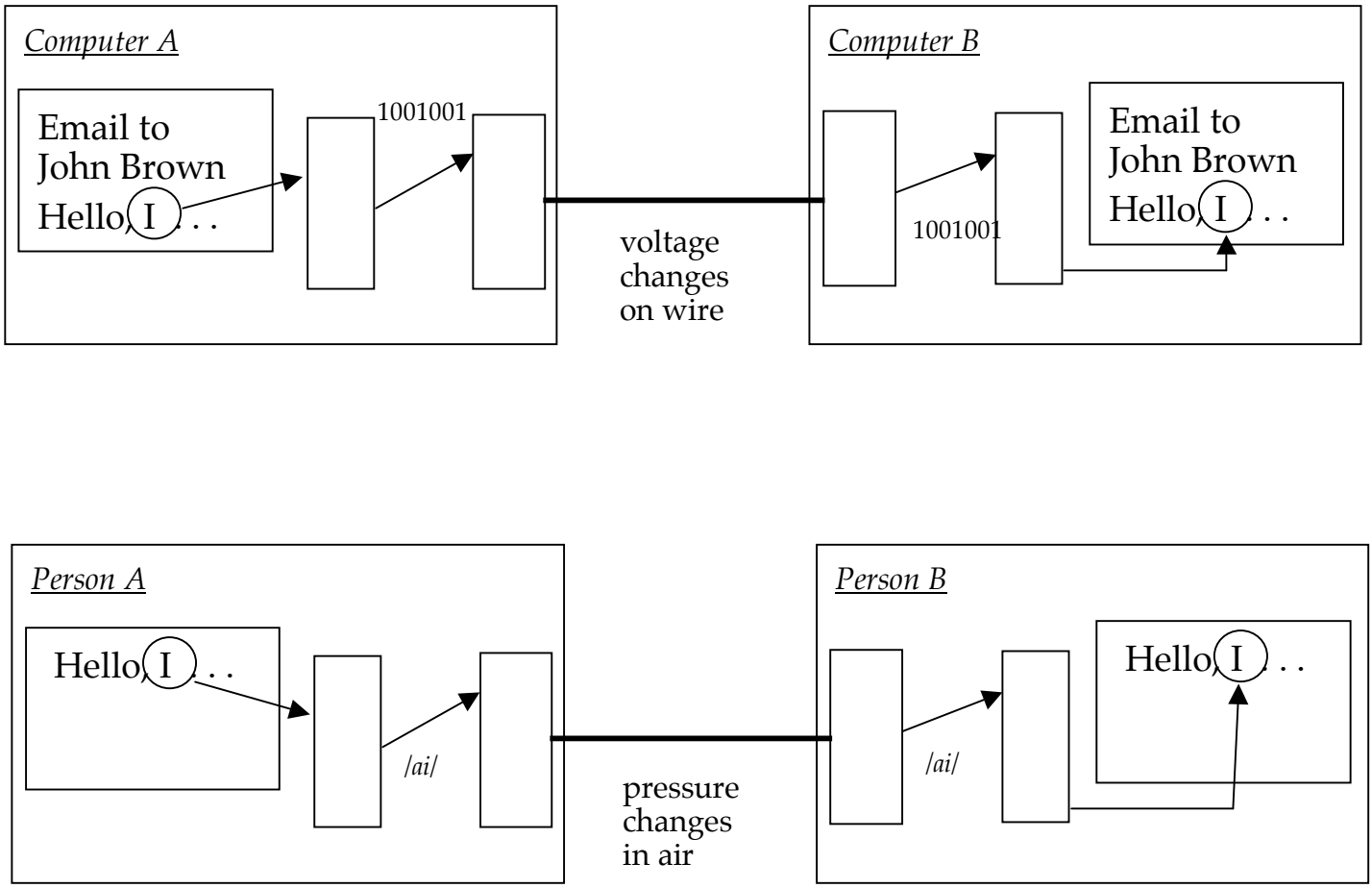

Figure 1. The information transfer model of communication, top panel as applied to communicating computers, bottom panel as the encoding-decoding model of human-human communication.

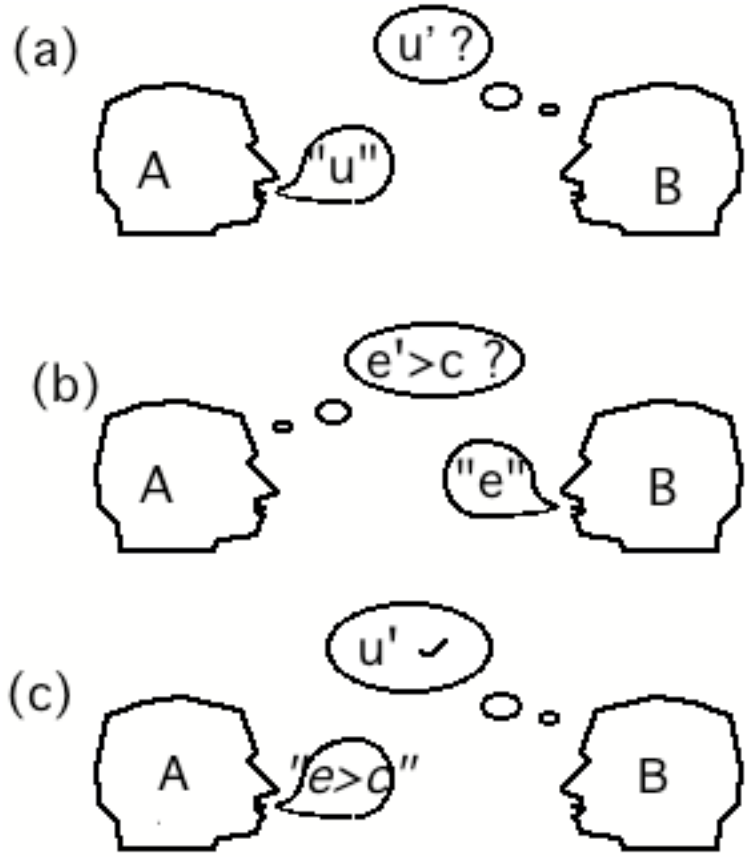

Figure 2. Clark's grounding process. u = utterance, u' = understanding of utterance, e = evidence of understanding sufficient for current purposes, e' = understanding of evidence of ..., c = grounding criterion.

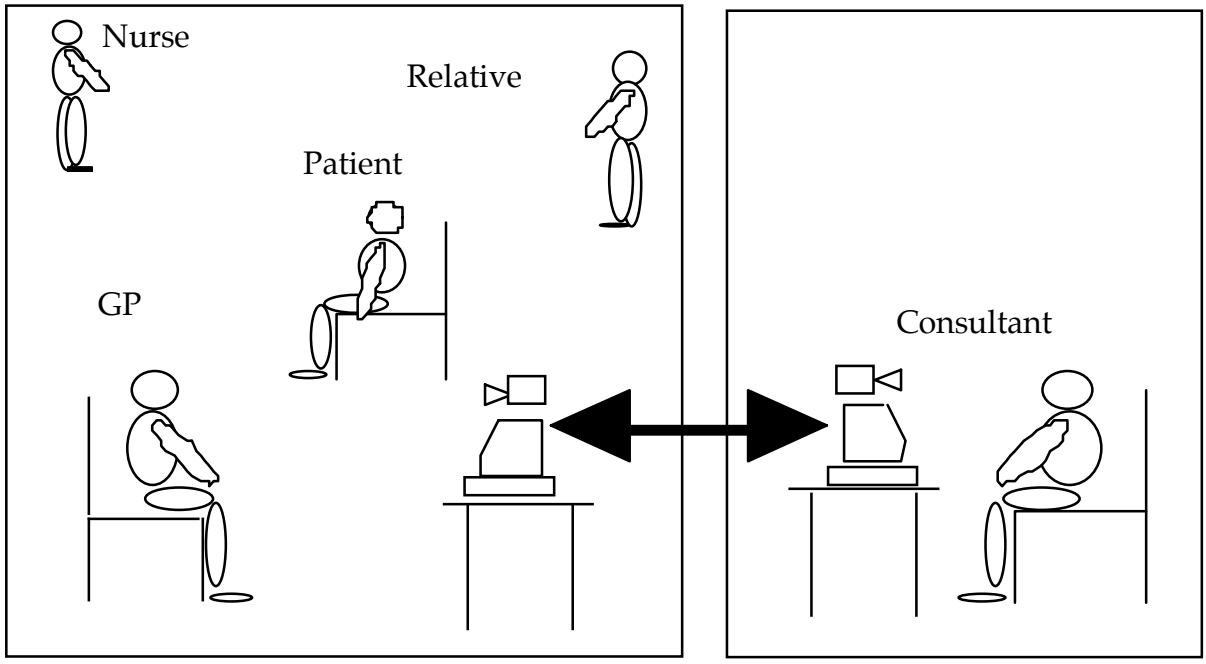

Figure 3. Schematic of the video conferencing context studied by Watts and Monk (1999).