Gaze and Gesture Activity in Communication[1]

Kristiina Jokinen

University of Helsinki

Non-verbal communication is important in order to maintain fluency of communication. Gestures, facial expressions and eye-gazing function as non-verbal means to convey feedback and provide subtle cues to control and organise conversations. In this paper, verbal and non-verbal feedback are discussed from the point of view of how they contribute to the communicative activity in conversations, especially the type of strategies that the speakers deploy when they aim to construct shared understanding of the tasks and duties in interaction in general. The study concerns conversational data, collected for the purposes of designing and developing more natural interactive systems.

1. Introduction

Recently verbal and non-verbal aspects of communication have become popular research topics in interaction technology. Understanding how intonation conveys the speaker's attitudes and emotional state, and how gestures, facial expressions and body posture support, complement, and in some cases, override verbal communication, is necessary for modelling interaction management. The knowledge is crucial also in the design and implementation of systems that can adapt themselves to different users in different environments and in different cultural contexts: intelligent, interactive and context-aware applications require knowledge on natural interaction strategies and what it means to communicate in a natural way. Often these aspects have been overlooked in dialogue system design because of technological constraints or lack of larger theoretical views of how human communication takes place. However, information management and multimodal user interface are considered among the main research challenges for enabling technologies, and interactions with smart objects, services and environments need to address challenges concerning natural, intuitive, easy, and friendly interaction. For instance, Norros et al. (2003) list speech, hearing, gesturing and gaze as serious modality candidates to enhance future Human Technology Interfaces, since these are extensively used in human-human interaction.

Active research is going on concerning the types and communicative functions of various gestures and facial expressions in order to learn more about natural human communication, and to enable more intuitive and flexible interactions between human

[1] In C. Stephanidis (Ed.): *Universal Access in Human-Computer Interaction.* Proceedings of the 5th International Conference, UAHCI 2009, Held as Part of HCI International 2009, San Diego, CA, USA, July 19-24, 2009.

users and computer agents. Besides corpus-based empirical investigations, ECAs (Embedded Conversational Agents, Cassel et al., 2003; André & Pelachaud, 2009) and virtual world characters have been used to experiment and develop more intuitive interaction techniques. Also robotic companions have been developed so that they can recognize speech and gestures, and so become engaged in multimodal communication (Bennewitz et al., 2007). New application areas have also appeared for non-verbal communication techniques: various games and educational toys that use the novel technology can allow users, especially children with special needs to enjoy and be empowered by the new technology. In second language learning and intercultural communication studies non-verbal communication also appears important: students need to learn to observe relevant communicative gestures and to produce suitable gestures themselves in order to communicate fluently.

In this paper we look at verbal and non-verbal feedback from the point of view of the joint activity that the partners show in their gestures, facial expressions, and body posture. Through the collection of large corpora and examination of communication in natural situations, we study the regulating function of non-verbal communication, and how the speakers use different non-verbal means to manage dialogues in multi-party settings. The study concerns especially the synchronisation of communicative activity that the partners show in their reactions to the partner's presentations, and the kind of gesturing they use to construct shared understanding of the tasks and duties in communication. Non-verbal activity is thus tied to the notion of feedback and to the general dialogue strategies modelled for the purposes of designing and developing more natural interactive systems. The specific research questions concern:

– What kind of non-verbal communication takes place in human interactions?
– What kind of interrelations can be found among the various types of non-verbal communication?
– Can the correlations be measured and modelled for interactive systems?

The paper is structured as follows. Section 2 discusses different non-verbal signals, gestures, face and body posture in providing feedback, and Section 3 presents the data and examples of the gestures and body communication which function in the coordination of communication. Section 4 presents their contribution to the dialogue activity and synchrony among the participants. Section 5 draws conclusions and points to further research topics.

2. Non-verbal Feedback

Natural and intuitive dialogue phenomena do not only include spoken words and utterances, but also vocal aspects and gesticulation which do not appear in written language. Such phenomena include:

- Hesitations (silence, sound prolongation, hesitation markers, repetitions)
- Discourse markers (*well, I mean, oh, ah*)
- Backchannels (*uhuh, mmhmm, ok*)
- Speech signals (changes in voice quality, speaking rate, pitch, intensity)
- Eye movements (focus of attention)

- Head movement (related to focus of attention, turn taking)
- Facial expressions (reflecting states of mind)
- Hand gestures (indicating rhythm of the speech, pointing, icons)
- Body posture (controlling their involvement in the discussions).

Often these phenomena work in synchrony. For instance, turn-taking and backchannelling are based on prosodic and syntactic features (Koiso et al., 1998) and can also be supported by eye-contact which signals the end of turn or the speaker's intention to take turn. The importance of gaze can also be seen in establishing the focus of shared attention in classroom interactions and meetings: gaze direction serves to frame the interaction and establish who is going to speak to whom, and about what. Moreover, non-verbal signals serve social functions, creating bonds and shared knowledge, as well as reflecting attitudes, mood and emotions of the speakers (Feldman & Rim, 1981).

Much of the conversational information exchange relies on the assumptions that are not necessarily made explicit in the course of the interaction. One of the main challenges in interaction management thus lies in the grounding of language: finding the intended referents for the partner's expressions and regulating the flow of information so as to construct shared knowledge of what the conversation is about (Clark & Schaefer, 1989; Traum, 1999; Jokinen, 2009a, 2009b). Non-verbal signals provide an effective means to contribute to the mutual understanding of the conversation, and to update one's knowledge without interrupting verbal presentation. For instance, looking at the conversational partner or looking away can provide indirect cues of the speaker's willingness to continue interaction, gazing at particular elements in the vision field tells what the focus of attention is, while gesturing usually catches the partner's attention and marks relevant parts of a message.

Feedback can be classified according to the strength of commitment on the feedback giver's side: at its weakest we talk about backchannelling as providing feedback about the basic enablements of communication (contact, perception and understanding), and at its strongest the speaker gives feedback by agreeing on the content of the utterance. Non-verbal feedback is usually backchannelling, i.e. automatic signalling of “the channel being open” rather than conscious and intentional exchange of symbolic information: it contributes to the fluency of communication and allows the participants to monitor the state of interaction quickly and unobtrusively. However, non-verbal signals are usually ambiguous and their interpretation requires that the communicative context is taken into account. Different levels of context are relevant, ranging from the utterance itself to the roles and cultural background of the participants (see Jokinen & Vanhasalo, 2009). One of the necessary conversational skills is thus to know how and when to enable the right type of contextual reasoning: participants need to observe each others’ reactions and changes in their emotional and cognitive states so as to draw appropriate contextual inferences.

The form of non-verbal signals also provides important information about the meaning of the signals in a given context. For instance, the form of gestures (hand shape, movement, use of fingers) can vary from rather straightforward picturing of a referent (iconic gestures) to more abstract types of symbolic gestures, up to culturally governed emblems (such as the sign of victory). Although gestures are often culture-specific, some forms seem to carry meaning that is typical of the particular hand shape itself. Kendon (2004), for instance, talks about different gesture families which

describe the semantics and function of gestures in their context. This supports multi-functionality of gestures and also the gestures forming a continuum rather than a classification of categories. Furthermore, it allows wider and more communicative interpretation of gestures: gesture families guide the partner towards certain general pragmatic interpretations (e.g. “I'm offering this information for your consideration and expect your evaluation of it” vs. “I think these points are important and expect you to pay attention to them”), and the verbal content then specifies the meaning of the gesture by expressing what detailed information is being offered or considered important. The gesture can of course also occur alone, in which case its interpretation is based on the semantic theme typical of the hand shape.

3. Conversational data and analysis

The corpus consists of conversations collected in an international setting at the ATR Research Labs in Japan, and includes videos of four participants engaged in free-flowing conversation. One of the speakers knows all the others while the others are unfamiliar with each other, although share some background knowledge of the culture and living in Japan. All the interlocutors speak English but represent different cultural backgrounds and language skills. In order to collect as natural data as possible the participants' topics or activities were not restricted in advance. The three about 1,5 hour long conversations were recorded during three consecutive days and consist of casual conversations in an unobstructed technical setting. The technical setup for the collection is similar to the one described in Douglas et al. (2003) and Campbell & Ohara (2005), while the corpus annotation is reported in Campbell & Jokinen (2008).

The ATR data is transcribed, and a small part of it is annotated with respect to non-verbal gesticulation. The analysis is based on the MUMIN coding scheme (Allwood et al. 2005), developed as a general tool for the annotation of communicative functions of gestures, facial expressions, and body posture. The communicative functions are related to the use of the three non-verbal modalities in turn-taking and feedback giving processes, and also in sequencing information. Annotation also takes into account the general semiotic meaning of communicative elements: they can be indexical (pointing), iconic (describing) and symbolic (conventional) signs. Also the form is annotated for each communicative element: for gestures this includes e.g. the shape of hand, palm, fingers, and hand movement, for face and head this includes the shape and combination of eyes, eye-brows, mouth, and the movement of head, and for the body posture, the leaning back- and forward.

One of the conversational situations is shown as an example in Figure 1. The speaker (second from left) had just explained how her friend had went to an electrical shop and bought a special massaging shower and also tried a massage chair that she had liked. The speaker in the back right had suggested it is a healing massage, aiming to share background information that the friend was interested in various types of Eastern healing practices. The gesturer (the one in the front right) now wants to make sure she has understood the concept correctly, and asks for a clarification *do you mean this chair or ... or herself*. The clarification is accompanied by a gesture with the open hand moving up-down and emphasizing the phrase *the chair* and, after a

short pause and hesitation (*or ... or*), also the second alternative *herself*, although less visibly. During the speaking of the first alternative, all the partners look at her, but they move their heads simultaneously towards the original speaker during the second emphasis in anticipation of a clarifying response (Figure 2).

Figure 1. Gesture emphasising "do you mean this chair or ... or herself"

Figure 2. Gesture and face turning to expect a response.

The gesture is a typical emphasizing gesture which does not only accompany or complement the spoken content but also functions as an independent means for interaction management. It is related to what McNeill (2005) calls catchment. Kendon (2004) interprets this kind of gesture as a pragmatic gesture on the meta-discursive level. The particular shape of the gesture belongs to the gesture family of Palm Open Supine Vertical with the semantic theme of cutting, limiting, structuring information. Jokinen & Vanhasalo (2009) have called these gestures stand-up gestures since they can be distinguished from the normal flow of information presentation (i.e. beats) so as to direct the partner's attention, structure the information flow, and create mutual

context (they also belong to the normal repertoire of gesturing in successful stand-up comedies). In this particular example, the gesture cuts information flow and marks a particular part of the verbal content as something that the speaker considers relevant in the context to be clarified. The gesture focuses the partners' attention onto the two particular alternatives, and at the same time it limits the conversational topic to the clarification of these items. Thus, instead of expressing a set of meanings explicitly in an utterance, the speaker uses a non-verbal gesture that conveys them in an unobstructed manner, and simultaneously with the verbal content: how the building of mutual context is progressing (clarification needed), which part of the message is in focus (the two alternatives), and how the conversation is to be structured (divided into segments: presentation – clarification – explanation).

Notable differences can be found in the manner and frequency in which the speakers provide feedback. The basic statistics in Table 1 shows how the speakers differ concerning the types of non-verbal signals produced: although the average number of gestures seems to be the same across the participants, the number of head and body movement differ greatly.

	Total	Average	D	Y	K	N
Head	295	74	25	119	88	63
Gesture	133	33	37	32	34	30
Posture	69	17	4	16	10	39
	497	124	66	167	132	132

Table 1. Basic statistics of non-verbal signals.

There are of course personal and cultural reasons for the differences, but they may also be due to the underlying neurocognitive properties of non-verbal communication. For instance, if gesturing is closely related to thinking and motor control of speech, it is understandable that most speakers produce the same amount of gestures on average during speaking. Similarly, head turns and facial expressions, described with the help of the shape and movement of eyes, eye-brows, mouth, etc. have connections to the speaker's emotional state and focus of attention: facial expressions, at least as signs of true emotions, are produced via an innate mechanism of a cognitive stimulus firing neurons that control facial muscles (although the recognition and production of emotional expressions may be culturally conditioned too, see Ekman, 1972). Face and head movements, however, show the participants' reactions to what has taken place, and thus vary among the individuals since they appear as automatic reactions based on their personalities; moreover, the underlying reactions seem to be related to those aspects of social life (emotions and attitudes) which are strongly controlled by cultural roles and expectations and which the speakers have internalised as acceptable communicative behaviour. Thus, as means of non-verbal feedback, face and head movement seem to belong to the behaviour patterns that are learnt in order to indicate appropriate level of understanding, interest, and willingness to listen to the partner, while gesturing is related to the speakers' own communication management which

may be controlled by the speakers' needs and intentions to express themselves more than by social norms. Body movement, which to a large extent expresses “personal space”, can also indicate participation in the conversation, either as an active participant or an onlooker, and is thus also governed by social norms. Consequently, the observed differences in the frequencies of the different non-verbal signals across the participants can be related to their origin in the production of the speaker's own speech vs. reaction to the partner's speech, and to the culturally conditioned politeness and social norms which govern the appropriate level of expressiveness.

4. Dialogue activity

As seen from the previous data, conversations are full of simultaneous activity which requires subtle coordination by the participants. Models of feedback giving processes are thus relevant in order to design and develop more natural interactive systems, and active research using annotated corpora and machine-learning techniques is going on. Some work in this respect, using the annotated MUMIN categories to classify and cluster the information, is reported e.g. in Jokinen (2008).

To see how the conversational activity is distributed as signal-level observations, we can visualize the conversational activity as activity bars along the speakers' speech. Figure 3 below shows visualization of the verbal and gesture activity during a 25 minutes long dialogue excerpt from the ATR data, for the two speakers Y and D.[2] The horizontally depicted activity bars indicate which of the speakers are verbally active at a particular time, i.e. either speaking, laughing or backchannelling, while the vertical peaks show the speakers' movement as detected on the video: gesturing, head turns and nods, as well as body leaning forward and backward.

2 I would like to Stefan Scherer for the video and speech analysis and producing the pictures.

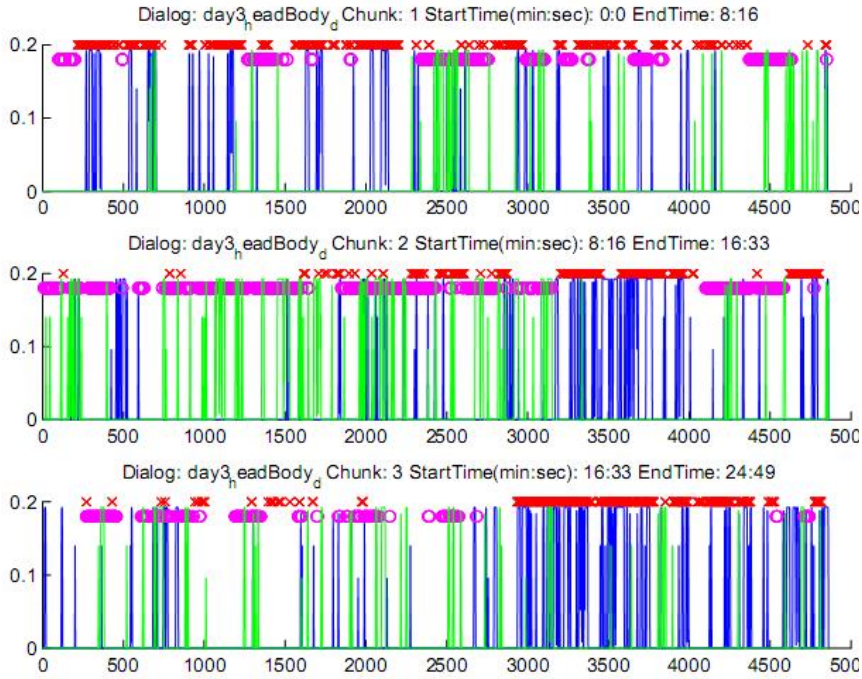

Figure 3. Conversational activity of speaker Y (above) and speaker D (below).

Speaker Y (the upper verbal bar) is the most active interlocutor concerning the annotated non-verbal activity, while Speaker D (the lower verbal bar) is the least active speaker. As can be seen, Y is very active in the beginning of the conversational excerpt, and has long stretches of speech in the middle and at the end, while D is fairly active in the beginning and end of the conversation, and has a long monologue type speech in the middle. Their speech overlap and also shows coordinated turn-taking (however, since the speech by the other two speakers is cut off, the actual coordination of the interlocutors turn taking and verbal activity is not shown in the figure). Speaker Y provides verbal feedback more than D as can be seen in the several crosses and circles on Y's speech bar, for instance, in the middle of the conversation where Speaker D has a long turn. These indicate that the length of Y's verbal activity was very short at these points (this kind of backchannelling is also supported by the top-down manual annotations).

What is clearly seen in the figure is the connection between speech and non-verbal body and gesture activity. There are clear peaks in the speaker's movement and speaking: when the speaker starts to speak, there is typically an action with their hands and/or body. Movement activity appears less when the speaker is listening although this also depends on the speaker. In general, these observations match with the assumptions made about the use of different non-verbal feedback signs in the conversation. As for the other speakers, their conversational activities against the other participants are shown in Figures 4 and 5 (next page).

5. Conclusions

In the beginning of the paper we asked three questions about the form and function of non-verbal communication, and we can now conclude the paper by providing answers on the basis of the research described above. The kind of non-verbal communication that we have focussed on concerns gestures, facial expressions, and body posture, and their functioning in different communicative functions. On the basis of real conversational examples, we have shown that the participants coordinate their activities in an accurate manner, and effectively use the signals to give feedback, coordinate turn taking, and build shared context. We have also visualized the speakers' non-verbal activity against their verbal activity, and thus contributed to the on-going research on specifying correlations and interrelations among the various types of non-verbal communication signals.

As future work, we plan to specify correlations between the non-verbal signals and speech properties such as intonation and the quality of voice. Work is also going on concerning the comparison of the top-down annotations and bottom-up signal processing, and we expect to learn more about the interplay between verbal and non-verbal interaction, and how observations of low-level signals match on the linguistic-pragmatic categories and human cognitive processing. As human-computer interactions get wider and more complex, the resulting models of interaction will provide us with a better understanding of the enablements for communication and the basic mechanisms of interaction that are crucial for flexible and intuitive interaction management. Intuitive communication strategies can improve the rigid and simple interactions that present-day systems exhibit, and thus the research also encourages interdisciplinary research where human and social sciences look into technological possibilities of application to the design and construction of interactive systems.

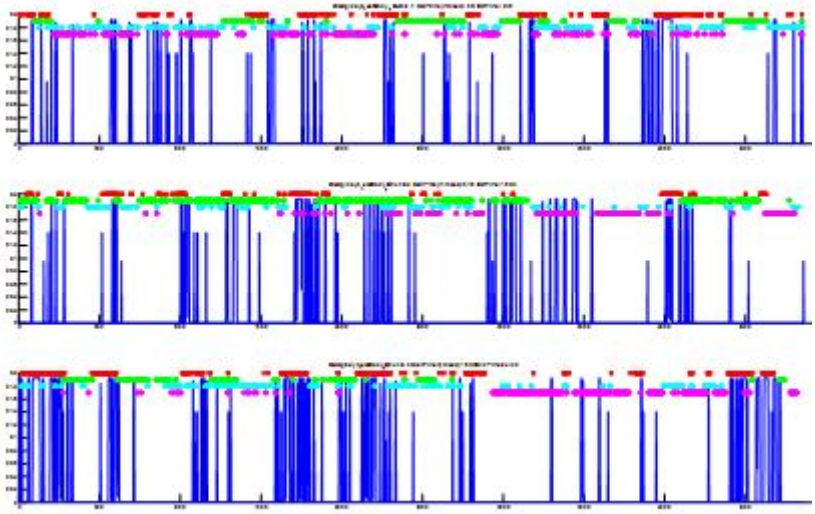

Figure 4. Speaker K's activity against the other participants' speech.

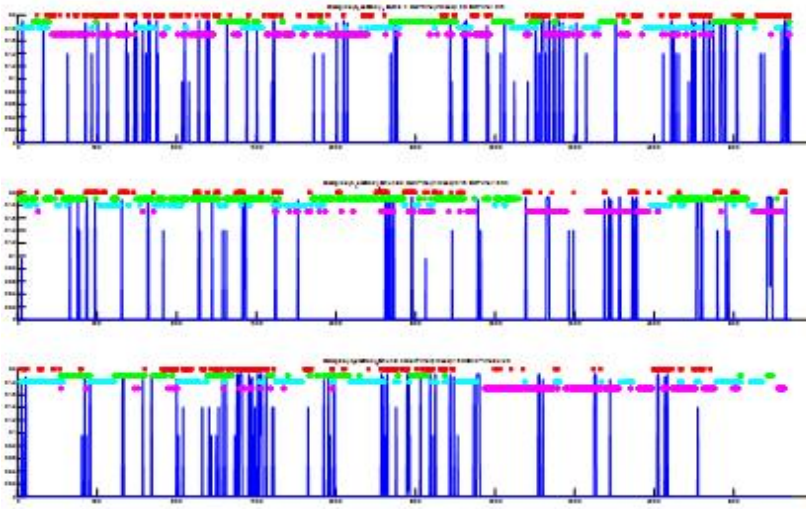

Figure 5. Speaker N's activity against the other participants' speech.

References

Allwood, J. Cerrato, L., Jokinen, K., Navarretta, C., Paggio, P. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. In J.C.Martin, P. Paggio, P. Kuenlein, R. Stiefelhagen, and F. Pianesi (Eds), Multimodal

corpora for modelling human multimodal behaviour. Special issue of the International Journal of Language Resources and Evaluation, 41(3–4), 273–287.

André, E., Pelachaud, C. 2009. Interacting with Embodied Conversational Agents. In Jokinen, K., Cheng, F. (Eds.) *Speech-based Interactive Systems: Theory and Applications*. Springer.

Bennewitz, M., Faber, F., Joho, D., Behnke, S. 2007. Fritz - A Humanoid Communication Robot. Proceedings of the 16th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).

Campbell, N. 2007. On the Use of Nonverbal Speech Sounds in Human Communication. In: Campbell, N., *Verbal and Nonverbal Communication Behaviors*, LNAI, 4775, pp. 117-128.

Campbell, N. 2004. Speech and expression; the value of a longitudinal corpus. The 4th International Conference on Language Resources and Evaluation (LREC), pp. 183-186.

Campbell, N., Ohara, R. 2005. How far can non-verbal information help us follow a conversation? Preliminary experiments with speech-style and gesture tracking. Proceedings of ATR Symposium on the Cross-Modal Processing of Faces & Voices. No laughing matter.

Carletta, J. 2006. Announcing the AMI Meeting Corpus. *The ELRA Newsletter* 11(1), 3–5.

Cassell, J., Sullivan, J, Prevost, S, Churchill, E (Eds.) 2003. *Embodied Conversational Agents*. MIT Press, Cambridge, MA

Clark, H. H., Schaefer, E. F. 1989. Contributing to Discourse. *Cognitive Science*, 13, 259–94.

Douglas, C. E, Campbell N., Cowie R., Roach, P. 2003. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40: 33-60.

Ekman, P. 1972. Universal and cultural differences in facial expression of emotion. In: Cole, J. R. (Ed.) *Nebraska Symposium on Motivation*, Nebraska University Press, Lincoln. pp. 207–283.

Feldman, R. S., Rim, B. 1991. *Fundamentals of Nonverbal Behavior*, Cambridge University Press.

Jokinen, K. 2009a. *Constructive Dialogue Management: Speech Interaction and Rational* Agents. John Wiley and Sons: Chichester.

Jokinen, K. 2009b. Natural Language and Dialogue Interfaces. In: Stephanidis, C. (Ed.) T*he Universal Access Handbook*. Chapter 31. Cenveo. pp. 495-506.

Jokinen, K. 2008. Non-verbal Feedback in Interactions. In: Tao, J.H., Tan T.N. (Eds.) A*ffective Information Processing*, Science+Business Media LLC, Springer, London. pp. 227-240.

Jokinen, K., Campbell, N. 2008. Non-verbal Information Sources for Constructive Dialogue Management. LREC-2008. Marrakech, Marocco.

Jokinen, K., Vanhasalo, M. 2009. Stand-up gestures – Annotation for Communication Management. *Proceedings of the Multimodal Workshop at Nodalida Conference*, Denmark.

Kendon, A. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., Den, Y. 1998. An analysis of turn taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs. *Language and Speech* 41(3-4), pp. 295-321.

McNeill, D. 2005. *Gesture and Thought*. University of Chicago Press, Chicago and London.

Norros, L., Kaasinen, E., Plomp, J., Rämä, P. 2003. Human-Technology Interaction Research and Design. VTT Roadmap. VTT Industrial Systems, VTT Research Notes 2220. Espoo. Also available: http://www.vtt.fi/inf/pdf/tiedotteet/2003/T2220.pdf.

Traum, D. 1999. Computational models of grounding in collaborative systems. In Working Papers of the AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems, AAAI, Menlo Park, CA., pp. 124–131.