13. Multimodal Communication

Jens Allwood and Elisabeth Ahlsén

1. Why is the topic of multimodal communication interesting?

Face-to-face communication is the evolutionary primary and most common way of communicating for humans. It is not unlikely that vocal-verbal language with more or less intentional control, but still to some extent systematic, developed from being action- and gesture-based to being vocally based (Arbib 2005). In contrast to the hundreds of thousands of years of natural development of face-to-face communication, communication technologies have been artificially developed in order to supplement, enhance or replace certain types of face-to-face communication and today we have a situation of steadily increasing use of communication technology with possibilities for using multiple modalities. In this chapter, we will give an account of human-to-human multimodal communication, in order to have a background for a discussion of the use of multimodality in new communication technology (see also Lücking and Pfeifer as well as Gibbon both in this volume). We will primarily deal with face-to-face communication, rather than multimedia in relation to written communication. First, we turn to a brief overview of human-human communication in order to see what features can potentially be used in human-computer interaction in general and more specifically in dialog and tutoring systems. Then, we will discuss some problems in making computer based systems multimodal and thereby more human like.

2. Human-human communication

2.1. What is multimodal communication?

The word "modal" has the fairly abstract meaning "pertaining to manner or mode" (e.g., Collins English Dictionary 2009), so it is not surprising that the term "multimodal communication" has been used in many different ways. For example, it has been used for the three phenomena that might more appropriately be called "multi-medial communication", "multi-representational communication" and "multi-mass-medial communication" (cf. Allwood 2008). Here we will take "multimodal communication" to mean communication involving more than two of the sensory modalities (sight, hearing, touch, smell and taste). For practical purposes, we will also, besides "sensory or perception modalities", talk about "production modalities", by which we mean human bodily

means that normally produce information for the different sensory modalities, that is, gestures, speech organs etc. Thus, gestures can produce information for the visual modality and the speech organs can produce speech sounds for the auditory modality. By "multimodal information" we mean information for several sensory modalities. Below, we will now consider some of the relevant features of multimodal communication under the four headings production, reception, interaction and context.

2.2. Production

In face-to-face human-human communication, a number of production modalities are used in interaction. From a linguistic-communicative point of view, they can be divided into:

(i) *Communicative body movements* (other than movements of speech organs), for example, gestures, facial expressions, body posture, gaze direction etc.,

(ii) *Aspects of the system of a language, which to some extent are common to speech and writing*, such as *phonemes* (to some extent corresponding to *graphemes*) *morphemes, lexemes (words)* and *phrases*, which make up the store and basis for the structure of meaningful units to be used in the language, *syntax* (word order), *semantics* and *pragmatics* (the meaning and use of language).

(iii) *Aspects of the system of a language*, such as *prosody* (variations in pitch, intensity and duration carrying information concerning word identification, emphatic stress and information structure, attitudes and emotions), *which are usually not directly captured in written language*. This, of course, is not to deny that information structure, attitudes and emotions in writing can be expressed by other means, such as bold face, underscoring, capitals, smileys or by non-prosodic means common to speech and writing, such as word order or rhetorical strategies.

The different communicative expressions that are produced can have different status with regard to what type of semiotic or representational sign relation they (or their various features) are based on (cf. Peirce 1931).

An *index* is a sign based on contiguity (closeness in time and space and by extension causality) to what it refers to. In this sense, clouds are an index of rain and a pointed index finger is an index of what it points to. An *icon* is, in addition to an indexical relation, based on a similarity relation, for example, photos and diagrams are icons, while a *symbol*, in addition to an indexical substratum, has an arbitrary (conventional only) relation to what it refers to. Most of the verbal vocal or written communication we engage in is mainly symbolic. But icons are, for example, used in the picture-like signs found in many interfaces to computers and cell phones. Indexical communication can, for example, be seen in arrows for pointing or so

called deictic expressions that depend on their context, like *I, you, he, here, there and now.*

Communicative expressions can also have different status with regard to the degree of intentionality and awareness involved in their production: *indicate*, *display* or *signal* (cf. Allwood 1976). When we *indicate* something, the signs we are producing are not intentional, but are still informative and convey something; blushing, for example, can convey nervousness. When we *display* something, we show it intentionally. A smile, for example, can display friendliness. A pointing gesture is a display of what it points to. Thus, indices (in the Peircean sense) do not need to be indicated, they can also by displayed, that is, a smile can be an indicated unintentional and perhaps unaware expression of friendliness, but it can also be a displayed, more intentional expression of the same emotion. An interesting question for multimodal communication research is to try to find out whether people react differently to indicated and displayed smiles and whether observable, registrable differences exist between them and could be picked up not only by humans but also by automatic sensors.

Most of what has traditionally been studied by linguists has, however, been focused on expressions that are *signaled* (i.e., shown to be shown). Typically this can be achieved by using spoken or written words, phrases and sentences or other types of communication with the help of symbols (i.e., signs with an arbitrary relation to their content – cf. above). The three types of representational relation (index, icon and symbol) and the degrees of intentionality and awareness (indicate, display and signal) provide us with two different ways of distinguishing what has been called "natural meaning" from "non-natural meaning" (cf. Grice 1957).

1. The first way suggests that natural meaning is provided by expressions (signs) that have a motivated non-arbitrary representational relation that is primarily motivated by contiguity or similarity (indices and icons). Non-natural meaning is provided by expressions (signs) that are not primarily motivated by contiguity or similarity, but by arbitrary convention (symbols).
2. The second way instead suggests that natural meaning is provided by expressions that are not exposed intentionally (indicative behavior). Non-natural meaning is manipulable, that is, produced by expressions that are expressed with some intentionality (displayed and signaled behavior). The two ways of drawing the distinction are not equivalent, even if there is an overlap. Indicated indices are natural and signaled symbols are non-natural in both variants of the distinction, while the remaining categories shift their belonging depending on the criteria the distinction is based on.

In face-to-face communication, the three types of representational relation and the three different degrees of intentionality and awareness often occur in a mixed or simultaneous fashion, which makes it possible, at the same time, to

communicate factual main messages, attitudes and emotions as well as to manage your own communication and the interaction. In this way, a great deal of the indicated and displayed communication, as well as some of the signaled communication, is mainly conveyed by other modalities than spoken words, phrases and sentences (Allwood 1976, 2008). We can, for example, smile, nod and speak, using both words and prosody at the same time.

The most common combinations of the three types of basic semiotic relations (indexical, iconic and symbolic) with the three degrees of intentionality and awareness (indicate, display and signal) are illustrated in Table 1 below. In principle, all the combinations are possible; the table only shows the most common combinations. Since all types of signs are based on causal relations, they are in this sense indexical.

Some examples of the other combinations are the following. If a smile is an unintentional automatic reaction, it is an indicated index. If it is expressed intentionally to show a type of attitude, it is a displayed index. If it is expressed with the intention that it be recognized as a display, it is a signaled index. This could occur, for example, when a friendly smile is a sign to an accomplice to initiate some action.

A sign is an indicative icon, when a likeness is produced accidentally: for example, I look sad because a grain of sand blew into my eye and made my eyes water. A sign is a displayed icon when the likeness is produced intentionally (e.g., a painted picture). A sign is a signaled icon when the intention to produce a likeness is intended to be recognized. This occurs very often with gestures when they are used to illustrate what is being said.

A sign can be an indicated symbol if it is used unintentionally, for example, spoken words in language X while sleeping might indicate a relation between language X and the sleep talker, for example, that the sleep-talker knows language X. This indicated connection can also be displayed, if I use a language where I have learned an utterance or two by heart, without understanding the meaning, in order to show my connection with the language (see Searle 1970).

Finally, a symbol may be signaled, which is its most common use, when we use words with the intention that it should be recognized that we are communicating something.

Table 1. Basic semiotic relations and degrees of communicative awareness and intentionality.

	Index	Icon	Symbol
Indicate	X		
Display		X	
Signal			X

If we compare *Human-Human Communication* (HHC) with *Human-Machine Communication* (HMC), it could perhaps be claimed that machines do not have awareness and intentions, they can only indicate, and that for this reason only humans exhibit the full range of variation in awareness and intentionality shown in Table 1. Even if this is correct, machines can, however still exhibit variation when it comes to employment of the basic semiotic relations of representation, that is, they can communicate with their users using indices (e.g., activated flashing lights), icons (e.g., pictures) and symbols (e.g., words) or combinations of all three relations.

2.3. Reception

Reception, just like production, includes processes on different levels of control (intentionality) and awareness. One reason for this is that human-human communication probably contains an element of what is often called "mirroring", that is, a kind of automatic reproduction or activation of the communicative actions of the interlocutor, either internally or externally visible (Ahlsén 2008; Le Bel et al. 2009). This means that all the aspects of production described above can be relevant for reception. This, in turn, means that reception can take place on different levels of conscious control or intentionality. What is indicated (and sometimes what is displayed), by body movements and the words and sentences produced, is often received and reacted to more or less *subconsciously*. This includes many aspects of emotions and attitudes as well as of communication management.

More generally, reception includes several degrees of processing from *subconscious reactions* to *perception* characterized by more conscious discrimination and identification, that is, hearing or seeing that certain signs are produced.

The receptive processes can also lead to *understanding* in which incoming information is *interpreted* in relation to the recipient's own stored background, for example, the communicative context and the activity context. One example of how a message is understood, interpreted and reacted to attitudinally is the recipient's *impression of the trustworthiness of* the person producing the message. This impression often has an influence on the effect the understood content has on the recipient (Komiak et al. 2004; Ruttkay and Pelachaud 2004).

In face-to-face communication *words, gestures and prosody* are all interpreted in interaction between two or more people and it is, thus, the multimodal totality of a contribution that is normally received, interpreted and reacted to. Let us therefore briefly consider this interaction.

2.4. Interaction

An important aspect of human-human face-to-face communication (as well as human-computer communication) is how the interaction between communicators is managed. The study of *Interactive Communication Management (ICM)* is the study of different systematic means for this. Often, interactive communication management makes use of different modalities that help to manage the progression of successful interaction.

The *turn management* system helps interlocutors manage the distribution of their contributions to the interaction. When a participant has the turn, he or she has the right to contribute by speech or by some other modality, that is, has "the floor". Participants manage turn distribution, for example, by showing (i.e., indicating, displaying or signaling) when they want to speak, when they accept an invitation to speak, when they want to continue, when they want to stop and often to whom they want to give the next turn (Sacks, Schegloff and Jefferson 1974). Turn distribution can be signaled, for example, in a formal meeting, where a chairman distributes the turns explicitly by means of names or gestures or by questions directed to a specific, named person (*What do you think, Bill*?). Turn management, however can also be less consciously achieved by changes in body movement, gesture, voice quality or facial expression.

The *feedback* giving system (Allwood, Nivre and Ahlsén 1992) helps interlocutors to express ability and willingness with regard to contact, perception, understanding and attitudinal reactions in relation to what is being communicated. This is done continuously, usually mainly by small, unobtrusive contributions, such as head nods, head shakes, smiles, and words like, *yes*, *no* and *mm*. These reactions from the listener guide the speaker concerning whether he or she can go on (continued contact), whether the communication is perceived, understood and concerning the attitudinal and emotional reactions of the interlocutor. When there is a need, feedback can also be elicited by special communicative means, such as use of tag questions or question intonation.

A third mechanism for interactive communication management is *sequencing*. Contributions often occur in fairly set sequences, Such sequences extend from "exchange types", sometimes also known as "adjacency pairs" (e.g., Levinson 1983), such as "question-answer" or "greeting-greeting" and preference organization (e.g., Pomerantz 1984), where a certain type of contribution activates a preference for a particular response among a certain selection of possible responses, for example, after an offer it is often preferred to express gratitude, to "scripts" identifying what is typically said in a specific type of activity, like a restaurant or a travel agency. This structure helps structure how the participants can inter- act (Shank and Abelson 1977).

2.5. Context

A relevant issue for how and when multimodal communication strategies can be used in human-human as well as in human-computer interaction is the *activity dependency* of the communication. Some aspects of human-human communication are relatively stable across different social activities, whereas other aspects show considerable variation and activity dependence (cf. Allwood 1995, 2000). This activity dependence also applies to multimodality.

In *most ordinary face-to-face communication* (also in distance video communication), multimodality is present in an integrated way. This applies to communication in everyday social activities, work, leisure activities and games as well as to communication through avatars in virtual reality environments etc.

Different communication functions can be more easily handled by one modality rather than by another. Depending on context, gestures, map drawing or diagrams, for example, can be used for giving explanations or instructions concerning how to find the way (e.g. Kopp et al. 2007; Tversky et al. 2008) and spoken words or other sounds can be used for alerting someone quickly.

Different social activities provide different contextual options for multimodal communication and in ordinary human-human communication, the participants more or less automatically adapt to these options. They communicate with actions, objects, tools, gestures, speech etc., according to what is possible and efficient (cf. Allwood 1995, 2000) in a particular activity context.

In communicative contexts where communication to a great extent is handled by *other modalities than spoken or written words*, multimodality is essential, for example, when movements need to be visualized by gesturing.

In a number of contexts, where multimodality is necessary, *flexibility in the choice of modality is needed*. This includes contexts where specific modalities are temporarily or permanently blocked. Examples of this are voice-only communication in phone calls, where the visual modality cannot be used or when sensory modalities are blocked as when one or more participants are blind. In communication with deaf persons, the auditory modality is blocked, so communication needs to be visual (see Kubina and Abramov in this volume). When a person's hands are occupied by other actions, communication needs to employ the auditory channel. In contexts such as these, one modality needs to compensate for loss of communication in another modality. Many of the contexts where communication needs to be multimodal are, thus, *communicatively challenging*, for example, because the participants do not share the same language and/or cultural background or because one of the participants has a communication disorder. Flexibility in choice of modality and the possibility of *redundancy* by using several modalities in interaction is often needed. Also when human-human communication takes place in *complex communication environments*, for example, in technically advanced activities, such

as controlling airplanes, nuclear power plants etc., enhanced efficiency makes multimodality involving the *distribution of communication over several modalities* necessary (Goodwin and Goodwin 1997).

3. Which features from multimodal human-human communication (HHC) are relevant for human-machine communication (HMC)?

One answer to this question is that all or most features of multimodal human-human communication are or will over time become increasingly relevant for human-machine communication. However, not all of the multimodal HHC features are relevant for all HMC systems. In the past, the dominating view has been that HMC is so different from HHC that perhaps it works more efficiently if it does not try to model all of the complexity of HHC (cf. Becker 2006). But this view is changing. Even though the complexity of HHC is considerable and not completely understood, described and explained, the rapid technical development enabling more complexity and multimodality, and the demands relating to usability and accessibility for all, together with the increasing dependency on HMC for all kinds of everyday communication activities drive development in the direction of trying to make HMC more HHC-like in general and in this, to make use of the available possibilities for multimodal communication strategies.

A more cautious answer than the one just given is that, although all or most features of HHC are relevant, some of them are probably more important than others. Above all, speech is better understood than other modes of production (cf. Partan and Marler 2005) and might for this reason come into more general use.

A further aspect is that the quality of the features of multimodal communication implemented in HMC systems can vary considerably and that there is often a threshold effect, so that a certain quality has to be fairly fully achieved in order to make a particular communication strategy more useful than disturbing. For some features, however, rough estimations can be sufficient for some communicative purposes, whereas other features need to be fine tuned in order to be useful. A friendly smile can work for many purposes, while a rough estimation of synchronization of lip movements and speech sounds can be more disturbing than helpful.

A complication is that in order to achieve a reasonable degree of "naturalness" in multimodality, the coordination and integration of different features has to be good and this is still not easy to achieve. (cf., however, Oberzaucher et al. 2008; Boukricha et al. 2009). The goals of a "natural" implementation are, for example, to make a virtual *Embodied Communicative Agent* (ECA; see also Kopp and Wachsmuth as well as Martin and Schultz both in this volume) have movements that are smooth and coordinated and do not violate any of the constraints on movement in humans, so that, for example, lip movements and

speech sounds should be perfectly synchronized. Facial expressions should also be fine tuned and have many variations, which then must be generated in coordination with the spoken message, since small deviations can be very distracting for the user. Some of the more abstract goals that have been expressed for the design of ECAs are to achieve "addressability", trust, "personality", "believability", "naturalness" and "flow of communication". All of these goals have to be seen in a long term perspective, since, in many cases, it is not clear how they are achieved in human-human, face-to-face communication and even less how they can be achieved in human-machine communication.

Let us now consider some of the features of multimodal Human-Human Communication that are relevant for Human-Machine Communication and therefore are important to consider in further research and development.

3.1. Activity dependency

Since human-human communication shows rapid adaptation to differences between social activities, this ability is also in a long-term perspective important to achieve in human-computer communication. If the ECA is supposed to be involved in more than one activity type, it should be able to flexibly adapt to a new activity type, including the multimodal strategies that would be expected and benefi- cent in HHC. If the ECA is only intended for one activity type, it could, of course be designed or adapted for only this activity. Achieving this kind of flexibility in HMC involves analyzing the relevant influencing activity factors (e.g., purpose, roles, instruments and environment – cf. Allwood 2000, 2007) and implement- ing the relevant type of multimodal behavior that they are connected with.

3.2. Cultural variation

Besides differences between activities, differences between cultures are often also relevant. Most societies of today are multicultural and agents on the Internet get involved in communication with humans who have widely different cultural backgrounds. So, like with activities, if the ECA is supposed to be involved in more than one culture, it should be able to flexibly adapt its communication to the culture at hand, including the relevant multimodal strategies. For example, Jan and Traum (2007) and Allwood and Ahlsén (2009) present models with parameters that can be varied for different cultures for types of conversa- tional behavior, such as proxemics, gaze and overlap in turn taking. These parameters can be set for a specific culture. If only one culture is involved, the ECA should be adapted to the more specific and local purposes that are relevant in this culture.

3.3. Flexibility and adaptability in choice of modality

Accessibility for all or most people to the services provided through an ECA also involves the need to adapt the system to the needs of users who can perhaps not make use of all modalities (e.g., blind users, deaf users). It is, thus, necessary to have the possibility to present information in a way that compensates for this. Like with the adaptation to activity and culture discussed above, flexibility and the possibility to choose and set different parameters for multi- modal strategies are required. Although strict "translation" between different sensory modalities is not often possible, the possibility to communicate with a system via speech, writing, or sign language means a substantial increase in accessibility. Attempts at rendering the content of pictures in web pages via text, which is presented as speech, are today obligatory in many contexts. When an ECA, for example, is pointing to something on the screen, like a piece of merchandise or a map of a shop, this information has to be accessible also in an auditory (text to speech) version. Another challenging task is to use alternative modalities more creatively. For example, (non-speech) sounds of different intensity, duration and "corresponding" variations in tactile interfaces can be used for presentation of otherwise visual information for persons with visual impairment and an increase in picture and auditory support for text presentation can be used for persons with reading difficulties. The use of systems with multiple combinations of modalities for presenting different types of information to user groups with and without specific disabilities is an interesting area for research in cognitive science, communication and information technology (cf., e.g., Caporusso, Mkrtchyan and Badia 2010).

3.4. Body communication on different levels of awareness and intentionality (gestures, body posture, eye gaze, facial expression etc.)

Besides vocal-verbal, written information and pictures, communicative body movements are the most important type of multimodal communicative behavior in both HHC and HMC. Communicative body movements are often coordinated with prosody, so special attention is needed with regard to prosody in speech. Other challenges are posed by the need for integration of modalities, and for being able to handle this integration, both in the recognition and understanding of multimodal information (sometimes known as "fusion") and in the production and distribution of multimodal information (sometimes known as "fission"), in this way, making multimodal input and output available to HMC systems.

Some of the most important communicative body movements and other means of expression that need to be modeled both for production and reception of communicative behavior include the following (see also Allwood 2002):

- *head movements* (forward movement, backward movement; both nods and jerks, shakes; both left turn and right turn, forward movement, backward movement, tilts)
- *facial gestures* (smiles, frowns, wrinkle, mouth movements other than speech)
- *direction of eye gaze and mutual gaze*, eye movements
- *eye brow movements*
- *pupil size*
- *lip movements*
- *posture and posture shifts*
- *arm and hand movements*
- *shoulder movements*
- *movements of legs and feet*
- *spatial orientation movements*
- *clothes and adornments*

For speech, prosody (measured and observed through variations in intensity, pitch and duration) is very important. For example, prosody often provides information like which part of the utterance is in focus and/or provides new information. Compare, for example, *It is an old WOMAN* (as opposed to *man*) and *It is an OLD woman* (as opposed to *young*). Prosody also often identifies the type of communicative act, for example, if the utterance is a statement or a question. Compare the utterance *It is raining*, pronounced with a statement or question intonation. Nonverbal sounds also have certain communicative functions. Examples of nonverbal sounds with communicative functions are laughter or sighs and smacking sounds, which can both, for example, indicate or display a reaction of boredom or dissatisfaction in relation to another person's utterance. All of these also provide interesting challenges for automatic recognition in automatic systems.

3.5. Robustness

Multimodality can contribute to making interactive systems increasingly more able to handle *Own Communication Management (OCM)* (cf. Allwood 2007), that is, communication processes that enable choice of what is to be communicated and change of what has been expressed (sometimes also called repair, editing or correction), and *Interactive Communication Management (ICM)*, that is, communication processes that enable turn-taking, feedback and sequencing. The system also needs to handle difficulties in text understanding, difficulties in speech recognition and difficulties in picture/gesture recognition in a sensible way. One of the hopes in constructing multimodal systems is that robustness could be increased and that some of these difficulties could then be handled by

making use of the assumed redundancy of multimodal systems, that is, since similar and closely related information is expressed in several modalities, failure of recognition in one modality could be compensated by recognition in another modality and the two data streams can then be integrated into a more holistic multimodally based recognition.

3.6. Interaction and interactive features

In order to handle interactive communication in a robust manner, the system has to be able to flexibly handle both own communication management and interactive communication management and these functions need to be potentially adaptable to different activities and contexts. The system, thus, needs to handle:

– *Turn management*

How do we indicate, display or signal that speaker change is about to occur? Is it OK to interrupt other speakers? If so, when should interruptions occur? How long should the transition time be from one speaker to the next speaker? Is it OK to do nothing or be silent for a while? What can the speaker or the system do to keep a turn? How can they signal that they don't want the turn, but rather want the other party to continue? (Sacks, Schegloff and Jefferson 1974; Allwood 1999; De Ruiter, Mitterer and Enfield 2006; Magyari and De Ruiter 2008).

– *Feedback*

How do speakers indicate, display and signal to each other that they can/cannot perceive, understand or accept what their interlocutor is communicating (cf. Allwood 2002). Is this done primarily by auditory means (small words like *mhm, m, yeah* and *no*) or by visual means (head nods, head shakes, posture shifts etc.)? What emotions and attitudes do primarily occur in giving and eliciting feedback? Is very positive feedback preferred or is there a preference for more cautious feedback? (See Kopp et al. 2008.)

– *Sequencing*

What opening, continuing and closing communication sequences are preferred in a particular activity or culture, for example: What is the preferred way of starting an interaction in different activities (opening sequence)? What is the preferred way of closing (closing sequence)? When and how are greetings used? (See also Allwood et al. 2006.)

In studying naturalistic interaction, the role of "mirroring", imitation, automatic alignment, priming, contagion etc., that is, relatively automatic coordination of behavior between interactive individuals has received considerable attention in recent years (cf. Wachsmuth, Knoblich and Lenzen 2008). Making ECAs that can mirror and align requires more research on human behavior as

well as design and programming of the type of perception and expressive behavior that is required for relatively automatic coordination. Work on designing such systems has, for example, been done by Grammer and Wachsmuth (cf. Oberzaucher et al. 2008).

3.7. Types of content

Different types of content require different combinations of modalities to be expressed. An emotional content in human face-to-face interaction is easier to perceive, and usually more trustworthy, if words, prosody and facial gestures reinforce each other. Information about train departure times, on the other hand, is less dependent on reinforcement by prosody and facial gestures. The situation might here be different in an interactive multimodal system, where other kinds of multimodal information could be available and information about train departure times would best be given by a visual presentation of a train table, accompanied by highlighted parts being read aloud. More generally, interactive multimodal systems, for example, ECAs, are employed for various functions and, depending on the function, to a varying degree need specific combinations of modalities. However, the quest for naturalness in interaction continuously makes the issue of how different types of content are expressed more complex. In human face-to-face interaction, all modalities are available and the possibility to oscillate between different topics and even different activities and subactivities is often utilized (Allwood 2000, 2001, 2007). So, in the process of making an ECA appear as a reliable, natural communication partner, it is important to notice that most institutional human-human interactions involve confidence promoting and social alignment strategies, such as jokes and references to personal experiences, remarks about the weather, news etc. and that multimodality plays an important role in this process (e.g., smiles, body posture, prosody, facial expressions etc., in addition to fluent speech). If we want ECAs to be as natural as possible, these elements should be included. This means that also content usually expressed by indexical and iconic communication is important and that the ECA has to be able to deal with a number of topics, other than the topic that is the focus of the actual task being pursued. It therefore has to be able to change between different domains in a flexible way. As we have seen, this, besides the often task-specific factual main message content, also includes variation concerning culture and activity. This variation can then, for example, concern the content and functional areas of:

– *Identity*. How should the body and body movements indicate, display or signal who the agent is?
– *Physiological states*. What "physiological state" should be indicated, displayed or signaled by the agent and how?

– *Emotions*. What emotions are acceptable and appropriate in different activities?
– *Attitudes.* What attitudes, for example, regarding epistemic stance, politeness or respect, are appropriate?
– *Factual information.* What information is communicated? To some extent, this question is related to what is often called "information structure". What information is explicitly verbally in focus? What information is backgrounded and presupposed (perhaps multimodally available) in a particular situation? I can point to a car and say *new brake system* or perhaps say *it has a new brake system* or perhaps without nodding or pointing *that car has a new brake system.* In another situation, I might be answering a question, *Which cars have new brake systems?* and answer *that* or *that car*, perhaps accompanied by nodding or pointing. The factual information in all these examples is on some level the same, but the way it is focused and presented differs.
– *Everyday topics.* Included in factual information, we can ask what topics are regarded as neutral and possible to address even for strangers (e.g., politics, the weather, job, income etc.)?
– *Common speech acts*. Finally, we can consider what types of speech act are the most appropriate to convey the above types of information: What types of speech acts are commonly used in different activities, e.g. greetings, farewells and other typical exchange types?
– *Communication management*. How should the different types of communication management be accomplished? (Cf. Section 3.5 and 3.6 above).

4. Types of human-machine interaction

4.1. Non-digital machines

Multimodal communication has probably always been characteristic of the human species. In paleolithic times, stones were, for example, decorated by carvings, made by cut stones and painted with colors. Images of animals, persons, objects, often guiding rites and other activities, were created. Later on, tools for imprints in clay and pens, brushes, printing, typewriters etc. were developed, as writing gradually became an important way of communicating over distances in time and space. In the last 150 years or so, communication technology for images, photography, film and video as well as telegraphy, telephony and different types of audio recording has made multimedia communication, for example in newspapers and films, possible. Finally, in the last 20–30 years, with the advent of digital computer based technology, different media channels have been combined in many different ways (see Waltinger and Breuing in this volume).

4.2. Computers

4.2.1. Text only interfaces

At first, communication with and via computers occurred using text only. This is still the case for some applications and can pose problems, like when students are not able to represent a figure in taking notes at a lecture or, perhaps more seriously, when applications are not accessible to persons with low literacy skills.

4.2.2. Interface with pointer (or touch screen)

At a more recent stage, user friendliness was increased by introducing the desktop interface, together with menus and icons as well as interfaces using pointers, for example, mouse, pen or screen pointing. This added iconic and indexical dimensions to human-computer interfaces.

4.2.3. Voice communication

Voice communication with computers, for a long time has posed and still poses a real challenge to developers and users and is, for this reason, a very critical issue in trying to achieve naturalistic HCI. Speech synthesis has progressed from monotonous, metallic sounding robot speech with poor assimilation of adjacent sounds to more natural sounding concatenated speech, based on recordings of humans. But still many problems remain in generating naturalistic speech, not least in the areas of prosody for expressing affect and information structuring. Speech recognition remains to a great extent a so far poorly solved but very challenging problem. While systems for very limited domains, such as booking trains or flights or enquiring about telephone numbers, are in use and work sufficiently well, although not perfectly, voice dictation systems still need adaptation to individual speakers and tend to produce many errors. So, better recognition of naturalistic speech by computers is an important problem, which remains a challenge in the quest for on-line speech communication between humans and computers (Edlund et al. 2008).

4.2.4. Communication with a multimodal agent

While voice communication would make human-computer interaction similar to telephone conversation, and this would be adequate for many purposes, it is still incomplete, especially concerning speech recognition, as we have mentioned above. The goal of more naturalistic HCI is to enable humans as well as computers to produce and recognize multimodality in similar ways to what hu-

mans do. For this, simulations and representations of functions of the human body are needed. The design of embodied communicative agents (ECAs) is therefore an important area for multimodal communication technology dealing with "face-to-face-like" spoken and more or less multimodal communication. It is, in fact, even more challenging than voice communication, since the task of recognizing and producing naturalistic body communication is even less developed than the production and recognition of speech.

Simple "embodied" communicative agents used as interfaces to databases, today occur fairly frequently. Most typically, such systems have a human-looking face and upper body connected to a written dialog system for enquiries about services, merchandise, prices etc. The face of the "embodied" communicative agent can often add a few facial expressions linked to key words in the written input and output. The multimodality of this type of communication is, therefore, still extremely limited. One example of this type of ECA is IKEA's interface "Anna", which comes in a few slightly varying versions, but is essentially the same in most countries, showing three possible facial expressions accompanying text answers to queries about buying furniture (cf. Allwood and Ahlsén 2009).

Examples of AI-based, more advanced screen based embodied Communicative Agents, which can handle many HHC-like aspects are REA (Cassell et al. 1999) and GRETA (De Rosis et al. 1999), MAX (Wachsmuth 2005) and GANDALF (Thórisson 1997) (cf. also André et al. 1999; see also Kopp and Wachsmuth in this volume). In such ECAs, features like emotions (based on a PDA, i.e., pleasure, dominance, arousal, model), interaction management, and gestures have been modeled.

There are a number of application areas for ECAs and different forms of multi- modal communication strategies, for example, in tutorial systems for education and learning, in systems of interaction for children with autism spectrum dis- orders or other learning difficulties, in systems for giving various types of ad- vice, in social care-giving services and in entertainment. Another interesting area of HMC, where human-human multimodal communication strategies are relevant and could serve the purposes of providing more communicatively efficient, naturalistic interfaces is the area of mobile dialog systems.

4.2.5. Mobile dialog systems for ECAs and robots

Mobile dialog systems mostly rely on text or menu based communication, especially on the human input side, since, as we have mentioned above, speech recognition is hard to achieve, except for in very limited domains. In recent years more and more research has therefore concerned speech interfaces, with the purpose of making the interaction with different systems more naturalistic.

The introduction and increasing use of embodied communicative agents (and robots) in dialog systems has, as we have seen, led to greater interest in phenomena related to face-to-face spoken interaction, such as interaction management and communication of emotions and attitudes. One of the consequences of this is that designers of dialog systems now strive to make systems, which for a long time have been *asymmetric with respect to initiative* (in order to facilitate the processing tasks of the system), more symmetric in this respect, so that they give more equal opportunities for initiatives to user and system (e.g., Johnston et al 2001; see below). Another consequence is that making a dialog system multimodal means that the system, like in human-human communication, should be able to make use of different modalities *in coordination* (cf. Wahlster 2003). This in turn means that in order to achieve human-human like properties, dialog systems should possess *symmetric multimodality*, so that the different modalities can be used both for input and output. An example of a multimodally symmetric system is SMARTKOM, which uses an embodied communicative agent that handles a number of spoken interaction phenomena, for example, emotional prosody, gestures and backchannel feedback. Another example is the dialog system of the mobile robot BIRON, which is based mainly on the speech modality but can augment semantic representations by hypo- theses based on other modalities, for example, gestures (cf. Toptsis et al. 2004). Recent systems are moving from the "interface metaphor" to the "human metaphor" in exploiting more characteristics of human-human communication. This provides advantages concerning what the system can achieve, for example, increased naturalness and socially oriented communication, as well as challenges concerning what to achieve, for example, on-line prosodic analysis, communication management and context sensitivity etc. (Edlund et al. 2008).

Mobile dialog systems for robots, for example, the WITAS dialog system for multimodal communication with a robot helicopter, involve processing highly dynamic environments. Multimodal telecommunication systems (Lemon et al. 2003), like the MOBILTEL (Cismár et al. 2009), involve the use of different input mechanisms in a handheld device with a speech and graphical interface that includes an integrated VoIP (Voice over IP; see also Waltinger and Breuing in this volume) client as well as a pen, touch-screen, keyboard input and display including icons, emoticons, hyperlinks and scrolling menus, but without the usual HHC face-to-face features, such as gestures. Similar systems, like MATCH (*Multimodal Access to City Help*) (Johnston et al. 2001), provide a mobile speech-pen interface where the user can choose the modalities of communication, which are then integrated by the system in a speech-act based multimodal dialog manager which is symmetric with respect to initiative, that is, allows mixed-initiative dialog.

4.2.6. Tutoring systems

Another challenging and interesting area is the area of *Intelligent Tutoring Systems* (ITS) (Self 1998) or *Intelligent Learning Environments* (ILE) (Fernandez-Manjon et al. 1998). Such systems are often based on BDI (*Beliefs, Desires and Intentions*) cognitive models that traditionally have as their basic components (i) *domain knowledge*, (ii) a *user* (student) *model* and (iii) *pedagogic strategies*. Sometimes the systems have an "agent paradigm" from cognitive AI (Rao and Georgeff 1991). They can include *Multi-Agent Systems* (MAS) involving more than one user. Advanced systems for teaching in dynamic multi-agent virtual worlds also exist (e.g., Marsella and Johnson 1997). A student model can be "affective" as well as "cognitive" in a more narrow sense. Pedagogical "agents" are either modeled as cooperative agents who work in the background as part of the educational system or as personal, animated agents for human-computer interaction using voice and gesture and showing emotional attitudes (e.g., *Vincent* (Paiva and Machado 1998), *Steve* (Rickel and Johnson 1999, 2000), and *Cosmos* (Lester et al. 1997, 1999; cf. also Person et al. 2001; Vicari et al. 2009). An "agent paradigm" can be used for exploration of interaction and dynamic changes in the environment, for teaching and learning and there is an ambition to make the software more flexible in relation to the user's needs and preferences. Some features that are often needed in tutoring systems are strategies for human-computer interaction and for handling of multimedia information. To achieve a pedagogical human-computer interaction is extremely important for the teaching and learning process. Traditionally the main pedagogical functions of explanation, education and diagnosis have been implemented as a one-way mechanism, that is, the system is in control. So in tutoring systems, the ambition is to replace an asymmetric communication mode with respect to initiative with a symmetric communication mode between human and computer, including conflict solving by real negotiation, when this is needed. The BDI models, mentioned above, are today being extended with more social aspects, where, for example, expectations, confidence, planning and emotion are also modeled and interaction modes like negotiation, competition and cooperation are focused on more. This calls for more social behavior, which in turn entails the use of multimodality in communication.

We can see that dialog systems, including those used in tutoring systems, are increasingly being directed toward resembling human-human communication by including more human-like features.

5. Potential problems with human-like naturalistic ECAs for some applications

We have seen above, that there is considerable complexity and difficulty in designing naturalistic multimodal ECAs and that many tasks lie ahead of us in trying to achieve this. All the same, the research is being pursued and striving for increasing naturalness is a research agenda that is both important and fairly generally adopted. In the meantime, more limited multimodal strategies are being implemented in simplified and not always so natural artificial agents in different interfaces and applications, for example, in tutoring systems.

However, in relation to the development of "social agents", that is, agents that often have as a main purpose to provide, teach or train social interaction, it is important to consider the ethical ramifications of making ECAs optimally naturalistic. Research on social agents is described, for example, by Beazeal (2002) and Louwerse et al. (2008). According to Becker (2006), there are clear and insurmountable problems related to the goal of naturalness. In some contexts where ECAs are (successfully) used, for example, for communication with children with autism spectrum disorders (Dautenhahn 2007; Dautenhahn and Werrym 2004) or for communication with elderly persons, there are problems connected with simulating natural interaction, since it is not the same thing as providing real natural interaction. The agent might simulate emotions by mirroring etc, but actually does not have them, and this is very different from a human conversation partner. Becker especially points to the importance of eyes and voice, where there is no true multimodal symmetry between the ECA and the user. Becker poses the question of what a person is learning with respect to human-human interaction, by interaction with an ECA. The possibility of mixups of a real person, that is, skyping on the computer screen, with a very naturalistic ECA, by elderly persons suffering from dementia (one of the target groups for the design of supportive ECAs) can potentially lead to prob- lems, involving over-confidence in the abilities of the ECA, which could be risky in some situations. Ethical concerns have to be addressed in using natural- istic ECAs, in some applications. As a result, Becker advocates the alternative of keeping to more limited tasks and domains where not all (or as many as possible) of the features of a human communicator are implemented. The potential ethical problems involved in using naturalistic ECAs are a controversial area where more research is needed.

6. Conclusion and outlook

The requirements on multimodal communication systems, if they are to be optimally naturalistic with respect to human-human communication, are considerable, as we have seen above in the description of features of human-human communication.

There are huge challenges for the enterprise of providing naturalistic communication. One of them is to provide automatic speech recognition which is not just limited to a restricted domain and which includes some processing of expressions for emotions and attitudes (see also Martin and Schultz in this volume). Another is the achievement of naturalistic gesturing, facial expressions and body posture as well as the recognition of these features.

Two lines of research and development are already obvious and are likely to persist in future work:

(i) the pursuit of greater understanding of natural features of human-human communication in different modalities, that is, basic research on what can then be potentially available to be modeled;
(ii) the development and implementation of application specific multimodal ECAs with limited repertoires of features that are judged to be efficient for specific purposes.

Both lines of research and development will benefit from an exchange of ideas, methods and findings. In both types of research, ethical considerations have to be taken into account. In the development of applications, research on how actual and potential users in fact respond and react attitudinally and emotionally is important for the success of the application and can probably also lead to a more realistic appreciation of the possible ethical problems by providing more specified insights and guidelines.

There is no doubt that multimodal communication technology will be increasingly used in our everyday social life, both in professional tasks and leisure activities. Especially in computer gaming, multi-modal interfaces are being developed rapidly (cf., for example, Sargin et al. 2008; Liu and Kavakli 2010) and also address the challenge of special needs (e.g., Caporusso, Mkrtchyan and Badia 2010). Finally, it is likely that research in this area, based on features of human-human communication, in turn, enhanced by the developments of computer- and internet-based functions, will bring about new and exciting ways of communication.

References

Ahlsén, Elisabeth
2008 Embodied communication – aphasia, apraxia and the possible role of mirroring and imitation. *Clinical Linguistics and Phonetics* 22(4–5): 1–5.

Allwood, Jens
1976 *Communication as Action and Cooperation. Gothenburg Monographs in Linguistics* 2. Gothenburg: University of Gothenburg, Department of Linguistics.

Allwood, Jens
1995 *An activity based approach to pragmatics. Gothenburg Papers in Theoretical Linguistics* 76. Gothenburg: University of Gothenburg, Department of Linguistics.

Allwood, Jens
1999 Are There Swedish Patterns of Communication? In: Hiroshi Tamura (ed.), *Cultural Acceptance of CSCW in Japan and Nordic Countries*, 90–120. Kyoto: Kyoto Institute of Technology.

Allwood, Jens
2000 An Activity Based Approach to Pragmatics. In: Harry Bunt and Bill Black (eds.), *Abduction, Belief and Context in Dialogue: Studies in Computational Pragmatics*, 47–80. Amsterdam: John Benjamins.

Allwood, Jens
2001 Capturing Differences between Social Activities in Spoken Language. In: Istvan Kenesei and Robert M. Harnish (eds.), *Perspectives in Semantics, Pragmatics and Discourse*, 301–319. Amsterdam: John Benjamins.

Allwood, Jens
2002 Bodily Communication – Dimensions of Expression and Content. In: Björn Granström, David House and Inger Karlsson (eds.), *Multimodality in Language and Speech Systems*, 7–26. Dordrecht: Kluwer Academic Publishers.

Allwood, Jens
2007 Activity Based Studies of Linguistic Interaction. *Gothenburg Papers in Theoretical Linguistics* 93. Gothenburg: University of Gothenburg, Department of Linguistics.

Allwood, Jens
2008 Dimensions of Embodied Communication – towards a typology of embodied communication. In: Ipke Wachsmuth, Manuela Lenzen and Günter Knoblich (eds.), *Embodied Communication in Humans and Machines,* 257–284. Oxford: Oxford University Press.

Allwood, Jens, Elisabeth Ahlsén., Johan Lund and Johanna Sundqvist
2006 Multimodality in Own Communication Management. In: *Proceedings from the Second Nordic Conference on Multimodal Communication. Gothenburg Papers in Theoretical Linguistics* 92(2): 10–19. Gothenburg: University of Gothenburg, Department of Linguistics.

Allwood, Jens and Elisabeth Ahlsén
2009 Multimodal Intercultural Interaction and Communication Technology – A conceptual framework for designing and evaluating Multimodal Intercultural Communicators. In: Michel Kipp, Jean-Claude Martin, Patrizia Pag-

gio and Dirk Heylen (eds.), *Multimodal Corpora*, 160–175. LNCS 5509. Berlin: Springer.

Allwood, Jens, Joakim Nivre and Elisabeth Ahlsén
1992 On the Semantics and Pragmatics of Linguistic Feedback. *Journal of Semantics* 9: 1–26.

André, Elisabeth, Thomas Rist and Jochen Müller
1999 Employing AI methods to control the behavior of animated interface agents. *Applied Artificial Intelligence* 13: 415–448.

Arbib, Michael
2005 From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics. Behavioral and Brain Sciences 28(2): 105–124.

Beazeal, Cynthia L.
2002 *Designing Sociable Robots*. Cambridge, Massachusetts: The MIT Press.

Becker, Barbara
2006 Social robots – emotional agents: Some remarks on naturalizing man-machine interaction. *International Review of Information Ethics* 6(12): 37–45.

Boukricha, Hana, Ipke Wachsmuth, Andrea Hofstätter and Karl Grammer
2009 Pleasure-Arousal-Dominance driven facial expression simulation. In: *3rd International Conference on Affective Computing and Intelligent Interaction, ACII 2009*, 119–125. Amsterdam: IEEE Press.

Caporusso, Nicholas, Lusine Mkrtchyan and Leonardo Badia
2010 A multimodal interface device for online board games designed for sight-impaired peope. *IEEE Transactions Information Technology Biomedicine* 14(2): 248–254.

Cassell, Justine, Joseph Sullivan, Scott Prevost, Scott and Elizabeth Churchill (eds.)
2000 *Embodied Conversational Agents.* Cambridge, Massachusetts: The MIT Press.

Cismár, Anton, Matús Pleva, Ján Papaj, Lubomir Dobos and Jozef Júhar
2009 MOBILTEL Mobile multimodal telecommunications dialogue system based on VoIP telephony. *Journal of Electrical and Electronics Engineering* 2: 134–137. *Collins English Dictionary – Complete and Unabridged 10th Edition*
2009 Fulham Palace: Harper Collins Publishers.

Dautenhahn, Kerstin
2007 Socially intelligent robots: dimensions of human – robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences* 362(1480): 679–704.

Dautenhahn, Kerstin and Iain Werrym
2004 Towards Interactive Robots in Autism Therapy: Background, Motivation and Challenges. *Pragmatics and Cognition* 12(1): 1–35.

DeRosis, Fiorella, Catherine Pelachaud, Isabella Poggi, Valeria Carofiglio and Berardina DeCarolis
2003 From Greta's mind to her face; modeling the dynamics of affective states in a conversational embodied agent. *International Journal of Human-Computer Studies* 58(1–2): 81–118.

De Ruiter, Jan Peter, Holger Mitterer and N. J. Enfield
2006 Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language* 82(3): 515–535.

Edlund, Jens, Joakim Gustafson, Mattias Heldner and Anna Hjalmarsson
2008 Towards human-like spoken dialogue systems. *Speech Communication* 50(8–9): 630–645.

Fernandez-Manjon, Baltasar, Juan Cigarran, Antonio Navarro and Alfredo Fernandez-Valmayor
1998 Using automatic methods for structuring conceptual knowledge in intelligent learning environments. In: *Proceedings of Intelligent Tutoring Systems, 4th International Conference, ITS' 98 San Antonio, Texas, USA, August 16–19*, 264–273. LNCS 1452. Berlin: Springer.

Gibbon, Dafydd
2012 Resources for technical communication systems. In: Alexander Mehler, Laurent Romary and Dafydd Gibbon (eds.), *Handbook of Technical Communication*. Berlin/Boston: De Gruyter.

Goodwin, Charles and Marjorie Harness Goodwin
1997 La Coopération au travail dans un aéroport [Cooperation at the workplace in an airport]. *Réseaux* 85, 129–62. (Special issue "La Coopération dans les Situations de Travail" edited by Dominique Cardon).

Grice, H. Paul,
1953 Meaning. *The Philosophical Review* 66: 377–388.

Jan, Dusan and David Traum,
2007 Dynamic Movement and Positioning of Embodied Agents in Multiparty Conversations. In: *Proceedings of ACL 2007 Workshop on Embodied Language Processing*, 59–66.

Johnston, Michael, Srinivas Bangalore and Gunarajnan Vasireddy
2001 Match: Multimodal Access to City Help. In: *Proceedings of the Automatic Speech Recognition and Understanding Workshop, Madonna di Campliglio*, 2001.

Komiak, Sherrie Y. X., Weiquan Wang and Izak Benbasat
2004 Trust building in virtual salespersons versus in human salespersons: similarities and differences. *e-Service Journal* 3(3): 49–64.

Kopp, Stefan, Paul A. Tepper, Kim Ferriman, Kristina Striegnitz and Justine Cassell
2007 Trading spaces: How humans and humanoids use speech an gesture to give direction. In: Toyoaki Nishida (ed.), *Conversational Informatics: An Engineering Approach*. Chichester: John Wiley & Sons.

Kopp, Stefan, Jens Allwood, Karl Grammer, Elisabeth Ahlsén, and Thorsten Stocksmeier
2008 Modeling Embodied Feedback with Virtual Humans. In: Ipke Wachsmuth and Günther Knoblich (eds.), *Modeling Communication with Robots and Virtual Humans*, 18–37. LNAI 4930. Berlin: Springer.

Kopp, Stefan, Kirsten Bergmann and Ipke Wachsmuth
2008 Multimodal communication from multimodal thinking – Towards an integrated model of speech and gesture production. *Semantic Computing* 2(1): 115–136.

Kopp, Stefan and Ipke Wachsmuth
2012 Artificial interactivity. In: Alexander Mehler, Laurent Romary and Dafydd Gibbon (eds.), *Handbook of Technical Communication*. Berlin/Boston: De Gruyter.

Kubina, Petra and Andy Lücking
2012 Barrier-free communication. In: Alexander Mehler, Laurent Romary and Dafydd Gibbon (eds.), *Handbook of Technical Communication*. Berlin/Boston: De Gruyter.

Le Bel, Ronald M., Jaime A. Pineda and Anu Sharma
2009 Motor-auditory-visual integration: The role of the human mirror neuron system in communication and communication disorders. *Journal of Communication Disorders* 42(4): 299–304.

Lemon, Oliver, Anne Bracy, Alexander Gruenstein and Stanley Peters
2003 An information state approach in a multi-modal dialogue system for human-robot conversation. In: Hannes Rieser, Peter Kühnlein and Henk Zeevat (eds.), *Perspectives on Dialogue in the new Millennium*, *Pragmatics and Beyond*, 229–242. Amsterdam: John Benjamins.

Lester, James C., Sharolyn A. Converse, Susan E. Kahler, S. Todd Barlow, Brian A. Stone and Ravinder S. Bhogal
1997 The persona effect: Affective impact of animated pedagogical agents. In: *Proceedings of CHI '97*, 359–366.

Lester, James C., Jennifer E. Voerman, Stuart G. Towns and Charles, B. Callaway
1999 Deictic believability: Coordinating gesture, locomotion, and speech in like-like pedagogical agents. *Applied Artificial Intelligence* 13: 383–414.

Levinson, Stephen, C.
1983 *Pragmatics*. Cambridge: Cambridge University Press.

Liu, Jing and Manolia Kavakli
2010 A survey of speech-hand gesture recognition for the development of multi-modal interfaces in computer games. In: *Proceedings of the 2010 IEEE International Conference on Multimedia and Expo, ICME 2010, 19–23 July, Singapore*, 1564–1569.

Louwerse, Maxim M., Arthur, C. Graesser, Danielle McNamara and Shulan Lu
2008 Embodied conversational agents as conversational partners. *Applied Cognitive Psychology* 23(9): 1244–1255.

Lücking, Andy and Thies Pfeiffer
2012 Framing multi-modal technical communication. In: Alexander Mehler, Laurent Romary and Dafydd Gibbon (eds.), *Handbook of Technical Communication*. Berlin/Boston: De Gruyter.

Magyari, Lilla and Jan Peter De Ruiter
2008 Timing in conversation: The anticipation of turn endings. In: Jonathan Ginzburg, Pat Healey, and Yo Sato (eds.), *Proceedings of the 12th Workshop on the Semantics and Pragmatics Dialogue*, 139–146. London: King's college.

Marsella, Stacey C. and W. Lewis Johnson
1997 An instructor's assistant for team-teaching in dynamic multi-agent virtual worlds. In: *Proceedings of the Fourth International Conference on Intelligent Tutoring Systems*, 464–473. Berlin: Springer.

Martin, Jean-Claude and Tanja Schultz
2012 Multimodal and speech technology. In: Alexander Mehler, Laurent Romary and Dafydd Gibbon (eds.), *Handbook of Technical Communication*. Berlin/Boston: De Gruyter.

Oberzaucher, Elizabeth and Karl Grammer
2008 Everything is movement: on the nature of embodied communication. In: Ipke Wachsmuth, Manuela Lenzen and Günter Knoblich (eds.), *Embodied communication*, 151–177. Oxford: Oxford University Press.

Paiva, Ana and Isabel Machado
1998 Vincent, an autonomous pedagogical agent for on-the-job training. In: Valerie Shute (ed.), *Intelligent Tutoring Systems*, 584–593. Berlin: Springer.

Partan, Sarah R. and Peter Marler
2005 Issues in the classification of multimodal communication signals. *The American Naturalist* 166(2): 231–245.

Peirce, Charles Sanders
1931 Principles of Philosophy. Vol. 1, Book 2. § 369, p. 195. In: *Collected Papers of Charles Sanders Peirce, 1931–1958*, 8 vols. Edited by Charles Hartshorne, Paul Weiss and Arthur Burks. Cambridge, MA: Harvard University Press.

Person, Natalie K., Arthur C. Graesser, Roger J. Kreuz and Victoria Pomeroy
2001 Simulating human tutor dialog moves in AutoTutor. *International Journal of Artificial Intelligence in Education* 12: 23–29.

Pomerantz, Anita M.
1984 Agreeing and disagreeing with assessment: Some features of preferred/dispreferred turn shapes. In: J. Max Atkinson and John Heritage (eds.), *Structure of Social Action: Studies in Conversation Analysis*, 57–101. Cambridge: Cambridge University Press.

Rao, Anand and Michael Georgeff
1991 Modeling Rational Agents within a BDI-Architecture. In: *Proceedings of Knowledge Representation and Reasoning (KR&R 1991)*, 473–484.

Rickel, Jeff., and W. Lewis Johnson
1999 Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. *Applied Artificial Intelligence* 13: 343–382.

Rickel, Jeff and W. Lewis Johnson
2000 Task-oriented collaboration with embodied agents in virtual worlds. In: Justine Cassell, Joseph Sullivan, Scott Prevost and Elizabeth Churchill (eds.), *Embodied Conversational Agents*. Cambridge, Massachusetts: The MIT Press.

Ruttkay, Zsófia and Catherine Pelachaud (eds.)
2004 *From Brows till Trust: Evaluating Embodied Conversational Agents*. Dordrecht: Kluwer.

Sacks, Harvey, Emanuel A. Schegloff and Gail Jefferson
1974 A simplest systematics for the organization of turn-taking for conversation. *Language* 50: 696–735.

Sargin, Mehmet E., Yucel Yemez, Engin Erzin and Ahmet M.Tekalp
2008 Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(8): 1330–1345.

Shank, Roger and Robert Abelson
1977 *Scripts, Plans, Goals and Understanding*. Hillsdale, HJ: Erlbaum.
Searle, John R.
1970 *Speech Acts: An Essay in the Philosophy of Language*. Cambridge, Massachusetts: Cambridge University Press.
Self, John
1999 The defining characteristics of intelligent tutoring systems research: ITS care, precisely. *International Journal of Artificial Intelligence in Education* 10: 350–364.
Thórisson, Kris R.
1997 Gandalf: An Embodied Humanoid Capable of Real-Time. Multimodal Dialogue with People. In: *Proceedings of the first ACM International Conference on Autonomous Agents, Marina del Rey, California, February 5–8*, 536–537.
Toptsis, Ioannis, Shuyin Li, Britta Wrede and Gernot A. Fink
2004 A Multi-modal Dialogue System for a Mobile Robot. In: *Proceedings of the International Conference on Spoken Language Processing*, 273–276.
Vicari, Rosa Maria, Patricia Augustin Jaques and Regina Verdin (eds.)
2009 *Agent-Based Tutoring systems by Cognitive and Affective Modeling*. Hershey, PA, USA: IGI Global.
Wachsmuth, Ipke
2005 "Ich, Max" – Kommunikation mit künstlicher Intelligenz. In: Ch. S. Herrmann, M. Pauen, J. W. Rieger and S. Schicktanz (eds.), *Bewusstsein: Philosophie, Neurowissenschaften, Ethik,* 329–354. München: Wilhelm Fink Verlag.
Wahlster, Wolfgang
2003 SmartKom: Symmetric Multimodality in an Adaptive and Reusable Dialogue Shell. In: *Proceedings of the Human Computer Interaction Status Conference 2003*, 47–62. Berlin: DLR.
Waltinger, Ulli and Alexa Breuing
2012 Internet-based communication. In: Alexander Mehler, Laurent Romary and Dafydd Gibbon (eds.), *Handbook of Technical Communication*. Berlin/Boston: De Gruyter.