

Gothenburg Papers in Theoretical Linguistics 92

# MULTIMODAL COMMUNICATION 2005



## PROCEEDINGS

from the

## Second Nordic Conference on Multimodal Communication

Goteborg 7-8 April, 2005

Edited by

Jens Allwood, Beatriz Dorriots  
& Shirley Nicholson



Department of Linguistics,  
SSKKII  
Göteborg University, Sweden



Gothenburg Papers in Theoretical Linguistics 92

Papers from The Second Nordic  
Conference on Multimodal Communication

2005



Department of Linguistics  
SSKKII  
Göteborg University, Sweden

## **GOTHENBURG PAPERS IN THEORETICAL LINGUISTICS**

Papers in GPTL cover general and theoretical topics in linguistics. The series contains papers both from single individuals and project groups. It appears irregularly in two subseries, blue for English papers and green for Swedish papers.

Jens Allwood  
Editor, GPTL

© 2006

Department of Linguistics, Göteborg University.  
ISSN 0349-1021

## **PREFACE**

This volume contains papers from the Second Nordic Conference on Multimodal Communication which was held in Göteborg, April 7-8, 2005.

In the volume there are 21 contributions to the conference covering a wide variety of topics related to multimodal communication such as normal adult face-to-face communication, children's multimodal communication, intercultural communication and attempts to construct devices that can support or carry out multimodal communication. As a whole the volume gives a good picture of the wide variety of approaches and interests that are present in Nordic research on multimodal communication.

Göteborg, September 2005

Jens Allwood



## **PLENARY SPEAKERS**

Dominic W. Massaro.....	1
Embodied Conversational Agents with Realistic Speech and Language	
Isabella Poggi.....	5
Social Influence Through Gesture and Face	
<hr/>	
Jonna Ahti.....	31
Between “Virtual” And “Real”: Multimodality In Finland-Swedish Chat Conversations	
Jens Allwood, Elisabeth Ahlsén, Johan Lund and Johanna Sundqvist.....	43
Multimodality in Own Communication Management	
Jens Allwood and Nataliya Berbyuk.....	65
Word-finding Problems in Medical Consultations between Non-Swedish Physicians and Swedish Patients	
Jens Allwood, Loredana Cerrato, Kriistina Jokinen, Patrizia Paggio & Costanza Navarretta .....	91
The MUMIN Annotation Scheme for Feedback, Turn Management and Sequencing	
R. Atladottir, J. Gay, K.L. Jensen, R.B. Jensen, I. Kun, L.B. Larsen, S. Larsen Multi Modal Interaction in an Automatic Pool Trainer .....	111
Tom Brøndsted.....	125
The Philosophy behind a (Danish) Voice-Controlled Interface to Internet Browsing for Motor-Handicapped	
Loredana Cerrato.....	137
Linguistic Functions of Head Nods	
Loredana Cerrato and Gunilla Svanfeldt.....	153
A Method for the Detection of Communicative Head Nods in Expressive Speech	
Fang Chen.....	167
Designing Multimodal Communication System for Firefighters	

Pierre Gander .....	179
Gesture and Speech Manifestations of Perspective on Memory of Events with Varying Degree of Participation	
Mia Heikkilä .....	191
Modal Differences in Children’s Communication	
David House .....	201
On the Interaction of Audio and Visual Cues to Friendliness in Interrogative Prosody	
Sari Karjalainen .....	215
Achieving Topic by Multimodality in Early Dyadic Conversation	
Knut Kvale, Narada Warakagoda, and Marthin Kristiansen.....	227
Evaluation of a Mobile Multimodal Service for Disabled Users	
Ann-Christin Månsson .....	241
Word-Finding Ability and Body Communication	
Gunilla Svanfeldt, Preben Wik & Mikael Nordenberg.....	257
Artificial Gaze Perception Experiment of Eye Gaze in Synthetic Faces	
Anna-Lena Rostvall and Tore West .....	273
Multimodality and Designs for Learning	
Karen Woodman .....	283
Multimodality Stimulation: Understanding Fluency in Study Abroad Programs	
Therese Örnberg Berlund .....	303
Multimodality in a Three-Dimensional Voice Chat	

# EMBODIED CONVERSATIONAL AGENTS WITH REALISTIC SPEECH AND LANGUAGE

*Dominic W. Massaro, Ph.D*

University of California, Santa Cruz, U.S.A

## **Abstract**

*Speech and language science and technology evolved under the assumption that speech was a solely auditory event. However, a burgeoning record of research findings reveals that our perception and understanding are influenced by a speaker's face and accompanying gestures, as well as the actual sound of the speech. Perceivers expertly use these multiple sources of information to identify and interpret the language input. Given the value of face-to-face interaction, our persistent goal has been to develop, evaluate, and apply animated agents to produce realistic and accurate speech (Massaro, 1998). Baldi<sup>1</sup> is an accurate three-dimensional animated talking head appropriately aligned with either synthesized or natural speech. Baldi has a realistic tongue and palate, which can be displayed by making his skin transparent.*

*To implement multilingual agents, we have developed a client/server architecture system (Massaro et al., 2005; Ouni et al., 2005). The client is the application controlling Baldi. It sends text from a variety of languages including Arabic, Mandarin, and many European languages as well as English to a general speech synthesis server. The server generates the appropriate phonemes in the appropriate language with all the information needed by the client (phonemes, duration, pitches, word boundaries, etc.) and the acoustic speech waveform, and then it sends them back to the client. Using this information, the client generates the appropriate*

---

<sup>1</sup> Baldi is a registered trademark of Dominic W. Massaro.

*language-specific visible phonemes synchronized with the synthesized speech.*

*Based on this research and technology, we have implemented computer-assisted speech and language tutors for children with language challenges and persons learning a second language. Our language-training program utilizes Baldi as the conversational agent, who guides students through a variety of exercises designed to teach vocabulary and grammar, to improve speech articulation, and to develop linguistic and phonological awareness (Massaro, 2004). Some of the advantages of the Baldi pedagogy and technology include the popularity and effectiveness of computers and embodied conversational agents, the perpetual availability of the program, and individualized instruction. The science and technology of Baldi holds great promise in language learning, dialog, human-machine interaction, education, and edutainment.*

## **References**

- Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, Massachusetts: MIT Press.
- Massaro, D. W. (2004). Symbiotic Value of an Embodied Agent in Language Learning. In R.H. Sprague, Jr.(Ed.), *Proceedings of 37th Annual Hawaii International Conference on System Sciences, (HICCS'04)*(CD-ROM,10 pages), Los Alimitos, CA: IEEE Computer Society Press. Best paper in Emerging Technologies.
- Massaro, D.W., Ouni, S., Cohen, M.M., & Clark, R. (2005). A Multilingual Embodied Conversational Agent. In R.H. Sprague, R.H.(Ed.), *Proceedings of 38th Annual Hawaii International Conference on System Sciences (HICCS'05)* (CD-ROM, 10 pages). Los Alimitos, CA: IEEE Computer Society Press.
- Ouni, S., Cohen, M. M., & Massaro, D. W. (2005). Training Baldi to be multilingual: A case study for an Arabic Badr. *Speech Communication*, 45(2), 115-137. A Multilingual Embodied Conversational Agent. In R.H. Sprague, R.H.(Ed.), *Proceedings of 38th Annual Hawaii International Conference on System Sciences (HICCS'05)* (CD-ROM, 10 pages). Los Alimitos, CA: IEEE Computer Society Press.
- Ouni, S., Cohen, M. M., & Massaro, D. W. (2005). Training Baldi to be multilingual: A case study for an Arabic Badr. *Speech Communication*, 45(2), 115-137.

## **Author's address**

*Daniel Massaro*  
*Animated Speech Corporation*  
<http://www.animatedspeech.com>  
*Perceptual Science Laboratory*  
*Department of Psychology*  
*University of California*  
*Santa Cruz, CA 95060 U.S.A.*  
*phone: 1-831-459-2330*  
*e-mail: massaro@fuzzy.ucsc.edu*  
<http://mambo.ucsc.edu/psl/dwm/>  
*fax: 1-831-459-3519*



# SOCIAL INFLUENCE THROUGH FACE, HANDS, AND BODY

*Isabella Poggi*

University of Rome, Italy

## **Abstract**

*The paper presents a view of persuasion in terms of a goal and belief model of mind and social interaction, and analyses some cases of persuasive multimodal discourse (a real political debate and two films of a famous Italian comic actor) in terms of two procedures: the analysis of discourse as a hierarchy of goals and the "musical score" of multimodal communication. The two procedures prove useful to capture the mechanisms of direct, indirect and multiple communication, and to disentangle, within the structure of multimodal persuasive discourse, the use of the different persuasive strategies discovered by Aristotle: the appeals to *lògos*, *èthos* and *pàthos*.*

**Keywords:** Persuasion, multimodal communication, gestures, gaze, face, body

## **1. Introduction**

Communication is a means to influence others through acting on their mind. Every time we utter a sentence or make a gesture we ask the other to do, to tell us, or to believe something. And this powerful way to act on the world passing through the minds of people resorts to the whole rich repertoire of states and actions of our body: we communicate through voice, face, gesture, gaze, trunk, posture... All of these instruments play together in a complex polyphonic music that conveys meanings to people

and through them induces to action. For many centuries, Rhetorics and Linguistics have in depth explored the ways in which words, sentences, discourses do this job; now, thanks to the modern visual recording technologies a similar endeavour could be undertaken by trying to single out the specific contribution that face, hands and body provide to the work of influencing through communication.

In this work I present a model for the analysis of multimodal communication aimed at specifying the different roles that speech and body communication play in persuasion.

## **2. Social Influence and the People's Goals**

According to a goal and belief model of mind and social interaction a system's life is regulated by goals (Castelfranchi & Parisi, 1980; Conte & Castelfranchi, 1995). A goal is a regulatory state generally represented in a system's mind: when the perceived state of the world is discrepant with the goal, the system performs actions until the goal is achieved.

Often, a single action is not sufficient to achieve a goal, so in order to achieve its goals, a system has to plan and perform more or less complex plans, that is, hierarchies of goals. An action is a means for Goal G1, but Goal G1 can in its turn be a means for a further goal (Supergoal) G2. But often a system does not have all the resources or cannot perform all the actions it needs to achieve its goals. Social exchange and interaction multiply the resources of people and their potential to reach more goals than they could by themselves (Castelfranchi, 1990). This is done through the device of goal adoption – the fact that people put their own resources in the service of other people's goals. A system B adopts the goal of a system A when B pursues A's goal as its own goal, when B “helps” A to achieve A's goal. In order to have B adopt some goal of A's, A may have the goal of influencing B, that is, of raising or lowering the likeliness for B to pursue some goal.

Of course, the goal of influencing others may be either Ego-centred or Alter-centred. Ego-centred influence holds when A has the goal of influencing B to adopt a goal of A's. (for example in a command like "Give me the newspaper please", A wants B to do something for A). In Alter-centred influence A's goal is to influence B to pursue a goal that is an

interest of B. (in an advice like "Put on your coat, it's cold", A wants B to do something that is good for B himself).

There are many different ways to influence people: the use of strength, suggestion, seduction, hypnosis, education, persuasion.

### **3. Persuasion as a case of social influence**

Persuasion, in particular, is a case of communicative and non-coercive way to influence another's goals obtained by influencing his/her beliefs: one leaning on the free choice of the person who is being influenced (Poggi, 2005). In persuading, A has the goal of influencing B, that is, of causing B to pursue a goal proposed by A, by linking the proposed goal to B's previous goals: by making B strongly believe that B is a useful sub-goal to B's actual goals, which B might also have chosen regardless of A's influence.

For example, a politician argues that his party can best recover the economical health of his country in order to make the Electors think that he is the most convenient candidate, and this, in turn, to make them vote for his party. Or, again, the Accuser is ironic about the credibility of the accused, so that the judge thinks it is reasonable not to trust the accused, and this aims in its turn to have the judge condemn the accused.

As Aristotle stated, in every act of persuasion, the persuader can use three different weapons: the rational argumentation (*lògos*), the Speaker's credibility and reliability (*èthos*), and the appeal to the Hearers' emotions (*pàthos*). All three strategies are needed 1) to convince B that the goal GA proposed by A has such a high value that it is worth being preferred to other possible goals, and 2) to convince B that there is a means-end link between GA (the goal proposed by persuader A) and GB (B's previous goals).

The *pàthos* strategy is important because if the proposed goal GA is a means not only for practical goals but also for emotionally loaded goals (e.g., if voting politician A also gives you the emotion of fighting for freedom), its value is raised, and the goal will be more likely pursued. The *ethos* strategy is relevant because if the link between the proposed goal and the previous goals of the B is made credible not only through arguments but also by the reliability of the Persuader, GA is more likely to be pursued.

But how can we disentangle the three strategies of *lògos*, *èthos* and *pàthos* in a persuasive discourse? For a written texts, an accurate analysis of its words and sentences might allow us to precisely find out what, in it, is exploiting each of the three strategies.

#### **4. The analysis of discourse as a hierarchy of goals**

One model of communicative behaviour that allows to do so is one proposed by Parisi & Castelfranchi (1975) that views discourse as a hierarchy of goals. They start from the intuition of Speech Act Theory (Austin, 1962; Searle, 1969), according to which producing a sentence to communicate is to perform an action, and combine it with the notion of the hierarchical structure of action (Miller, Galanter & Pribram, 1960), by concluding that, if a sentence performs an action, and every action has a goal, then also every sentence aims at a goal. So, according to Castelfranchi & Parisi (1980), each sentence has a "literal" goal, which is its literal meaning, explicitly communicated by its lexicon and syntax. But, just as an action may have further goals that are hierarchically super-ordinate to the first, a sentence too may have, beyond its literal goal, one or more super-ordinate goals, called "super-goals", for which the first goal of the sentence is only a means: they are "indirect" meanings of that sentence, and the Speaker, by definition, does not communicate them explicitly through lexical or syntactic features of the sentence, but wants the Hearer to understand them by inference. For example, if my colleague asks me "Are you going home?" the literal goal of his sentence is simply to ask me if I am going home, but his super-goal (the indirect meaning of the sentence, the goal he might want me to infer) may be to ask for a lift. In this view, a discourse has the structure of a plan in that it is a sequence of sentences governed by a hierarchy of goals: for all sentences, the speaker wants the listener not only to understand the single literal meanings, but also to reconstruct inferentially, starting from those literal meanings, a more or less complex hierarchy of super-goals - indirect meanings -, which all aim to communicate the final goal of the whole discourse.

However, as pointed out by Poggi & Magno Caldognetto (1997), also nonverbal behaviours, such as a gesture, a gaze, a picture, may be, if not speech acts, communicative acts. As such, they may have goals and super-goals, and can constitute, in themselves or in combination with verbal communicative acts, a multimodal discourse, that is, a combination of

verbal and nonverbal communicative acts that all aim at a common goal (Poggi & Magno Caldognetto, 1997; Magno Caldognetto & Poggi, 1999). To analyse a discourse in terms of hierarchy of goals you graphically represent it as a hierarchical structure by making explicit the literal meanings of its sentences or other nonverbal communicative acts, but also their indirect meanings - the inferences they induce. After segmenting the discourse into its communicative acts, for each you write down its literal goal and its super-goals, that is, the inferences it aims at inducing, and you link them through arrows representing means-end relationships, finally singling out the final goal of the whole discourse plan and the goals of the intermediate sub-plans.

As far as persuasive discourse is concerned, this analysis allows, for example, to assess how much (for example, in how many communicative acts) the specific discourse analysed appeals to *lògos*, *èthos* and *pàthos*, and the degree of directness in doing so (showing, for example, how many meanings are conveyed literally - appearing in the bottom part of the analysis - , and how many are conveyed only through inference - represented only in the upper levels of the hierarchy).

## **5. Persuasion in political discourse**

In this Section I present the analysis of a verbal fragment of a political discourse in terms of hierarchy of goals; then I will show how some of the persuasive strategies used in a verbal fragment of this discourse are also present in the multimodal communication of another fragment of the same discourse: the strategies used in the two are utterly analogous.

The first fragment I will analyse in terms of hierarchy of goals is taken from a political discourse delivered on an Italian TV Channel in 1994. It is taken from a pre-election debate held between Silvio Berlusconi and Achille Occhetto at the end of March 1994. Berlusconi is the owner of several TV channels and of a billionaire financial business, and in March 1994 he became the leader of a new party, "Forza Italia", confronting the leftist coalition. Achille Occhetto was at that time the leader of PDS, the Left Democratic Party (former Communist party). After this debate, Berlusconi's coalition won the election and he was Prime Minister for 9 months. The debate was held on the political show "Braccio di ferro", broadcasted on Channel 5 (one of Berlusconi's TV channels). These are the background events immediately preceding the debate: just a few days

earlier, the clubs supporting Forza Italia (Berlusconi's party) had been subject to a blitz by the Financial Police, aimed at discovering suspected collusion with the mafia. At that time Berlusconi told a newspaper that a leftist white putsch was being plotted, for election purposes. But then, Berlusconi denied having said that. A similar incident happened with Luciano Violante, an important judge and a member of the Italian Parliament belonging to the PDS party. Violante too said something to the newspapers and then denied it. Here is the verbal fragment, with the part to analyse in italics.

Let me remind you that yesterday Berlusconi asked what face I could wear today, at this meeting, given that I was in charge of a plot against Forza Italia. *Here I am then, wearing this face (S1). That is, the face of an honest man (S2); and when the other day Berlusconi claimed that the left had supposedly organised a putsch (S3), but then he recanted (S4), I didn't say a word (S5), this is the way serious and fair opponents confront each other (S6). I do not understand why one should accept Berlusconi's recantation (S7), but not that of a gentleman such as Violante (S8) who has fought openly (S9), who is willing to take risks, even putting his own life in peril as Violante has done in confronting the mafia (S9).*

Fig. 1 represents the hierarchy of goals of this fragment:

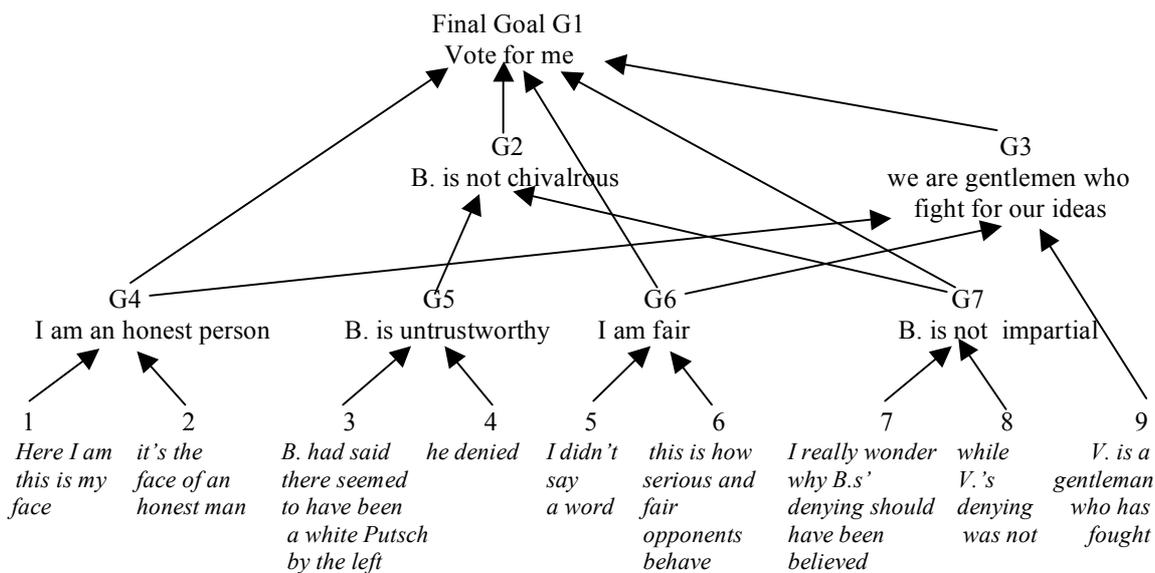


Figure 1. The persuasive strategy of a political discourse

In this fragment, Occhetto says he is presenting himself while *wearing this face* (Fig. 1, Sentence 1), *that is, the face of an honest man* (S2): thus he has the goal of letting people infer (but just infer – he does not say it explicitly) that *he is* an honest man (Goal 4). Then he mentions (S3) that Berlusconi had told (S3) the media of a leftist putsch and then recanted (S4), thus implying (again, just implying) a double-faced, untrustworthy behaviour on his part (G5). After that, he reminds us that he never publicly commented on this (S5), and says this is the way serious and fair opponents confront each other (S6), thus implying that he is a fair opponent (G6). Finally, he rhetorically says he does not understand why Berlusconi's recantation should be taken seriously (S7), while Violante's recantation should not (S8) – thus insinuating that Berlusconi is not impartial (G7).

Obviously, the super-goals G5 and G7, concerning Berlusconi's untrustworthiness and partiality aim at the further goal G2 of showing that Berlusconi is not fair, while the super-goals G4 and G6, along with sentence 9, aim at showing Occhetto's and his party's honesty and fairness. Since the final goal of any pre-election discourse is by definition to persuade electors to vote for the speaker, we can say that here the goal proposed by Occhetto (G1, vote for me), that is, the GA he asks electors to adopt, is pursued by soliciting evaluations from the electors: negative evaluations about his opponent (G2) and positive evaluations for himself and his own party (G3). In sum, therefore, the strategy used by Occhetto in this excerpt does not make an appeal so much to *lògos*, the argumentation about his own political program and about why it would be convenient for the electors to vote for him: it is more of an appeal to *èthos*, that is, the Self-presentation of the Speaker, his moral rectitude, and the political and moral qualities of the people in his party. Thus Occhetto pursues an *èthos* strategy, aimed at raising the electors' trust in himself and his party, in five out of nine sentences, but in an indirect way, because no sentence of his discourse explicitly mentions his reliability.

But do the same features – salience of *èthos* and indirectness – hold also in the other modalities of this persuasive discourse?

To answer this question, we must find a way to analyse multimodal communication in all its aspects.

## 6. The “musical score” of multimodal communication

The tool for analysing multimodal communication I use is the “musical score”, a procedure for the transcription, analysis and classification of multimodal communication (Magno Caldognetto & Poggi, 1994; Poggi & Magno Caldognetto, 1996). In the “score” (now also implemented in a software, the Anvil-score, see Magno Caldognetto et al., 2004), signals delivered in five different modalities are reported on parallel lines:

- v. **verbal** modality (the words and sentences uttered);
- p. **prosodic-intonational** modality (speech rhythm, pauses, intensity, stress, intonation);
- g. **gestural** modality (movements of the hands, arms and shoulders);
- f. **facial** modality (head and eye movements, gaze, smile and other facial expressions);
- b. **body** modality (trunk and leg movements, body posture, orientation and movements in space).

For each communicative item in each modality five levels of analysis can be provided:

- SD. signal description:** the acoustic or visual features of the movement, gesture, gaze, or vocal element at issue are described. For the prosodic signal, length, intensity, fundamental frequency, pauses; for gestural, facial, body signals, handshape and hand movement, facial muscles at work, eye direction, posture and so on.
- ST. signal type:** any item may be classified according to some typology proposed in the literature: a symbolic gesture, a deictic gaze....
- MD. meaning description:** a verbal formulation is provided of the meaning attributed to each communicative item. For instance, the gesture of raising the right hand can be glossed as: "just wait, be careful".
- MT. meaning type:** each communicative item is classified in terms of a semantic typology as to the kind of information it conveys: Information on the World (events, entities, their properties and relations), on the Speaker’s Mind (the Speaker’s beliefs, goals, emotions concerning the discourse being delivered), and on the Sender’s Identity (sex, age, ethnicity, personality, claimed identity).
- F. function:** on the basis of the meaning attributed to each item, it is stated what semantic relationship it bears to a parallel verbal item. A signal is **repetitive** if it has the same meaning of the other signal, **additive** when it adds congruent information, **substitutive** if it

conveys information not borne in the other modality, **contradictory** if its meaning is incompatible or contrasting with the other signal, and **independent** if the two signals make part of two independent communicative plans.

While many other tools now exist for the analysis and classification of multimodal discourse (e.g., Kipp, 2001; Ingenhoff & Smith, 2003; Jokinen et al., 2005), the most characterizing feature of this annotation device is the attempt to find out the precise meaning of each item and to find a verbal paraphrase or synonym of it. This is what allows to classify (level of analysis MT) and to state its function (level F). More: since, as mentioned, it is maintained that not only a word or a sentence, but also a gesture, a gaze, a posture may have, beside its literal meaning, further indirect meanings (Poggi & Magno Caldognetto, 1997), the lines for Meaning Description, Meaning Type and Function allow an analysis on two layers: for each item (gaze, gesture, posture...) not only the literal but also the possible indirect meaning is written down, that is, the meaning that is to be inferred from the literal one.

## 7. Multimodality in political discourse

Let us analyse a multimodal fragment of Occhetto's discourse.

*I want to remind that when... it came out the story of Berlusconi's brother... I was at "Red and Black"; they often asked me to take a stand, may be a polemic one. I did not, I told them I was not a vulture.*

While still arguing about his chivalrous behaviour, Occhetto reminds that when Berlusconi's brother was accused of a dirty financial management, he was being interviewed at "Red and Black", (a political talkshow), where he could well have blown a heavy attack to him; but he did not, he said he was not a vulture; thus again globally implying an outstanding difference between Berlusconi's and his own moral behavior.

Once more, a global analysis of this fragment (Fig.2) shows us a heavily evaluative import, that is, one strongly leaning on an *èthos* strategy. (The first and second score include, respectively, prosodic and gestural, and prosodic, gestural and facial communication, and skip the level of analysis of Signal Description).

v. *Voglio ricordare che quando è venuta fuori la vicenda del fratello di Berlusconi io ero a rosso e nero*  
 I want to remind you that when the story came out of Berlusconi's brother I was at "Red and Black"

p.SD I *raising int. pause* *raising int. pause* *raising int.*

MD I important I'm looking for words important  
 II important important different but connected to B.

MT I ISM ISM ISM  
 II ISM IW

F I Add. Add. Add.  
 II Add. Add.

g.SD *thumb and index ring ring ring* *index f. points at right*  
*up and down*

MD I I state precisely I state I state Berlusconi  
 II I accuse him

MT I ISM ISM ISM IW  
 II ISM ISM

F I Rep.  
 II Add.

v.			
<i>mi s'e' più volte chiesto di prendere una posizione, magari polemica. Io non l'ho presa, ho detto che non facevo l'avvoltoio</i>			
they often asked me		to take a stand,	may be a polemic one.
I did not,		I said	I was not a vulture
p.SD	I	<i>raising int.</i>	<i>raising int. raising int.</i>
MD	I	they insisted	contrast
	II	I resisted	contrast
			I am not this like I am not what they thought
MT	I	IW	ISM
	II	ISI	ISM
			ISI
F	I	Rep.	Add.
	II	Add.	Add.
			Add.
g.SD			
I		<i>open hand back &amp; forth</i>	<i>moves hand forward</i>
II		<i>moves hand forward</i>	<i>moves hand forward</i>
MD	I	argue	refuse
	II	I am not polemic	I am not unfair
			I am not unfair
MT	I	IW	ISM
	II	ISI	ISM
			ISI
F	I	Rep.	Add.
	II	Add.	Add.
			Add.
f.SD			
I		<i>lowers face</i>	<i>lower lip corners</i>
II		<i>lowers face</i>	<i>lower lip corners</i>
MD	I	resolute person	I am disgusted
	II	argue	I am noble
MT	I	IW	ISM
	II	IW	ISM
			ISI
F	I	Add.	Add.
	II	Rep.	Add.
			Add.

Legenda:

v.=verbal modality; p-i.= prosodic-intonational m.; g.= gestural m.; f.=facial m.; b.=bodily m.; SD=Signal Description; ST=Signal Type; MD=Meaning Description; MT=Meaning Type; F=Function; IW=Information on the World; ISM=Information on the Sender's Mind; lh.=left hand; rh.=right hand; I = literal layer; II = indirect layer.

*Figure 2. The two-layers score of Occhetto multimodal communication*

In the verbal modality Occhetto is saying: "*Voglio ricordare che quando... e' venuta fuori la vicenda... del fratello di Berlusconi...*" ("I want to remind you that when... it came out the story... of Berlusconi's brother").

In the prosodic modality, he utters the word "*quando*" ("when") with a raising intonation, and then pauses. These are two different ways to emphasize (to tell it is important) what is being said: with the former (raising intonation) the Speaker gives information on his mind (ISM), quite explicitly remarking "what I am saying now is important"; with the latter, the pause, we can find two layers of meaning. First, pausing means "I am looking for the right words"; but looking for the right words may imply that what one is going to say is particularly important; so, both "looking for right word" and "important" are Information on the Speaker's Mind, but the 2nd layer meaning of the pause is equivalent to the 1st layer meaning of the raising intonation.

Again, Occhetto raises intonation while uttering the word "*fratello*" (brother), which literally means, once more, "it is important what I am talking about now": but an inference required about this phonological emphasis, and then an implied meaning, is: "This is important because it is connected to Berlusconi". At the same time, in the gestural modality Occhetto points at Berlusconi, which means, literally, "I am referring to Silvio Berlusconi", a simply deictic gesture; but it may indirectly be an accusing index finger, a performative gesture of throwing in Berlusconi's face that, while Occhetto was fair at that time, Berlusconi is not now.

Here is where the score captures the difference between literal and indirect meaning: the two meanings on the two layers can be classified differently as to both meaning typology and function. The simply deictic meaning of the gesture is information on the World (IW), with a repetitive function with respect to the concomitant word (Occhetto utters "*Berlusconi*" and at the same time points at Berlusconi); but on the other side, the second layer meaning of accusation is information on the Speaker's mind (ISM), conveying the performative of Occhetto's communicative act ("I accuse, I throw in your face"), with an additive function against the verbal signal: the idea of throwing on Berlusconi's face is only conveyed by gesture, not by words.

While saying: "*They often asked me to take a stand, may be a polemic one*", the raising intonation over "*polemica*" ("polemic") has a first

meaning of a contrastive emphasis, which might imply, on a second layer: "this is different from how I am, I am not that sort of person". If this interpretation is plausible, while the first layer meaning is Information on the Speaker's Mind (ISM), the second layer is an Information on the Speaker's Identity (ISI; here, more specifically, an act of Self Presentation, then communication about the Speaker's *èthos*). The same holds for the raising intonation over "l'ho" in "*Io non l'ho presa*" ("I did not" (take a stand)).

In the gestural modality, parallel to the word "*polemica*", Occhetto makes an iconic gesture with his right hand open, palm to speaker, moving back and forth as if miming two confronting opponents. It means something like "arguing". But the tension of the gesture movement, giving an annoyed impression of violence, might imply disapproval and hence again a Self Presentation information like: "this is not the sort of things I do", "I am not polemic". Then he says: "*Io non l'ho presa. Ho detto che non facevo l'avvoltoio*" ("I did not (take a polemic stand): I said I was not a vulture"). At the same time, he makes an iconic gesture twice. With the same hand configuration as before, moving forward he acts like removing something from himself. This has a performative meaning of "refusing" (an ISM meaning), but this also may have, at the indirect level, an implication of Self Presentation, "I am not unfair". In fact, at the same time in the facial modality he lowers lip corner, an expression of "disgust". But this literal meaning (layer I) "I feel disgust (about this)" again implies a Self Presentation information to be inferred (layer II): "I am not that sort of people, I am a noble person".

We analysed two fragments of Occhetto's persuasive discourse with two different methods of analysis: 1. the representation of the hierarchy of goals underlying a the verbal fragment and 2. the "musical score" of a multimodal fragment. Both methods remark the distinction between literal and indirect meanings, and from both analyses, a common feature emerges underlying the two fragments. While the literal meaning usually regards neutral (non-evaluative) information (Information on the World or on the Speaker's mind), it is just the indirect meaning that is more frequently loaded with evaluative information. From our hierarchy of goals analysis it results that most of Occhetto's sentences, while literally conveying factual or neutral information, indirectly convey negative evaluation of his opponent or positive evaluation of himself and his party. In the "musical score" representation we have seen that while literal meanings most frequently consist of Information on the World or Information on the

Speaker's Mind, the indirect meanings of prosodic, gestural and facial communication most frequently convey Information on the Speaker's Identity: more precisely, Information aimed at the Speaker's Self Presentation, at eliciting positive evaluation about the Speaker. In both fragments, therefore, the literal meanings of signals try to persuade mainly by pursuing a *lògos* strategy; but the indirect meanings, not only of sentences but also of voice, gesture and face, pursue an *èthos* strategy.

## 8. Totò's score

So far we have seen how, to analyse all the nuances of persuasive discourse, both for the verbal signals and for the other modalities one has to take into account not only the literal meaning of each word or sentence, but also the indirect meanings; and we have shown that the "score" of multimodal communication can just be exploited to annotate the literal and indirect meanings of all signals. Let us see one more example of this in another type of persuasive discourse: a complex pantomime of apology and justification in a comic movie by Totò. Totò (Antonio De Curtis) was a very famous Italian comic actor: he was from Naples, and he superimposed all the Neapolitan expressiveness to his playing, as well as all the creativity of an actor who was always improvising on the stage. In fact, he was not simply a comic actor, but a great actor, who only rarely was valued at his best by directors in the theatre and the movies.

One could object that it is not worthwhile to analyse fiction materials: after all, we are interested in how real everyday human communication works. But in my view, examining a (good) actor's art is a useful occasion to test the analytical power of the "score" and of a hierarchy of goals account: only through sophisticated tools like these, in fact, can you account for the complex and multi-level intertwining and the rich possibilities of human expressivity.

I will first analyse a fragment from the film "*Totò a colori*" ("Totò in technicolor") and, particularly, a scene of what we would now call "sexual harassment".

Totò has just tried to court the maiden in his sister's home, but on her refusal he feels he needs to offer an apology and a justification. He says: "*You are right, you are right, I apologise so much!*", and at the same time he performs a complex pantomime of apology and

justification: he beats his fists on his breast, like in the ritual catholic gesture for "*mea culpa*" ("I am guilty, it's my fault!"); at the same time he lowers his gaze thus expressing shame, he pulls the inner parts of his eyebrows up and closer expressing helpless sadness, and therefore a request for forgiveness. Then he adds: "*I let myself be carried by the impetus, by the impulse of flesh!*", and at the same time he closes his fists with a strong muscular tension, while also frowning with his eyebrows: with these movements he iconically expresses how he is striving to contrast this strong impulse of the flesh. Then he closes and lowers his arms with still closed fists, he relaxes his eyebrows and pulls his chin up, somehow meaning that he suddenly recovered his calmness and self-confidence. Gesture and facial expression do not convey shame any more, but pride: he (as a typical Italian "macho" of the fifties) is proud of having let himself be carried by the impulse of the flesh! "*By this bloody bad flesh!*" he adds, by rising wrists with palms up; then he points at his own body like if accusing it, while, with his head bent, gaze down, eyebrows frowning and a nasty grin, he expresses disgust and hatred to himself. Finally he says: "*Now I'll hurt myself! I don't know what I would like to do to myself! Now I'm going to blind me an eye!*", and he starts slapping himself with both hands, he scratches himself, he punches his own temples, he grasps and wrenches his nose, he puts his finger into his eye, while at the same time producing grins of pain and of anger.

This is the verbal description of Totò's acting in this scene. Such a performance can be accurately analysed through a two-layers score: each communicative signal of each modality considered goes through two levels of analysis, where it is attributed two different (sometimes, possibly, very different) meanings and meaning types (see Figure 3).

v. SD

*hai ragione, hai ragione, scusa tanto! mi sono lasciato trasportare dall'impeto, dall'impulso della carne*  
 You're right, you're right, I apologise! I let myself be carried by impetus, by the impulse of the flesh!

g. SD *hands before trunk opens and raises hs. closed fists in tension extends forearm with closed fists palms forward*

M I	I leave you alone	I strive hard	I'm relaxed
II	I apologise	impetus	I'm serene
MT I	IW	ISM	ISM
II	ISM	IW	ISM
F I	Add.	Add.	Add.
II	Rep.	Rep.	Add.

f. SD *looks down, inner brows up frowns raises chin*

M I	I am sad	I strive hard	I am proud
II	I apologise	impetus	
MT I	ISM	ISM	ISM
II	ISM	IW	
F I	Add.	Add.	Add.
II	Rep.	Rep.	

Legenda:

v.=verbal modality; p-i.= prosodic-intonational m.; g.= gestural m.; f.=facial m.; b.=bodily m.; SD=Signal Description; ST=Signal Type; MD=Meaning Description; MT=Meaning Type; F=Function; IW=Information on the World; ISM=Information on the Sender's Mind; lh.=left hand; rh.=right hand; I = literal layer; II = indirect layer

Figure 3. The score of a multimodal apology and justification

While saying: "*Hai ragione, hai ragione, scusa tanto!*" ("You're right, you're right, I apologise so much!"), Totò raises his open hands, palms forward, before his breast. The literal meaning of this gesture is "I leave you alone", "I don't want to hurt you", but the indirect meaning is "I apologise". The former provides Information on the World (IW) and has an additive function; the latter provides Information on the Sender's Mind (ISM) and has a repetitive function, because it tell the same thing as the parallel verbal signal, "I apologise". Also in analysing his gaze looking down we can find differences between the first and the second layer of meaning: the literal meaning is "I am sad" (which is additive with respect to speech), but since showing sorry is semantically a part of the act of apologising (Poggi, 1994; 1997), its indirect meaning is again one of

apology, and then ones more a repetition of his verbally apologising. Next, when Totò says: "*Mi sono lasciato trasportare dall'impeto*" ("I let myself be carried by the impetus"), he closes and squeezes his fists strongly. The literal meaning is "effort", "strive hard": information on a physical sensation of the Sender (hence, an ISM) which is additive with respect to speech; but the indirect meaning is "impetus", since a "hard strive" is metonymically linked to the notion of "impetus": impetus is something you have to strive hard to oppose it. This indirect meaning concerns Information on the World and has a repetitive function with respect to the word "*impeto*" ("impetus").

From this example one can see that according to which one it is the layer of meaning we analyse, the same signal can be classified in different ways as for both its meaning type and its function; and since the meanings conveyed in the two layers are not always the same, also the relationships of congruence, addition or contrast among the different modalities in a discourse cannot be accurately stated if we take into account only one layer of meaning. As it is clear from this two-layers score, what looks as an addition to speech is instead often a case of repetition, since it communicates, sometimes through indirect meaning, something that in speech is expressed literally.

## 9. Totò's hierarchy of goals

Totò's multimodal discourse in this fragment has the final goal of providing an apology and a justification. But this goal is pursued, as we have seen, through a quite complex combination of signals in different modalities: speech, gestures, facial expression. So far we have analysed them only in their temporal sequence; but their sequential and simultaneous combination is the ultimate result of a complex communicative plan. Also for a multimodal fragment, then, we can represent its hierarchy of goals, which encompasses verbal and body behaviour together with their respective meanings and their hierarchical discourse relationships. Figure 4 represents the hierarchy underlying the fragment, whose verbal text is the following:

"You are right, you are right, I apologise so much! I let myself be carried by the impetus, by the impulse of flesh! By this bloody bad flesh! Now I'll hurt myself! I don't know what I would like to do to myself! Now I'm going to blind me an eye!"

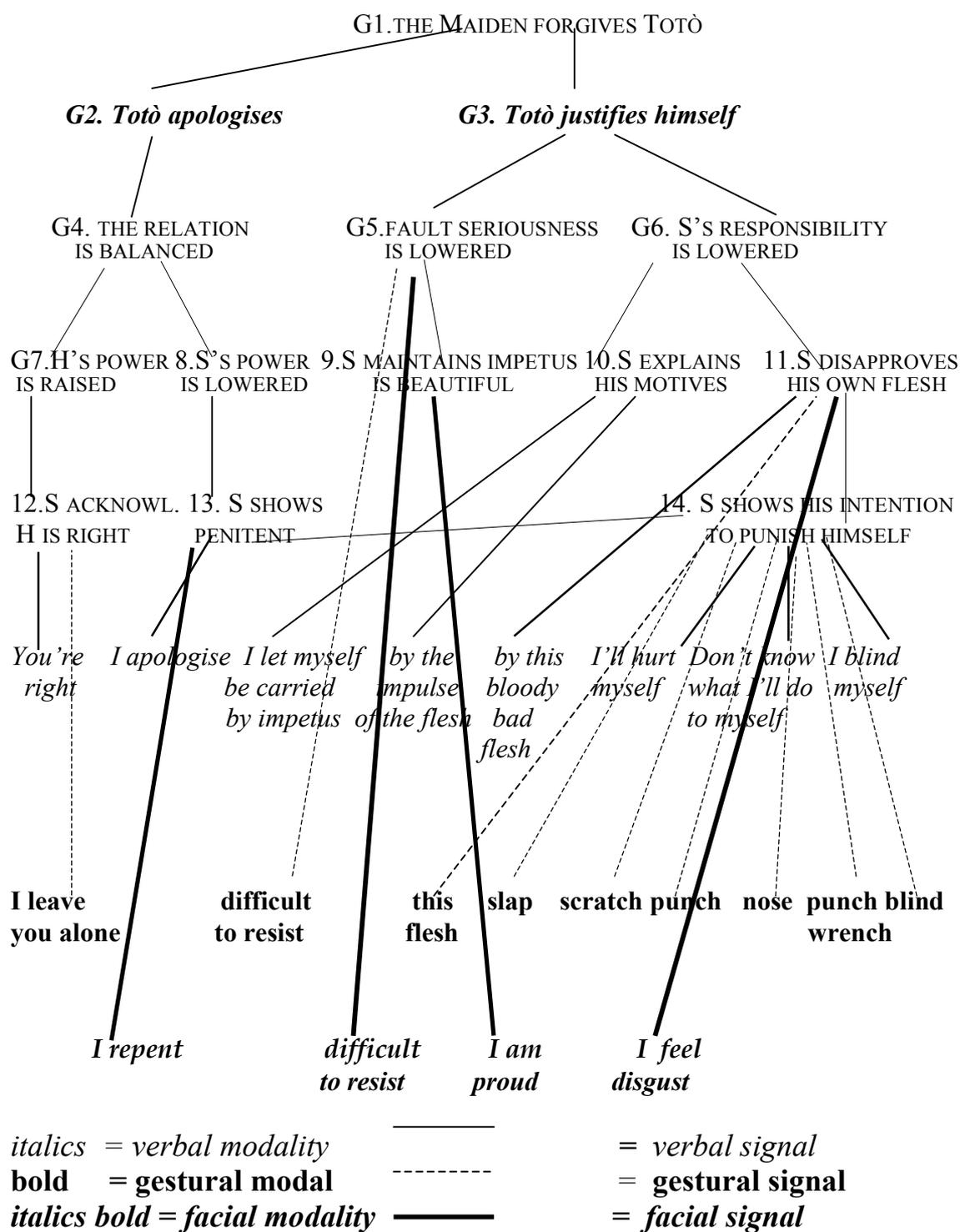


Figure 4. Totò's hierarchy of goals

The final goal of Totò's plan (G1) is to be forgiven by the maiden for his tentative sexual harassment. In order to this, on the one side he apologises (G2), on the other side he provides a justification (G3). When we provide an apology or a justification (Poggi, 1997) we try to re-balance our relationship with the other, that was challenged by some damage we caused to him, in two ways: on the one side (G7) we attribute him more power, on the other side we lower our own power, by showing we submit to him (G13 and then G8). The former is done by Totò by saying "*You're right, you're right*", the latter by the words "*I apologise so much*", the gesture of *raising hands* (= "I leave you alone"), the *sad and repenting facial expression* and the *ashamed posture* ("I repent") (Fig. 5a). After apologising, though, Totò also feels he needs to provide a justification. To provide a justification we can use two different strategies (Poggi 1994): arguing that the damage caused to the victim is less than one could think (G5), and arguing that we are less responsible in it than may be thought (G6): that is, both by extenuating the damage caused and extenuating our own responsibility in it. Totò mainly uses the second strategy, by arguing (G10) that it is not he who is guilty of his indecent proposal, but his impetus, his "*impulse of the flesh*", his "*bloody bad flesh*" (Fig.5c). With the words "*impetus*", "*impulse of the flesh*", the gesture of squeezing his fists and the expression of the eyebrows frowning (Fig. 5b) out of his strive (both meaning "difficult to resist"), he shows (G10) that it was not in his power to resist that impulse, and therefore he is not to be held responsible if he did not resist (G6). After that, he almost splits into two persons, in order to discharge all the guilt of the sexual harassment onto another being, one separated from him, who is the cause of everything and that he disapproves (G11). And in order to make this split more credible he shows (G14) he wants to punish this other self by inflicting him a number of pains: slaps (Fig.5d), scratches (Fig.5e), punches (Fig.5f), nose wrenchings (Fig.5g). By this pantomime of self-punishment Totò wants to demonstrate how he disapproves his own flesh, how he wants to dissociate and take distance from it (G11). And he makes things worse with the words "*by this bloody bad flesh*", and with his expression of disgust (Fig.5g). Therefore he chooses, for his justification, the strategy of lowering his responsibility (G6).

However, there is a move in his multimodal communication that does not look congruent with this strategy. While naming the impulse of the flesh, Totò raises his chin in a proud expression, as if saying "But I am proud of this impulse!". In the straight no more ashamed posture, in the relaxed

gesture and in the proud expression, Totò reminds a Defender while haranguing the Court by arguing for the noble motivation of the Accused's behaviour.

This communication could be interpreted in two different ways. On the one side, Totò could be using also the other justification strategy, by lowering the severity of the damage (G5): sexual harassment is ugly, but impulse of the flesh is beautiful (G9). Further, one more possibility is that Totò, in only this frame, does not want to acknowledge his guilt: what others can contempt, he, as an Italian macho, is proud of. In this hypothesis we should once more think of a split in his communication. If below he splits into a victim and an executioner, here there could be, instead, a third person, the actor, who blinks at the audience by showing ironic on the whole scene of apology and expiation. If this interpretation holds, then here we have one more plan of Totò's communication: here he is not communicating to the maiden, the other character of his movie, but to the movie audience. On the screen, in fact, we by definition have a case of "multiple" communication: the Sender is addressing both to the other characters and to the spectator.



a. "apology"



b. "flesh impulse"



c. "bloody bad flesh"



d. "slap"



e. "scratch"



f. "punch"



g. "nose wrench"

*Figure 5. A pantomime of apology and justification*

## 10. Persuading through spatial behaviour

Let us see another fragment of Totò's playing, this time not a comic film. In "Guardie e Ladri" ("*Cops and Thieves*"), even a small fragment of multimodal communication can provide us with the key to interpret the whole film, its persuasive goal.

A cop (the actor Aldo Fabrizi) has been pursuing a thief (Totò) for long time, without succeeding; if he catches him, it will be bad times for the thief and his family, but if he does not, he himself will be fired. In the last scene, Fabrizi finally stops Totò on the stairs of his house. Totò understands he cannot escape, but first he tries to convince him to let him go, by criticising his persistence and by minimising his own dangerousness: "*Ma si puo' sapere perche' ce l'hai tanto con me? Nemmeno fossi un brigante di quelli che vanno ad assaltare le banche... Ma non lo vedi che sono un poveraccio? Ma che t'ho fatto?*" "But why are you so nasty with me? I'm not a robber who goes around robbing banks... Can't you see I'm a poor man? What did I do to you?". The cop answers: "*A me niente, ma è agli altri che fai danno*". "You do not do anything to me, it's to others that you do damage". Again, Totò ridicules him by extenuating his own responsibility, and tries to induce him to compassion: "*Oh! Capirai che danno! E poi brigadiere mio lo debbo fare per forza. Cosa crede lei, che sia una cosa facile mandare avanti una famiglia? Fargli trovare la minestra tutti i giorni, mandare i bambini a scuola (...) Comprare le medicine quando stanno male, pagare i vestiti!*" "Oh! What a big damage! And moreover, cop, I must do it. Do you think it's easy to feed a family? To let them have supper every day, send kids to school (...) buy medicine when they are sick, pay for suits!".

Also in this case Totò is providing justifications: he is shifting from a "lower damage" to a "lower responsibility" strategy; and a parallel shift holds in his spatial behaviour. At first he is on the top of the stairs looking down to the cop; his intonation is one of someone who feels he has the right to judge the other: you're the cop I'm the thief, but I am a man like you, and as such I can judge you, and disprove of your persecution against a "poor man". But then he switches from the ridicule to the compassion strategy: he steps down twice, as if he wanted to be on the same level as the other; then he steps down again, he sits down on a step and looks up at the cop, finally imploring. Thus, Totò's spatial behaviour is pursuing a *pàthos* strategy, by attempting to induce different emotions in Fabrizi: respect when he stares at him from upstairs, then sympathy when he stops at the

same level, and finally compassion, when he sits down on the step and looks at him from bottom up.

This is what he communicates, in his role of actor, to the other character. But at the same time, he is also communicating to the movie audience, and his spatial behaviour also provides the key to the whole film. One of the film main goals, its moral goal, is to sanction the “war between poor people”, the conflict between two members of contrasting social categories, whose common feature is instead that they are all men and all oppressed by the social system that maliciously makes them struggle against each other. Actually, this thesis is just expressed by Totò’s behaviour from the top of the stairs: at the moment he stares down to the cop he feels superior, and able to judge him; he refuses that they are opposed in contrasting social roles of a cop and a thief, he tells they are simply two persons that can respect and judge each other.

So, once more, the hierarchy of goals and the complex arrangement of Totò's multimodal communication in this film is, as well as for fiction in general, a case of "multiple" communication: on the one side it aims at the character's final goals, on the other side to the film goals.

But this complex intertwining of multimodal signals, and of the goals they pursue singularly and in their combination, can better emerge thanks to annotation systems specific and sophisticated enough as to capture all the richness of human communication.

## **11. Conclusion**

In everyday interaction we influence each other in complex and sophisticated ways, and we do so by exploiting all the richness and subtlety of our body communication. A model that is capable to account for human interaction and communication, and to simulate them in Artificial Agents, should be sophisticated enough to capture and manage all these subtleties.

In this paper I have presented a model of persuasive social influence, and tried to show how persuasion is carried out not only through words, but by all communicative signals produced by our body. I have then presented two ways to analyse the complex intertwining of voice, hands, face and body signals implied in the task of persuading, the analysis through the "musical score" of multimodal communication and the analysis of verbal and multimodal discourse in terms of hierarchy of goals, while showing how

subtle nuances of human interaction can be highlighted thanks to these tools.

## References

- Austin, J. L. (1962). *How to do things with words*. Oxford: Oxford University Press.
- Castelfranchi, C., & Parisi, D. (1980). *Linguaggio, conoscenze e scopi*. Bologna: Il Mulino.
- Castelfranchi, C. (1990). Social Power: a missed point in DAI, MA and HCI. In Y. Demazeau & J. P. Mueller (Eds.), *Decentralized AI*. North-Holland: Elsevier, 49-62.
- Conte, R., & Castelfranchi, C. (1995). *Cognitive and social action*. London: University College.
- Ingenhoff, D., & Schmitz, H. W. (2003). Com-Trans: A multimedia tool for scientific transcriptions and Analysis of communication. . In M. Rector, I. Poggi & N. Trigo (Eds.), *Gestures. Meaning and Use*. Porto: Edicoes Universidade Fernando Pessoa.
- Jokinen, K., Allwood, J., Cerrato, L., Paggio, P., & Navarretta, C. (2005). "The MUMIN Annotation Scheme for Feedback, Turn Management and Sequencing". Proceedings of the *Second Nordic Conference on Multimodal Communication*. Goeteborg, 7-8 April, 2005.
- Kipp, M. (2001). From Human gesture to synthetic action. In C. Pelachaud & I. Poggi (Eds.), *Multimodal Communication and Context in Embodied Agents*. Proceedings of the Workshop W7 at the 5<sup>th</sup> International Conference on Autonomous Agents, Montreal, Canada, 29 May 2001, pp.9-14.
- Magno Caldognetto, E., & Poggi I. (1994). Il sistema prosodico intonativo e l'analisi multimodale del parlato. In P.L.Salza (a cura di), *Gli aspetti prosodici dell'italiano. Atti delle IV Giornate di studio del Gruppo di Fonetica Sperimentale (A.I.A.)*, Roma: Esagrafica, 1994, pp.143-154. Also in E. Magno Caldognetto & I.Poggi (a cura di): *Mani che parlano*. Padova: Unipress, 1997.
- Magno Caldognetto, E., & Poggi, I. (1999). The score of multimodal communication and the goals of political discourse. *Quaderni dell'Istituto di Fonetica e Dialettologia*, Vol.1, 1999 (CD-Rom). Consiglio Nazionale delle Ricerche, Padova.

- Magno Caldognetto, E., Poggi, I., Cosi, P., Cavicchio, F., & Merola, G. (2004). Multimodal Score: an ANVIL Based Annotation Scheme for Multimodal Audio-Video Analysis. Workshop on *Multimodal Corpora LREC 2004*. Centro Cultural de Belem, Lisboa, Portugal, 25th May 2004.
- Miller, M., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior*. New York: Holt, Rinehart and Winston.
- Parisi, D., & Castelfranchi, C. (1975). Discourse as a hierarchy of goals. *Working Papers*, 54-55. Urbino: Centro Internazionale di Semiotica e Linguistica.
- Poggi, I. (1994). Giustificarsi. In C. Castelfranchi, R. D'Amico & I. Poggi (Eds.) *Sensi di colpa*. Firenze: Giunti, pp.180-202.
- Poggi, I. (1997). Scuse, alibi e pretesti. *Sistemi Intelligenti*, IX, 1, 47-65.
- Poggi, I. (2005). The goals of persuasion. *Pragmatics and Cognition*.
- Poggi, I., & Magno Caldognetto, E. (1996). A score for the analysis of gesture in multimodal communication. In L. Messing (Ed.), *Proceedings of the Workshop on the Integration of Gesture in Language and Speech*, Newark, Delaware and Wilmington, Delaware. 7-8 October 1996. Newark, Del. & Wilmington, Del.: Applied Science and Engineering Laboratories, pp. 235-244.
- Poggi, I., & Magno Caldognetto, E. (1997). *Mani che parlano. Gesti e psicologia della comunicazione*. Padova: Unipress.
- Searle, J.R. (1969). *Speech Acts: an essay in the philosophy of language*. Cambridge: Cambridge University Press.

## **Biography**

**Isabella Poggi** teaches General Psychology and Psychology of Communication at Roma Tre University. She has published books and papers, also with C. Castelfranchi, E. Magno Caldognetto and C. Pelachaud, on different fields of communication: Teaching of Italian as a first Language, Pragmatics (interjections, deception, persuasion), Psychology of Emotions (guilt, shame, humiliation, pity, enthusiasm, emotions at work and at school), Multimodal communication (gestures, facial expression, gaze, touch, music), Embodied Agents. One of the founders of the International Society for Gesture Studies, she is in the board of the journals "Gesture", "Journal of Intercultural Communication", and of the "Digital Semiotic Encyclopedia".

## **Author's address**

*Isabella Poggi  
Dipartimento di Scienze dell'Educazione  
Università Roma Tre  
Via del Castro Pretorio 20  
00185 Roma - Italy  
phone: +39 06 49229296  
e-mail: [poggi@uniroma3.it](mailto:poggi@uniroma3.it)  
<http://host.uniroma3.it/docenti/poggi>*



# BETWEEN “VIRTUAL” AND “REAL”: MULTIMODALITY IN FINLAND-SWEDISH CHAT CONVERSATIONS

*Jonna Ahti*

Department of Scandinavian Languages and  
Scandinavian Literature  
University of Helsinki, Finland

## **Abstract**

*Defining multimodality is not an easy task. It may include a variety of different perspectives from embodied agents to the body language of small children. In this paper I will study the specific phenomenon in which the participants of a chat conversation also share the same physical environment at the time of chatting. Typically this means that they are sitting in the same computer class at school. The questions raised are: How can the relationship between “real world” and “virtual world” be described in cases where the participants choose to use computer-mediated communication instead of face-to-face interaction? Can this smooth switch between two media be seen as two modes of the same communicative act? My data shows that the chatters use on-line interaction side by side with face-to-face interaction and that these two ways to communicate do not exclude each other but complement each other. As a result I interpret the situation as two modes of the same communicative act instead of communication in two different media.*

**Keywords:** Multimodality, communication, chat, Finland-Swedish, minority, virtual community.

## **1. Background**

In this paper my aim is not to try to define or redefine multimodality but rather to study a specific phenomenon I have found in my data and discuss the questions it raises. It is a phenomenon in which the participants of a chat conversation (dyad or polyad) also share the same physical environment at the time of chatting. Most typically this means that they are sitting in the same computer class at school. My data includes several such cases and I will use three of them as examples. The questions raised are: How can the relationship between "real world" and "virtual world" be described in cases where the participants choose to use computer-mediated communication instead of face-to-face interaction? Can this smooth switch between two media also be seen as two modes of the same communicative acts?

This study is a part of a larger research project on if and how a specific Finland-Swedish chat-room can be defined as a virtual community. The data in my study consists of 8 days of chat conversations in a chat-room called Extrem Chat. It is administrated by the state-owned youth radio channel Radio Extrem (X3M) that sends out its programs all over Swedish-speaking Finland. Theoretically all Finland-Swedish teenagers have the opportunity to listen to Radio X3M and also be aware of the channel's services on the internet ([www.x3m.yle.fi](http://www.x3m.yle.fi)). Most of these services, like voting for the best song of the week or creating one's own weblog, are only open to registered members of the X3M Community, a virtual community of active listeners. At the time of writing this paper the community had about 7,000 registered members. The members and chatters are mostly teenagers, between 13 and 19 years old, and they come from different areas of Swedish-speaking Finland.

## **2. The Finland-Swedish minority**

Finland has 5 000 000 inhabitants of which 290 000 (5,6%) have Swedish as their mother tongue (Finnäs 2003). They are culturally integrated in the majority so practically the mother tongue is the only factor that separates Finnish and Swedish speaking people from each other. Due to historical reasons, the Swedish speaking minority has a strong position in Finnish society. Swedish is an official language in Finland and all services on governmental and local level should be offered both in Finnish and Swedish. The new Language Act came into force January 1, 2004.

The chief aim of the new act was to ensure that “the cultural and societal needs of the population will be met in either of the two national languages on an equal basis. The public authorities are thus obliged to provide services in Finnish and Swedish, and there should be social welfare services, primary education, other education and comprehensive information in both languages” (Folktinget: Language Act 2004). In everyday life many Finland-Swedes are yet forced to use Finnish when dealing with authorities. The influence of Finnish can be seen as a threat to Finland-Swedish but at the same time it is one of the things that make the situation linguistically interesting.

Because of the official status of the Swedish language it is relatively easy for Finland-Swedes to interact as well as to create and maintain contact within their language group even though the geographical area of Swedish Finland is quite large reaching from North Ostrobothnia along the coastline to East Nyland. Even if Swedish Finland is not a unified territory there are studies that show that the social networks of Finland-Swedish people are stronger and more vivid than those of the Finnish speaking majority (Saarela & Finnäs 2002).

The concept of a virtual community is interesting when it is studied from a minority perspective. The Finland-Swedish minority shares the same language as the 10 million people living in Sweden. Hence there are differences in these two variants of the same language. Finland Swedish has preserved some linguistic features that can be regarded as old-fashioned by Swedish people. There is a variety of Finland-Swedish dialects that have not been presented in a written form before the age of the Internet and other forms of electronic communication. Finland Swedish has been influenced by Finnish and code-switching is not uncommon especially in the varieties spoken in Southern Finland. Code-switching from Swedish to Finnish is also a large area of interest in studies in Finland Swedish (Saari 2003). These features can also be seen in the data I’m studying and I believe they have a specific role in creating a sense of a Finland-Swedish community.

### **3. Multimodality and sociality on the internet**

Like all communication also communication on the internet has several modes. For an ordinary internet surfer they usually appear as follows (Ahtikari & Eronen 2003):

1. Linguistic mode
  - words and phrases
2. Auditive mode
  - speech and music
3. Spatial mode
  - lay-out
4. Gestural mode
  - moving elements
5. Visual mode
  - pictures and colours

A chat-room can also include many of these elements like three-dimensional voice chat Traveler (see Örnberg, this volume). The chat-room I am studying is a "traditional" web chat that does not include speech or music. The chat-room opens up in a pop-up window which makes it possible to browse other web pages while chatting. Pictures and colours can only be seen if the colour of the font changes or if a chatter has put his/her picture on his/her personal information file. There are no moving elements except for the chatters utterances that keep on scrolling whenever there is a new message coming. This means that also the lay-out in the chat-room is not as vivid as on the administrator's main page. This is not necessarily a bad thing, rather it gives space for peaceful chatting. It also keeps one specific feature alive, a feature that I find most fascinating in chat conversations: the chatters can only use the linguistic mode to compensate the lack of other modes.

Even if the different modes can be listed like those above, it is in my opinion impossible to study communicative modes without taking into account the context and other participants. As regards to chat conversations I believe that also the physical world outside the virtual environment plays an important role in how and why people interact on the Internet. The clearest example on this in my data is the fact that the chatters have decided to chat in a Finland-Swedish chat-room. It would be linguistically possible for most of them to chat either in a Finnish chat-room or in a Swedish chat-room. Chatting in a Finland-Swedish chat-room implies that language is not the only criterion for choosing this specific chat-room but that there are social factors that also affect the choice.

Virtual communities and social worlds have usually been studied as autonomous entities that only exist under certain circumstances, like when a group of Star Trek fans create a chat-room where they can discuss events

in the films and tv-series. Special interest groups, like Star Trek fans, often organize real-life gatherings where they can meet each other, wear specific costumes and share their experiences and their knowledge. In addition to such subcultures it is easy to forget that there are also people who do not have any specific visible reference group in which they belong that would somehow authorize them to participate in chatting. For them, like the participants in X3M Chat, chatting might be just another way to keep in contact with real-life friends, an extension of their everyday communication.

#### 4. Analysis of the data

In this section I am going to present three examples of cases where the participant's face-to-face interaction continues or takes place in the virtual environment. For ethical reasons the names of the participants have been changed. The discussion is in Swedish and the translation in English is in italics.

In the first example two girls, Ada and Ida, discuss their next lesson and the day's school lunch. In this case the girls use the virtual environment to discuss things that take place in their contemporary reality.

##### *Example 1*

- 1 **Ada:** hoppas att lärarn komer ida sen  
*I hope the teacher comes today*
- 2 **Ida:** jo fan! men vi har väl Sanna nu o hon brukar ju no koma  
*sure hell! but we have Sanna now and she usually comes*
- 3 **Ada:** e du på väg å äta fisksoppa ida??  
*will you go and eat fish soup today??*
- 4 **Ida:** jo  
*yep*
- 5 **Ada:** bra... ja hänger me..  
*good... I'll come along..*
- 6 **Ida:** ok
- 7 *Ada has left the building*
- 8 *Ida has left the building*

In example 1 Ada and Ida are discussing their next lesson and wondering if their teacher will come (lines 1-2). This is obviously a subject they both

know something about and their way of talking about it reveals that some of their teachers do not arrive for the lesson on time. They also mention the teacher of their next lesson by name (line 2, Sanna) and even make their judgment about her habits by saying that she usually comes on time. This shared knowledge shows that these girls know each other in “real life”, which makes it difficult for other participants on the channel to participate in the discussion if they are not students in the same class.

After discussing their teacher, Ada takes up a new topic, fish soup. Her question about whether Ida is going to eat fish soup today or not shows that this topic, too, is recognizable to both girls and does not lead to any misunderstandings. A free warm meal is served in all Finnish schools everyday and it is a common topic at school. The most interesting feature in this sequence is how the situation is resolved (lines 3-8). In line 5 Ada says that she will go with Ida to eat fish soup and after Ida’s agreement in line 6 they both leave the channel immediately, probably to get their portion of soup in the school’s diner. This example shows that being in the same physical space (classroom) does not automatically mean that communication is face-to-face interaction. In example 1 the chat conversation leads to physical activity when girls leave the channel and go to eat. It shows that a clear distinction between “virtual” and “real” in communicative activity cannot be made. The participants glide smoothly between these two modes and use them to complement each other, not to exclude each other.

Example 2 shows another similar case where the participant’s real-life relationships also take place in a virtual environment. The first lines (1-8) consist of a variety of greetings complemented with the recipients nicknames.

### ***Example 2***

- 1   **Anni:** moi moi Mia  
      *hi hi Mia*
- 2   **Benny:** å heej eva!!  
      *and hello eva!!*
- 3   **Calle:** moi Disa  
      *hi Disa*
- 4   **Disa:** tjenis eva  
      *cheers eva*

- 5 **Calle:** heeeeej maricken!!  
*heeeeello maricken!!*
- 6 **Calle:** moi jenny!!!  
*hi jenny!!!*
- 7 **Jenny:** hej Calle...  
*hello Calle*
- 8 **Eva:** heejjj hopp:)  
*hi hopp:)*
- 10 **Disa:** nu e klassn här igen  
*now the class is here again*
- 11 **Jenny:** jepps.. vi e alla samlade.. nästan.. =)  
*yeah.. everyone's here.. almost =)*
- 12 **Disa:** mn så e d allti  
*but that's the way it always is*
- 13 **Jenny:** jo .. tyvärr.. vi tycks trivas här...  
*yeah .. unfortunately.. we seem to like it here...*
- 14 **Gabriel:** måste dra till 1timme  
*have to go to 1lesson*
- 15 **Calle:** jah.. vi e gäng  
*yeah.. we're a gang*

The contents of this example can be defined as meta-chatting where the participants gather in the chat-room to chat about how they gather in the chat-room to chat. It introduces a group of teenagers who know each other in the physical world but use the virtual environment as a meeting place equal to the physical world. For them the Net, just like in example 1, seems not to be a separate media for communication but rather an extension of their communicative reality. They join the discussion as a group (lines 1-8) and highlight the sense of community in their action (lines 10-12, 15). Even the quasinegative utterance in line 13 implicitly states that there must be something special about the chat-room because “*we seem to like it here*”. The contact between the physical and the virtual world can also be seen in line 14 where Gabriel informs others that he has to leave the chat-room because of a lesson. He feels a social need to explain to others why he has to go even if one might think that chatting does not include such social responsibilities as participating in a face-to-face conversation. Example 2 shows that such an assumption is false and that both communicating and socializing are extremely meaningful to the participants.

In example 3 below there are two participants, Sarah and Peter, chatting about how big Peter's feet are. Their conversation is influenced by things they do and see in their physical environment at the time of chatting.

### *Example 3*

- 1 **Sarah:** ja e int rädd för dina fötter  
*I'm not afraid of your feet*
- 2 **Peter:** du borde bli oxo  
*well you should be*
- 3 **Sarah:** någå små 40or..=)  
*some small size 40s..=)*
- 4 **Peter:** de faktiskt sant..int ha ja större fötter än d=)  
*it's actually true..my feet are not bigger than that =)*
- 5 **Sarah:** nå hur stora..dom ser no int stora ut  
*well how big..they don't look very big*
- 6 **Peter:** nå dom e typ 40  
*well they're about 40*
- 7 **Sarah:** haha..  
*haha..*
- 8 **Peter:** säg int de du håller på att skriva  
*don't say what you're writing*
- 9 **Peter:** voi voi voi  
*oh oh oh*
- 10 **Sarah:** smör smör smör  
*butter butter butter*
- 11 **Peter:** LOL

The conversation between Sarah and Peter has a feeling of intimacy that can be established even when chatters are total strangers. In this case there are yet some things that reveal that the participants know each other, that they can see each other but also that they can hear each other. Referring to someone's feet (line 1) instantly gives one a feeling of physical presence. This feeling is being confirmed in line 5 where Sarah says that Peter's feet don't look very big. It shows that Sarah is able to see Peter's feet at the time of chatting or that she has at least seen them previously. Peter admits that Sarah is right, his feet are only size 40 (line 6). After this Sarah starts to laugh and types *haha* (line 7). The most revealing utterance in this sequence is Peter's next utterance on line 8 where he says that Sarah should not say out loud the same thing she is writing. One can assume that Sarah has said *haha* at the time of typing and that Peter has been able to both hear

and see her laughter. The end of the conversation in lines 9-11 can be interpreted as a sign of intimacy as Sarah and Peter continue to joke. The joke in lines 9-10 is not possible to translate into English but it is typical for this group as it plays with words from both Finnish and Swedish. In line 9 Peter sighs and uses the Finnish word *voi* (*oh*). The joke is made by Sarah in line 10 about the double meaning of the word *voi*, also meaning butter in Finnish. She uses the word for butter in Swedish, *smör*, that does not have the same double content as the Finnish word. In line 11 Peter shows that he has understood the joke by typing LOL (Laughing Out Loud). In this example, like in examples 1 and 2, the participants share the same physical environment but move their conversation from off-line to on-line even if the topics come directly from their physical reality. This example also shows the participant's ability to use both Finnish and Swedish with creativity in everyday conversations.

## 5. Conclusions

Computer-mediated communication is a fascinating area of research that constantly gives researchers new dilemmas to solve. For me, one of them has been to figure out why so many teenagers in my data decide to chat on the internet even when they have the opportunity to talk face-to-face. They chat with each other at the same time they sit side by side in a computer class, where they are able to see and hear each other. My solution has been to interpret chatting in those situations as one mode of communication instead of one medium of communication. The participants in my data seem to use internet chatting as an equivalent to face-to-face communication and its different modes. They sometimes drop in to the chat-room just to exchange a few words and then continue their interaction in the physical environment. The unifying factor for all of them is that they are a part of the Finland-Swedish minority and that is also the biggest reason for these teenagers to chat in this specific chat-room. They have created a Finland-Swedish virtual community where participation, communication and interaction have their roots in the physical world and in the participant's cultural background. These roots lead to a specific way of communication where "virtuality" and "reality" become multimodality in a communicative activity. My interpretation breaks down the boundaries between "real world" and "virtual world" and raises new questions about how to define different ways and modes of communication. The questions might be more philosophical than practical, but I believe that the answers will give us a better understanding of interaction on the internet.

## References

Ahtikari, Johanna & Eronen, Sanna (2003), Netro.

[http://kielikompassi.jyu.fi/resurssikartta/netro/pankki/parametrit\\_modaali\\_multi.shtml](http://kielikompassi.jyu.fi/resurssikartta/netro/pankki/parametrit_modaali_multi.shtml)

Finnäs, Fjalar (2003). Finlandssvenskar 2002. En statistisk rapport. Folktinget.

<http://www.folktinget.fi/pdf/finlandssvenskarna2002.pdf>

Folktinget: Language Act 2004.

[http://www.folktinget.fi/en/language\\_act.html](http://www.folktinget.fi/en/language_act.html)

Saarela, Jan & Finnäs, Fjalar (2002). Language-group differences in very early retirement in Finland. In: *Demographic Research*, volume 7, article 3.

<http://www.demographic-research.org>

Saari, Mirja (2003). På väg mot ett blandspråk? Finska inslag i svenskan förr och nu. In Ivars, A-M., Maamies, S., Slotte, P. & Tandefelt, M. (Eds): *Boken om våra modersmål*. Festskrift till Mikael Reuter på hans 60-årsdag den 17 maj 2003. Esbo: Schildts Förlags Ab.

## **Biography**

**Jonna Ahti** is a doctoral student in Scandinavian languages at the University of Helsinki. Her research project deals with specific linguistic features in Finland-Swedish chat-rooms and how these features are used when creating a sense of a virtual community.

### **Author's address:**

*Jonna Ahti  
Department of Scandinavian Languages and Scandinavian Literature  
P.O.Box 24  
00014 University of Helsinki  
Finland  
e-mail: [jonna.ahiti@helsinki.fi](mailto:jonna.ahiti@helsinki.fi)*



# MULTIMODALITY IN OWN COMMUNICATION MANAGEMENT

*Jens Allwood, Elisabeth Ahlsén, Johan Lund & Johanna Sundqvist  
Göteborg University, Department of Linguistics  
and SSKKII Center for Cognitive Science*

## **Abstract**

*This study studies how gestures (here defined in a wide sense, including all body movements which have a communicative function) are used for Own Communication Management (OCM), an interesting and not completely well described part of the language system. OCM concerns how a speaker continuously manages the planning and execution of his/her own communication and is a basic function in face-to-face interaction. It has two main functions, i.e. “choice” and “change”. The study investigates how much of OCM involves gestures and whether there is a difference between choice and change OCM in this respect. It also concerns what kinds of gestures are used in OCM and what the relation is between vocal and gestural OCM. Some of the main findings are that roughly 50% of all speech based OCM cooccurs with gestures and that most of the OCM involving gestures (about 90%) is choice directed. Gestures occurring with OCM can illustrate the content of a sought after word, but also more generally induce word activation. They can also signal to an interlocutor that the speaker needs time. Gestures are often multifunctional and, thus, both choice and change are often integrated with more interactive functions. A final observation is that gestural OCM either precedes or occurs simultaneously with verbal OCM.*

**Keywords:** Gestures, Own Communication Management, choice, change, illustration, activation

## 1. Why study multimodality in Own Communication Management?

In order to function optimally, humans have evolved mechanisms for managing their communication. We can distinguish two main kinds of Communication Management (CM) – Interactive Communication Management (ICM) and Own Communication Management (OCM). Both of these types of management are continuously interwoven with each other and with the main message (MM) that is being communicated. (See below, figure 1.)

Own Communication Management concerns how a speaker continuously manages the planning and execution of the speaker's own communication and is a basic function in face-to-face interaction, while Interactive Communication Management concerns managing the interaction with other communicators through systems for turntaking, feedback and sequencing. Both types of management serve to share the main messages with other communicators and make communication more flexible and fluent by adapting it to face-to-face interaction demands on production and comprehension. Both are also fairly systematic (cf. Allwood, Nivre & Ahlsén 1990, 1992), and exhibit systematic variation between different activities, individuals, languages, cultures and other conditions.

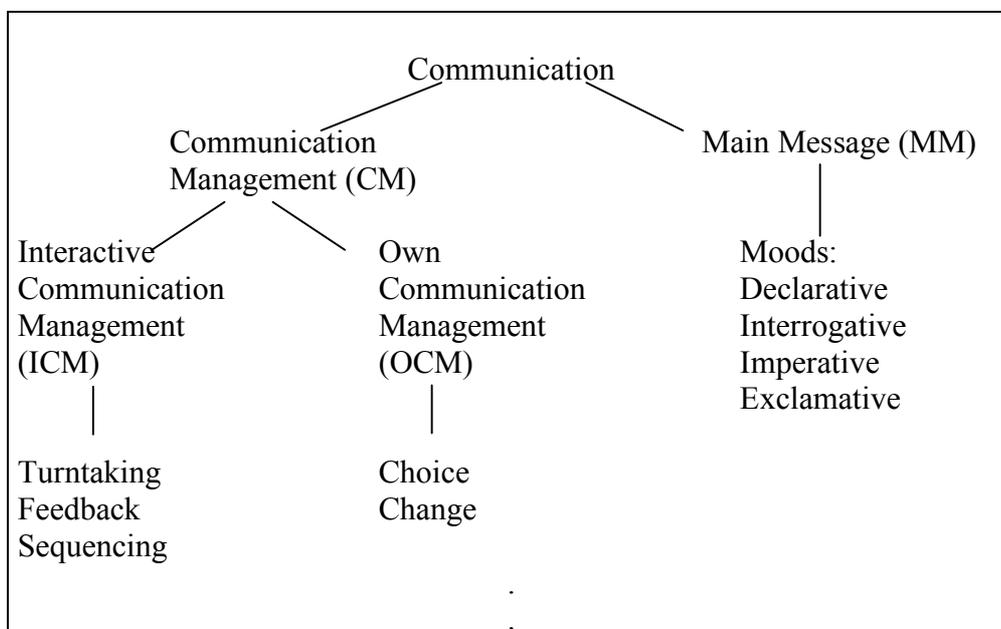


Figure 1. Main functions of Communication

The purpose of the present study is to study how gestures (here defined in a wide sense, including all body movements which have a communicative

function) are used for Own Communication Management, an interesting and not completely well described part of the language system.

As we have already mentioned, OCM concerns processes whereby a speaker manages his or her linguistic contributions to communicative interaction. Other terms that have been used for OCM are hesitation, planning, disfluency, (self) correction, editing and (self) repair (cf. Allwood, Nivre and Ahlsén 1990). OCM has also been described as “performance errors” (Chomsky) and “parole” (de Saussure). The term “disfluency” is here particularly noteworthy, since OCM, contrary to what this term suggests, often contributes to the fluency and flexibility of speech.

OCM has two main functions, i.e. “choice” and “change”. Choice mechanisms enable the speaker to gain time for processes having to do with the continuing choice (planning) of content and expression. Among other things, this involves prompting of memory, search of memory, hesitation, planning and keeping the floor. Change mechanisms enable the speaker, on the basis of various (internal and external) feedback processes, to change already-produced content and expressions. However, as we shall see, many cases of OCM are combinations of choice and change or of OCM with interactive communication management (ICM) and main message (MM) production. OCM is thus a central component in human spoken language interaction and a better understanding of how speech and body movements interact in OCM is a precondition for a theoretical account of spoken language. In addition, a better understanding of OCM will also have many practical applications, such as language teaching, speech therapy or computer based systems for speech and gesture recognition or speech and gesture synthesis.

## **2. Purpose, research questions and background**

Some of the questions we are trying to answer are the following:

- Question 1  
To what extent does OCM involve gestures?
- Question 2  
Is the distribution of OCM involving gestures (gestural OCM) different for OCM with choice as the main function and OCM with change as the main function

- Question 3  
Which kinds of gestures are used in OCM?
- what body parts?
- what types of movements?

- Question 4

What is the relation between vocal and gestural OCM?

The study will also address some of the current claims and hypotheses about gestures and semantic processes, especially in relation to word finding difficulties, e.g. the following:

1. Nouns and verbs (at least certain types of nouns and verbs) partly differ in activation areas in the brain and this is most likely related to different types of encoding, with association more related to areas for vision for nouns and motor areas (related to movement) for verbs (e.g. Pulvermüller 1999, Pecher et al. 2004). Can this, in turn, be related to differences in the types of gestures that are used together with nouns and verbs?
2. According to McNeill (2000), gestures and words are generated from a common growth point. An example of this can be seen in iconic illustrations, where words and gestures describing the same phenomenon are produced together. Does the role of iconic illustrations in OCM throw any light on this hypothesis? Related to this is the question of whether the analysis of different types of OCM gestures can give us better information about the semantic processing involved in speech planning. See, for example, the suggestions made by Kendon (1972), De Ruiter (2000) and Raucher, Kraus and Chen (2000).
3. Gestures help to activate and package information in a way that makes it easier to verbalize (Kita 2000). This has been seen in children doing cognitively complex tasks and in children with word finding problems. In the present study we investigate whether a similar tendency can be found in adults in spoken interaction.

Thus, the purpose of the study is to analyze what gestures occur in OCM. In order to study this, we first used a database consisting of a sample of 100 instances of speech based OCM from two video recordings of informal discussions. This sample is used as a basis for finding out how often speech

OCM involves gestural OCM. It is also used to find out what proportion of speech OCM is used for choice and change (speech OCM).

In a second step, we have used a sample of 100 examples of OCM involving gestures (gestural OCM). The examples have been extracted from video recorded interviews and discussions in the GSLC (Göteborg Spoken Language Corpus) (Allwood et al. 2000). The two samples were then used for a further analysis of OCM to be reported below. On the basis of this analysis, we attempt to answer the questions stated above and discuss the three claims and hypotheses concerning differences in gesturing between noun and verb activation, gestures as clues to semantic planning and gestures as facilitating packaging of information.

### 3. Two examples of OCM with gestures

In order to give a better understanding of the nature of OCM, we start by considering two examples.

#### *Example 1. Choice OCM*

Speaker: *å där så de e som en e // sportspår där som vi springer*

(and there so it is like a eh // sportstrack where we run)

A closer look at what gestures co-occur with the phrase *en e // sportspår* (a eh // sports track) is presented below, in Table 1.

Table 1. Choice OCM

Speech	en	e	//	sportspår
Type	Article	OCM word	pause	Noun
Gesture	hand circling, illustrating track	turns away head and gaze		head and gaze back

If we start by examining the temporal relation between the vocal-verbal and gestural production, we see that an illustrating gesture occurs before the OCM word *e* and pause which, in turn, precede the possible target word (*sportspår*). Since an indefinite article, *en*, is produced, that, however, is of the wrong gender for the noun actually produced later, we interpret this as indicating that the speaker has a problem in choosing and producing the right noun and that this is reflected in the production of the OCM word *e*

and a pause. Additional evidence that there is a word finding problem is provided by the fact that the illustrating gesture occurs when the article preceding the OCM is produced. This means that the gesture is probably not an illustrating iconic gesture, which could have occurred even if the speaker had no need for support in finding the word. Rather, it probably has a self-activating word finding function for the speaker, while at the same time keeping the floor and giving a clue about the meaning of the coming noun to the listeners.

In accordance with the “common growth point” hypothesis, McNeill (2000) claims that the peak of an iconic (i.e. based on similarity) gesture co-occurs with the stress of the iconically illustrated word, thus indicating a close semantic and articulatory relationship between speech and gesture production. Since the gesture here, however, precedes the corresponding noun, it probably therefore rather has a facilitating function (cf. Kita 2000).

Simultaneously with the OCM word *e* and pause, the speaker turns his head and gaze away from the interlocutors, perhaps indicating memory search and turnkeeping. When he produces the noun he moves his head back facing the interlocutors.

Thus, the example indicates that OCM contains a number of elements, vocal-verbal (OCM word *e* + pause) as well as gestural, and that the temporal relation between the modalities is not simple. The example is in this way fairly typical of the complex relation between OCM, speech and gesture.

Let us now consider a second example that illustrates how choice related OCM is related to change related OCM and how OCM is related to ICM and MM.

### *Example 2. Multifunctional OCM*

Speaker: *jo fö+ för att inte: eh //eh för att hålla en del grödor vid liv*

(yes fo+ for (in order)no:t eh // eh for (in order) to keep some crops alive)

This instance of OCM involves a complex combination of choice and change related OCM. Below, we first summarize the choice related parts in Table 2.

Table 2. Choice OCM

Speech	fö+ för att	inte:	eh	//	eh
Type	self repetition	vowel lengthening	OCM word	pause	OCM word
Gesture and function	turntake	two hands, turntake ICM, emphasis MM, activation OCM	head turn away-down, gaze away-down		

It also involves an instance of change related OCM (see Table 3). The speaker changes her mind from saying *in order not to (for not to)* to saying *in order to keep (for to keep)*.

Table 3. Change OCM

Speech	<u>för att inte: eh // eh för att hålla</u>	en del grödor vid liv
Type	substitution	
Gesture	head turn back, gaze back (eye contact), hand ICM	head nod affirmation elicit understanding

As she makes the substitution *för att hålla*, she turns her head and gaze back towards the listener, establishes brief eye contact and makes an offering gesture with her hand. The fact that the occurrence of this gesture is more or less simultaneous with the verbal utterance *hålla* indicates that no time is needed for activation or memory search, rather the hand gesture pinpoints the delivery of the substitution.

The example also shows how OCM functions often are integrated with ICM and MM functions. The use of the word *jo*, combined with self repetition and vowel lengthening of *inte:*, helps the speaker to take the turn and maintain it. Her main message (MM), at this point *inte:*, is strengthened by the lengthening, which gives emphasis. In a similar way, her hand gesture, in carrying out the substitution, functions as a contact maintaining aid to the listener (ICM) and her closing head nod functions to affirm her statement as well as to elicit understanding and possible agreement from the listener (ICM).

After having considered two examples of OCM, we will now turn to a finer classification of OCM and a consideration of what the data in our two samples reveal about OCM and gestures, especially in relation to the research questions listed above.

#### 4. Expressive features of OCM

If we look at OCM units produced in spoken interaction, we can classify them into different types, depending on their expressive features (cf. Allwood, Nivre and Ahlsén 1990). Our first classification separates units with single OCM features from units with several combined OCM features. There are two main types of unit with single OCM features, i.e. Basic OCM expressions and other units influenced by Basic OCM operations:

Basic OCM expressions:

- A. Pauses
- B. Simple OCM expressions, for example hesitation words, like *eh*, *uh* or *m*
- C. Explicit OCM phrases, like *what's it called*
- D. Other OCM sounds, like sighing, smacking or hissing

Basic OCM operations:

- A. Lengthening of continuants
- B. Self interruptions
- C. Self repetitions

The difference between the two kinds of units is that the first kind “basic OCM expressions” have as their type meaning an OCM function, while the second kind can have any type meaning, but are given an OCM function through the OCM operation, e.g. lengthening *inte*: (no:t) or self interruption + repetition *fö+ för* (fo+ for) in example 2 above.

The second main class consists of OCM units which involve combinations of basic OCM expressions and operations. An overview is given in Figure 2 below. For a more detailed explanation, see Allwood, Nivre and Ahlsén (1990).

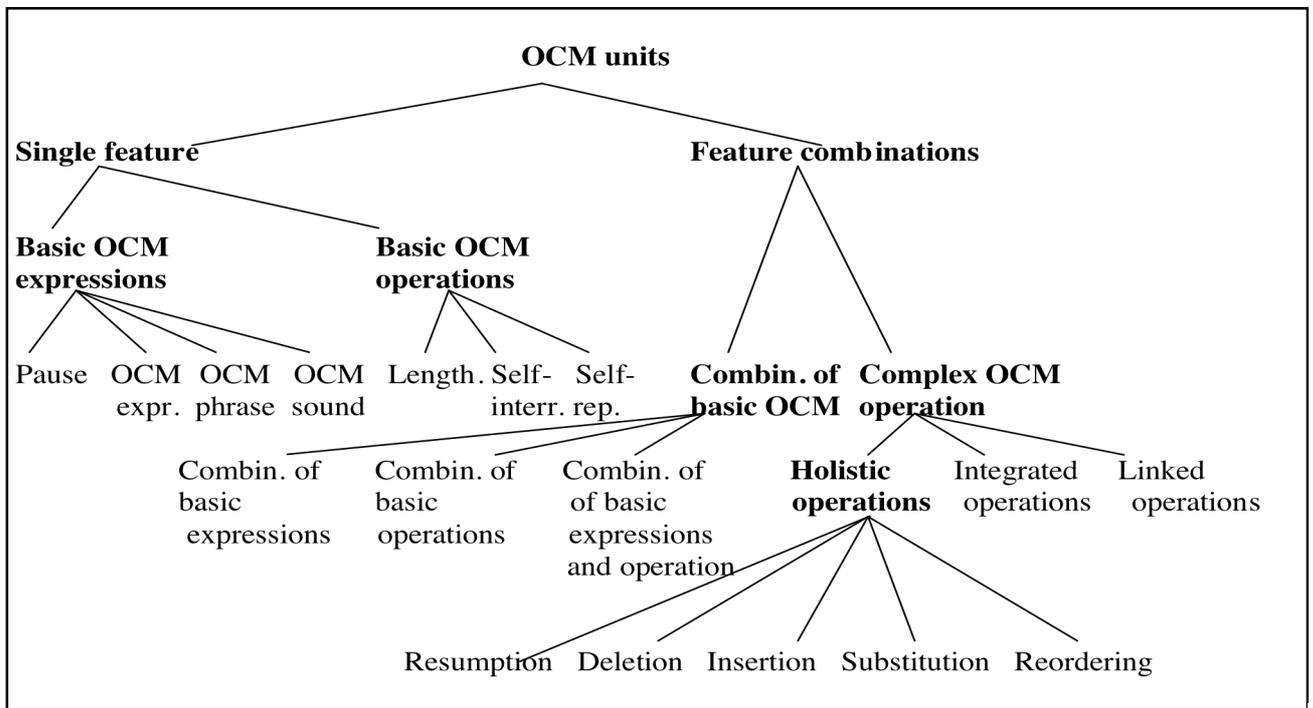


Figure 2. Taxonomy of OCM features

## 5. Functions of gestures in OCM

### 5.1. Choice, change and gesture

If we start by examining functions for speech based OCM in general, we find that 90% of the examples in the first sample have choice as their main function, whereas only 10% have change as their main function. As expected, the two functions very often cooccur. The distribution of functions is the same if we look at the second sample of OCM, involving gestures (89% vs. 11%). Choice is thus, by far, the most common OCM function. For a summary, see Table 4 below.

Table 4. Distribution of choice and change functions in two samples of OCM

Function	Speech based OCM		OCM involving gestures
	Sample 1	Sample 2	
Choice	90%	89%	
Change	10%	11%	

If we continue by examining to what extent gestures occur in connection with speech based OCM in Sample 1 and what features, in accordance with the taxonomy in Figure 1 above, occur in the sample, we get the following results, summarized in tables 5a and 5b, below .

Table 5a. Gesture involvement in speech based OCM

Gesture involvement	
Gesture involvement	55%
Only OCM related gestures	45%

Table 5b. Types of OCM involving gestures.

Types of OCM	
Basic OCM expression	39%
Combination of basic OCM features	47%

Table 5a shows that in Sample 1 (100 speech based OCM units), 55% of all instances of OCM involve gestures. If gestures that are not clearly related to OCM functions are removed, we find that 45% of the speech based OCM instances involve an OCM related gesture, while 10% involve gestures not clearly related. A gesture is not clearly related to OCM if it has some other identifiable main function, such as waving to a friend or drinking coffee etc. There are also gestures, like scratching the head or touching the face that can be related to both OCM and ICM, but these have not been counted as clearly related. We can, thus, conclude that about half

of the instances of speech-based OCM include gestures. In Table 5b, we see that in the 100 cases of speech based OCM, the most popular OCM speech features to combine with gestures are Single basic OCM expression (39%) and Combination of basic OCM (47%).

So in answer to the first question posed in section 2 above, i.e. to what extent OCM includes gestures, we found that in the sample of speech OCM approximately 50% of OCM instances include gestures, where hand and head gestures are the most popular. Considering the main functions of choice and change, about 90% of OCM is choice related, while only about 10% is change related. This holds in both the sample of speech based OCM and the sample of gestural OCM.

In Table 6, we present how gesture involvement is distributed between OCM with change and choice function.

Table 6. Gesture involvement in speech based OCM

	<i>Speech based OCM</i> - choice function	<i>Speech based OCM</i> - change function
Gesture occurrence	55	45
OCM related gesture	40	15

Out of the instances of speech based OCM with a choice function, 55% involve a gesture (of these, 15% are unrelated). For speech based OCM with a change function, 45% involve gesture (of these 30% are unrelated): For speech based OCM, it thus holds for both change and choice functions, that they are more common without related gestures. (60% of choice related OCM and 85% of change related OCM do not involve directly related gestures.)

### ***5.2 Types of gestures involved in choice and change***

If we look at the types of gestures occurring in OCM with choice function, the most frequent ones (more than 5%) are, in order of frequency:

Table 7. Proportions of different gesture types in OCM with mainly choice function.

---

Hand gesture	55%
Gaze down	21%
Head shake	12.5%
Gaze up	7%
Gaze to side	6%
Head nod	5%

---

*(The table sums to more than 100%, since some gestures occur simultaneously as a complex unit, but all the features have been counted separately here. Only numbers over 5% are included.)*

The most frequent choice related gestures are hand gestures, followed by gaze changes and head movements.

The most frequent (more than 5%) gestures occurring in OCM with change function are, in order of frequency:

Table 8. Proportions of different gesture types in OCM with mainly change function

---

Hand gesture	65%
Gaze down	5%
Gaze to side	5%
Gestures types that occur less than 5%	25%

---

Again, we can see that hand gestures of different types are the most frequent, followed by gaze down and gaze to the side. Combining both table 7 (choice) and table 8 (change), the most popular OCM gestures involve hand movement (56%), followed by gaze (31%) and head movement (15%).

### **5.3 Gaze Aversion**

As tables 7 and 8 show, gaze aversion is more related to choice than to change. This could point to a difference in the need for memory activation, where gazing away from the interlocutor indicates a greater degree of memory activation.

#### 5.4 Functions co-occurring with choice and change

In table 9, below, we give an overview of some of the other functions we have found correlated with choice and change functions in OCM.

Table 9. Functions that correlate with choice and change functions in OCM

<b>Gesture functions accompanying Choice OCM</b>		<b>Gesture function accompanying Change OCM</b>	
Illustrating the content of a sought after concept by iconic gesture	25%	Illustrating the content of what is changed	21%
Activation inducing gesture, moving hands to mobilize energy	61%	Activation - often with both hands	71%
Less clearly OCM related function, like scratching or supporting head	14%	Less clearly OCM-related function	8%

The table shows that the related function types are fairly similar for choice and change. If we combine the results for choice and change, we find that 24.6% of all OCM related gestures also have an illustrative function which, however, often also has an activating role in word finding, 62% have an activation function that is not illustrative and 13.4% have a function that is less clearly relatable to OCM.

## 6. More about functions of gestural OCM

### 6.1 Iconic illustration and general activation

The kinds of gestures we find with OCM include gaze, head and hand gestures. For types of gaze we find (in order of frequency of occurrence) gaze down, up or to the side. We find head nods, head shakes and head movements to the side. Often choice or change function are combined with other functions. For hands, we find both illustrative OCM related hand gestures and non-illustrative but OCM related gestures, as well as gestures that are less clearly related to OCM, with the hand close to the face. The distribution of hand gestures concerning these functions is shown in Table 10.

Table 10. Distribution of illustrative gestures, related but non-illustrative gestures and less clearly related hand gestures in a sample of 100 instances of gestural OCM.

Type of hand gesture		N 100
Hand lifted/waving (non-illustrative)	One hand	41
	Both hands	15
Illustrating hand gesture	One hand	13
	Both hands	8
Less clearly related gestures		23

Over and above choice and change, table 10 probably reflects two other main functions of gestural OCM. The first is “general activation”, which probably is also related to interactive communication management functions, such as turn keeping and attention holding. This function is more frequent and is in most cases accomplished using one hand. The other function is probably related to content activation and/or illustration. This function is less frequent but we can see that compared to activation gestures, it is relatively speaking more often made with both hands.

## 6.2 *Hands, gaze, nouns and verbs*

We also find a difference in activation of nouns and verbs in that OCM for choice related to verbs is more often accompanied by gestures requiring both hands. OCM for choice related to nouns is, on the other hand, often related to gazing down, something that does not occur in our database with verb related OCM (see Table 11).

Table 11. Gestures with choice OCM: noun and verb related

	Gaze	One hand	Both hands	Illustrating	Head
Choice of N/NP	9	17	6	7	1
Choice of V/VP/predicate	2	9	12	5	1

Possible explanations for these differences could be that nouns are more visually encoded and that this encoding is activated by a downward gaze trying to retrieve a visual image, whereas verbs are encoded more in relation to physical action, sometimes involving both hands, the movement of which activates the encoding. This would be in accordance with the findings by Pulvermüller 1999 and others who have claimed that gestural differences of the type discussed exist between noun and verb activation.

### ***6.3 Hand gestures - the main functions related to choice and change***

The main hand gesture functions related to choice OCM are: illustrating the content of a sought after concept by iconic gesture (25%), activation gesture, moving hands to mobilize energy (61%) and possibly non-OCM related function, like scratching or supporting head (14%). If we turn to change OCM, the main gesture functions are: illustrating the content of what is changed (21%), activation, often with both hands (71%) and possibly non-OCM-related function (8%). We can, thus, conclude that the distribution of gesture functions is fairly similar for choice and change OCM. Both of these functions can be related to two further, partly different cognitive functions, i.e. activation and illustration. A closer analysis of illustrating OCM gestures could very well reveal that many of them also have a more specific information packaging or activating function, as claimed by Kita (2000).

One reason for the assumption that illustrating gestures in an OCM context also can serve an activating and possibly information packaging function, is that in a number of examples, illustrating gestures relating to the sought for “target word” preceded the target word and co-occurred with OCM. Sometimes, as in Example 1 above, an illustrating gesture even precedes the verbal-vocal OCM and can be seen as the first part of the OCM. In this context, we may also consider the following example, where a man describes a toy museum with windows to protect the toys from the children.

### Example 3

Speaker: *de e alltså / de e fantastiskt å se just lek+//saksmuseum som skyddas / ö: från barnen med f+ // från med med fönster / glasade fönster*

(it is then / it is fantastic to see precisely to+ /+ymuseum that is protected / uh from the children with f+ // from with with windows / glass windows)

In Table 12 below, we give an analysis of the utterance, in terms of the function of the vocal verbal, as well as the gestural expressions.

Table 12. Analysis of function, vocal verbal and gestural aspects of an OCM sequence.

Function	Vocal verbal (Swedish)	Vocal verbal (English translation)	Gestural expression/function
<b>Change (substitution)</b>	de e alltså...de e fantastiskt	it is then ... it is fantastic	ICM elicitation/affirmation nod
<b>MM (emphasis)</b>			MM hands together focus
<b>Choice + MM (emphasis) + ICM (contact)</b>	lek+ /+saksmuseum	to+//y museum	Gaze - eye contact
<b>Choice</b>	/ ö:	/uh	Head/gaze turn back-down
<b>Choice</b>	f+ //	f+//	Hand gesture up & down, illustrating window
<b>Aborted resumption</b>	från	from	Quicker hand movement up and down
<b>Choice/change (deletion)</b>	med med fönster	with with windows	Head nod Gaze downwards ICM gaze turn to listener 3 Hands in resting position
<b>Change (insertion) MM (emphasis)</b>	fönster / glasade fönster	windows / glazed windows	Turn back to listeners 1 & 2

(/ = short pause, // = medium pause, + = self interruption)

We can see that the temporal relation vocal verbal – gestural is often simultaneity, but that the illustrating gesture for window comes well before the word *window*.

Although the timing of the gestural OCM and speech based OCM, ICM and MM deserves a more comprehensive and detailed analysis than has been possible in this study, we can see in our examples that gestural OCM, when it is used, can precede both vocal OCM words and the vocal related main message words. This opens up the perspective that OCM gestures, even elaborated ones with a more specific semantic content, are, at least sometimes, more easily and spontaneously produced than verbal-vocal OCM or MM output. This provides additional support for OCM gestures as an interesting object of study when trying to understand the speech planning and processes of speech production.

#### ***6.4 Head shakes, choice and change***

The function of head shakes in choice and change OCM differs. Choice-related headshakes seem to indicate uncertainty, searching for words or that the situation is perceived as strange. Change-related head shakes, on the other hand, indicate the rejection of one expression in favor of another. Typical cases are when *eller* (or) or *nä* (no) is uttered with a headshake in a change context.

#### ***6.5 Choice and change of words vs choice and change of clauses***

If we turn to the second question in section 2 above, concerning the role of gestural OCM in speech planning, we find that 65% of the choice related and 75% of the change related OCM occur before choice or change of a word, whereas only 35% of the choice related and 25% of the change related occur before the choice or change of a clause. Words, thus, perhaps are a more basic units in planning than clauses.

### **7. Summary and conclusions**

Perhaps the main observation of this study is that many gestures are multifunctional, they can simultaneously support the main message (MM), e.g. by iconic illustration, while at the same time managing the interaction (ICM) by keeping the floor, maintaining contact and attention, and thirdly

facilitating the speaker's own communication (OCM) by activating his/her memory and providing time for planning.

Some of the main findings, answering the initial questions posed in section 2 above are:

- (i) Roughly 50% of all speech based OCM co-occurs with gestures (Question 1).
- (ii) Roughly 90% of both vocal verbal and gestural OCM is choice directed, whereas about 10% is change directed (Question 2).
- (iii) Roughly 40% of all speech based choice related OCM involves gestures. The corresponding share for speech based change related OCM is 15% (Question 2).
- (iv) In choice OCM, gestures can illustrate the content of a sought after word by iconic gesture (25%), but they can also more generally induce activation by moving some part of the body to mobilize energy (61%). Finally, they can signal to the interlocutor that the speaker needs time for planning by scratching or supporting the head (14%) (Question 2).
- (v) In change OCM, gestures also illustrate the content of what is sought after in order to make a change (21%). In this way, choice is often integrated in change, as one of the means whereby change is achieved. The gestures can also be used to mobilize energy (71%) or be used for some unrelated function (8%) (Question 2).
- (vi) Hand movements, gaze change and head gestures are the most popular types of gesture involved in OCM (Question 3).
- (vii) Concerning the relation between vocal and gestural OCM, we have noted that gestural OCM either precedes or occurs simultaneously with vocal OCM. We have also noted that gestural OCM can be multifunctional and have an iconic illustrative MM function connected with the OCM function (Question 4).

Thus, in answering our initial questions, we have found a number of possible functions for OCM gestures. In addition to choice and change, we have found hand gestures that are more generally activating and hand gestures that are illustrating the content of a word the speaker wants to express. It is important to note that these functions need not be distinct, rather they generally reinforce each other. Thus, an illustrating type of hand gesture might also serve an activating and information packaging function. We have also found more gaze aversion in choice OCM than in change OCM and suggested that this might be related to a need for memory search. Finally, we found that head shakes were used with different functions for

choice and change and that words more often than clauses seem to be the units that are subject to gestural (as well as vocal-verbal) OCM.

As regards the three hypotheses and claims also mentioned in section 2, we have found that there was a difference in OCM gestures used when searching for nouns and verbs (gaze downwards only for nouns and more gestures with both hands for verbs) (Hypothesis 1, Pulvermüller 1999, Pecher et al. 2004). As regards hypothesis 2, that gestures and words are generated from a common growth point (McNeill, 2000), our observations show that this is not always the case, since iconic gestures referring to some phenomenon sometimes precede the vocal word referring to the same phenomenon. In fact, this, instead supports Hypothesis 3, i.e. that gestures help to activate and package information (Kita 2000).

We have, thus, described some of the features of gestures in Own Communication Management. We believe that giving detailed information on human multimodal communication, including gestural as well as vocal-verbal communication is both an empirical and a theoretical challenge, affecting how we describe, understand and explain the structure of spoken language. We also believe that such descriptions will be exploitable in practical applications based on speech and gesture; both in systems for production and generation/synthesis and in systems speech/gesture recognition and understanding.

## References

- Allwood, J., Björnberg, M., Grönqvist, L., Ahlsén, E. & Ottesjö, C. (2000). The spoken language corpus at the department of linguistics, Göteborg University. *FQS – Forum Qualitative Social Research*, 1.2, December 2000.
- Allwood, J., Nivre, J. & Ahlsén, E. (1990). Speech management - On the non-written life of speech, *Nordic Journal of Linguistics* 13, 3-48.
- Allwood, J, Nivre, J, & Ahlsén, E. (1992). On the semantics and pragmatics of linguistic feedback, *Journal of Semantics*, vol. 9, no. 1, 1-26.
- De Ruiter, J. (2000). The production of gesture and speech. In McNeill, D. (Ed) *Language and Gesture*. Cambridge University Press, pp.284-311.
- Kendon, A. (1972). Some relationships between body motion and speech: an analysis of an example. In Siegmann, A & Pope, B. (Eds), *Studies in dyadic communication*, 177-210.
- Kita, S. (2000). How representational gestures help speaking. In McNeill, D. (ed) *Language and Gesture*. Cambridge University Press, 284-311.

- McNeill, D. 2000. (ed) *Language and Gesture*. Cambridge University Press
- Pecher, D, Zeelenberg, R. & Barsalou, L. W. 2004. Sensorimotor simulations underlie conceptual representations. Modality-specific effects of prior activation. *Psychonomic Bulletin & Review*, 11, 164-167.
- Pulvermüller, F. (1999). Words in the Brain's Language. *Behavioral and Brain Sciences*, 22, 253-336.
- Raucher, F. H., Krauss, R. M. & Chen, Y. (1996). Gesture, speech, and lexical access: the role of lexical movements in speech production. *Psychological Science* 7: 226-230.

## **Biography**

**Jens Allwood** is since 1986 professor of Linguistics at the department of Linguistics at Göteborg University. He is also director of the interdisciplinary cognitive science center SSKKII at the same university. His research primarily includes work in semantics and pragmatics. He is investigating spoken language interaction from several perspectives, e.g. corpus linguistics, computer modelling of dialog, sociolinguistics and psycholinguistics as well as intercultural communication. Presently he is heading projects concerned with the semantics of spoken language phenomena, multimodal communication, cultural variation in communication and the influence of social activity on spoken language.

**Elisabeth Ahlsén** is professor of neurolinguistics at the Department of Linguistics, and SSKKII Center for Cognitive Science, Göteborg University. Her main areas of research are neurolinguistics, pragmatics and communication disorders. She teaches neurolinguistics, psycholinguistics, cognitive science, communication analysis and linguistic research methods. She coordinates a number of research projects on communication disorders in adults and children, focusing on pragmatics, semantics, gesture and ICT support.

### **Johan Lund**

Studied cognitive science at Göteborg University as well as psychology and computer science. Develop research tools at Göteborg University for multimodal communication, i.e. Multitool and Baldi.

Researcher and producer (5 years experience as computer consultant within the industry and 10 years experience in the engineering trade)

Present occupation: Compose and produce music, videos and manage artists.

### **Johanna Sundqvist**

MA in computer science at Göteborg University

Present occupation: Project assistant with a company working with technical translations.

**Authors' addresses:**

*Jens Allwood, Elisabeth Ahlsén,  
Göteborg University, Department of Linguistics  
and SSKKII Center for Cognitive Science  
Box 200  
S-405 30 Göteborg  
Sweden  
phone: +46 31 773 1867/1923  
e-mail: jens@ling.gu.se  
eliza@ling.gu.se*

*Johan Lund  
Skäpplandsgatan 5  
414 78 Göteborg  
e-mail: johan\_lund@bredband.net*

*Johanna Sundkvist  
Redbergsvägen 9B  
41665 Göteborg  
phone: 0704-372308  
e-mail: yoyo@goteborg.utfors.se*

# WORD-FINDING PROBLEMS IN MEDICAL CONSULTATIONS BETWEEN NON-SWEDISH PHYSICIANS AND SWEDISH PATIENTS

*Jens Allwood and Nataliya Berbyuk  
Göteborgs University, Sweden*

## **Abstract**

*One of the most common difficulties in medical consultations between non-Swedish physicians and their Swedish patients is the problem of finding the right word during the interaction, often as a result of using Swedish as a foreign language. In this study, some of the ways this problem is handled during interaction are presented and the influence of the ways of handling the problems on the physician-patient relationship is discussed.*

**Keywords:** physician, patient, communication, word-finding problems, gesture, joint production, patient involvement

## **1 Introduction**

Successful communication between physician and patient is essential for a high quality in health care and for the well being of a patient. In the course of a consultation, the physician and patient exchange information that is important to both of them - the patient provides information about experienced health problems, on the basis of which the physician makes a diagnosis, suggests a treatment and subsequently informs the patient. Consequently, the ability of both participants to contribute and share enough relevant information is essential.

In spite of the fact that the number of studies that focus on factors that lower the quality of information exchange in physician-patient communication is high, most of them concentrate exclusively on problems

stemming from the patient, e.g. a patient's language impairment in the case of aphasia (Demeurisse, 1999), a patient's deafness (McEwen & Anton-Culver, 1988) or from the fact that the patient is a foreigner, e.g. an immigrant, a representative of a minority ethnic group, etc with poor cultural and language competence (Fernandez *et al.*, 2004), who requires the help of an interpreter (Davidson, 2001). The physician, on the contrary, is usually seen as the dominant participant in the interaction, whose strong position as a health care provider is even more strengthened as a result of a patient's "weakness" resulting from physical and/or psychological problems or from cultural and language difficulties.

Research relating to problems a physician might have, e.g. a physician's language problems or due to the physician being a foreigner, are rare, in spite of an increased migration of health care personnel (Mejia, 2004). There are only a few such studies, some examples are the overview of problems experienced by foreign/ international medical graduates (FMG/ IMG) in the USA (Miller *et al.*, 1998; Steward, 2003), overseas-trained doctors (OTD) in Australia (McGrath, 2004), utländska läkare (foreign physicians) in Sweden (Ekström, 2004, Allwood, Berbyuk & Edebäck, 2004, Berbyuk, Allwood & Edebäck, 2004, Berbyuk, 2005), etc. Some studies on "foreign physician-native patient" communication worth mentioning here are (i) a brief overview of the problems reported by FMG (Fiscella *et al.*, 1997), (ii) a chapter in Handbook for Foreign Medical Graduates on cultural and language issues essential for successful communication in American health care (ECFMG, 1976), and (iii) a study on power issues in interaction between Polish and Vietnamese physicians and American patients (Erickson & Rittenberg, 1987).

The recent increase of foreign physicians in Sweden (Lindberg, 2005) makes the issue of intercultural communication between them and their Swedish patients and colleagues important for society. The case of a physician being a foreigner and using Swedish as a foreign language raises a number of questions, one of which concerns the influence of the physician's lack of language competence on the interaction with patients. In the research project 'Communication and Interaction in Multicultural Health Care,' initiated in 2003 at the Department of Linguistics, Gothenburg University, the communication between non-Swedish physicians and their Swedish patients and colleagues is analyzed. The purpose of the above mentioned project is to describe and to analyze the difficulties arising from, and possible positive effects of, cultural differences and the use of foreign language. The project also considers the

influence of gender on communication. In addition to this, language learning at work, i.e. the ways in which non-Swedish physicians get ‘informal tuition’ from their communicative partners, is being studied. The methods used in the project are interviews and questionnaires directed to non-Swedish physicians, health care personnel and Swedish patients as well as video and or/audio recordings of medical consultations and working meetings of different kinds.

The results of the interviews and the questionnaire studies show that non-Swedish physicians experience some difficulties in interaction (Allwood, Berbyuk & Edebäck, 2004, Berbyuk, Allwood & Edebäck 2004, Berbyuk 2005). A very frequent example of this is the problem in finding the right word during interaction with patients (Allwood, Berbyuk & Edebäck, 2004). In this paper, we will exclusively present a taxonomy of the strategies to solve the word finding problems we have found in the data. The question arises: In what situations does this problem occur? How and by whom is it solved? What influence does it have on interaction and on the physician-patient relationship?

## **2 Data and Methods**

Transcriptions of recordings of 33 medical consultations (31 video and 2 audio recordings) between 13 non-Swedish physicians (7 male and 6 female) and their Swedish patients (33 patients, 13 male and 20 female) have been analyzed. The recordings have been made in health care centers and hospitals in Western Sweden. Both physicians and patients gave their oral and/or written consent. The physical examinations were not video recorded for ethical reasons. If participants gave their consent, it was audio recorded only, i.e. the lens lid was placed on the camera, which meant no video recording was made. No researcher was present during the recording of a consultation.

All patients are native speakers of Swedish, aged between 20 up to 89 years. The physicians come mainly from Iran (5) and Hungary (4). Other countries represented are former USSR (Russia), Colombia, Germany and former Yugoslavia, represented by one physician each and included in what will be referred to as the Mixed group below. All physicians completed their medical education and gained some professional experience in their native countries before coming to Sweden. However, the time they have lived in Sweden, their professional experience and specialties vary (see

Table 1 below). The Hungarian doctors and a German doctor were recruited in a recruitment program started by the Västra Götaland region (Western Sweden) and, since they had their medical licenses automatically approved under EU/EEA (European Union/European Economic Area) regulations, they began work directly after coming to Sweden. They attended a three-month Swedish language course in their native countries. The physicians from outside the EU, i.e. the Iranian physicians, the Russian, the Yugoslavian and the Colombian physicians started working in the Swedish health care system 2–4 years after coming to Sweden, having supplemented their medical education and passing a compulsory language examination for physicians from outside the EU/EEA in order to have their medical licenses approved. In Table 1 below a brief overview of the non-Swedish physicians that participated in the study is presented.

Table 1. Overview of the non-Swedish physicians

<b>Participants' code</b>	<b>Age</b>	<b>Gender</b>	<b>Time in Sweden (years)</b>	<b>Work as physician in Sweden (years)</b>	<b>Specialty</b>
<b>Hungarian group</b>					
HuD1 <sup>1</sup>	45	male	1	1	anesthesiology
HuD2	34	female	1	1	
HuD3	36	male	1.5	1.5	
HuD4	44	male	2	2	
<b>Iranian group</b>					
IraD6 <sup>2</sup>	49	female	13	10	geriatrics rehabilitation
IraD7	40	female	7	>1	general medicine
IraD8	45	male	14	12	surgery
IraD9	48	male	17	13	ophthalmology
IraD10	50	female	18	15	obstetrics, gynecology
<b>Mixed group</b>					
ColD17	39	male	12	10	surgery
GerD12	56	male	1	1	orthopedics, rehabilitation
RusD19	45	female	14	10	general medicine
YuD20	43	female	2	4	anesthesiology

Transcriptions of the recorded interactions have been made using MSO transcription standard (Nivre et al 2004) and have been checked by two

<sup>1</sup> Abbreviations: Hu=Hungarian, resp. Ira=Iranian; Col=Columbian, Ger=German, Rus=Russian and Yu = f Yugoslavia

<sup>2</sup> Participants' codes are taken from the database of the forthcoming PhD thesis, in which in total 20 non-Swedish physicians are included. To avoid confusion in future references, the original numbering from the database is preserved here.

independent checkers. Relevant sequences where the physicians' word-finding problems occur have been selected and analyzed. The transcription conventions applied in MSO and GTS used in the examples in the article are presented in Table 2 below:

Table 2. Transcription conventions

<b>Symbol</b>	<b>Explanation</b>
<b>\$P, \$D,</b>	participant (patient, doctor)
<b>[ ]</b>	overlap brackets; numbers used to indicate the overlapped parts
<b>( )</b>	transcriber's uncertain interpretation of what is being said, e.g. (pritsche)
<b>/, //, ///</b>	a short, intermediate and a long pause respectively
<b>+</b>	incomplete word, a pause within word
<b>CAPITALS</b>	contrastive stress
<b>:</b>	lengthening
<b>&lt; &gt;, @ &lt;&gt;</b>	comments about non-verbal behavior, comment on standard orthography, other actions

### 3 Results

Sequences in which non-Swedish physicians have word-finding problems have been identified and analyzed in the material. The analysis reveals the variety of strategies used by the non-Swedish physicians to manage interaction with patients.

The non-Swedish physicians often have problems in formulating their messages. The statistical corpus analysis of data, obtained for the PhD dissertation shows a higher number of pauses and OCM (own communication management) in the speech of the non-Swedish physicians as compared to the Swedish physicians. This indicates a slower tempo of interaction as well as language problems (Berbyuk, forthcoming). The non-Swedish physicians, in the majority of cases, attempt to solve the lexical problem themselves, i.e. without help from their patients or other participants present during interaction, with varied success. The physicians recall the words by taking extra time, using gestures for retrieval as well as by using substitutes for the sought words. They also tend to paraphrase their messages, use equivalents from other languages, as well as medical terms. Below, we first give an overview of the word finding procedures used by non-Swedish physicians and secondly we present the procedures used by the Swedish patients (and other participants involved in the consultations) to talk to non-Swedish physicians.

### 3.1 Strategies used by non-Swedish physicians for handling word-finding problems in production

#### 3.1.1. Using own communication management (OCM) features to recall a word

We will consider the case where a physician successfully uses OCM strategies (cf Allwood, Ahlsén in this volume) to recall words. He uses such strategies to recall both the complex verb *titta i* (look in) and the noun *spegel* (mirror), where it is likely that the noun is being searched for already in the activation of the verb.

Speaker	Transcription	Translation into English
ŞD:	får ja fråga / e / har du MÄRKT att e: / <1 dina ögon >1 blev <2 lite >2 / gu:lfärgad // om du om du <3 ö: tittar i >3 <4 e: å:h va heter de / >4 spegel	may i ask / er / have you NOTICED that er: / <1 your eyes >1 became <2 slightly >2 / ye:llow coloured // if you if you <3 er: look >3 <4 er: oh what do you call it / >4 mirror

@ <1 hand gesture: pointing with left hand at left eye >1

@ <2 hand gesture: with left hand >2

@ <3 hand gesture: with left hand >3

@ <4 gaze: looking down >4

Example 1. "Mirror" (HuD3)

The physician starts his word finding with a sequence of typical OCM-behaviors indicating a need for planning and activation – an intermediate pause //, followed by self-repetition *om du om du* (if you if you), followed by an OCM (hesitation) sound *ö:* (er:) accompanied by a hand gesture and a verb *tittar i* (look). This is followed by another OCM (hesitation) *e:* (er:), an interjection *å:h* (oh) displaying frustration, an OCM phrase *va heter det* (what do you call it) and a short pause all of which are accompanied by gaze aversion. When the desired word *spegel* (mirror) is found, gaze is returned to the patient.

The example shows that word finding often requires effort and might encroach on the valuable consultation time, often limited to 15-20 minutes, and might lead to stress and anxiety for both physician and patient, with regard to both the completeness and correctness of the information. Furthermore, patients often report being nervous and unsure about the physician's ability to understand what they say (Berbyuk, forthcoming). The non-Swedish physicians, in their turn, in spite of reporting the Swedish patients' relative tolerance to their language problems, experience uneasiness, among other things, of such lexical problems as these that

might affect the patient's opinion of their professional competence as well as the patients' confidence in them as health care providers.

### 3.1.2. Substitution of a partially recalled word for another word by moving to a more general concept

The next example illustrates how a physician abandons an attempted word, which is causing problems in favor of another word, which captures more or less the same meaning. The second word *andningsproblem* (breathing problem) signifies a concept, which has the concept signified by the attempted word *andfåddhet* (breathlessness) as a special case. This is a fairly common strategy when specific information is lacking.

Speaker	Transcription	Translation into English
\$D:	vid lätt ansträngning eller vid större ansträngning inget problem	<i>in case of light exertion or more exertion no problem</i>
\$P:	nej	<i>no</i>
\$D:	ingen <1 anf+ >1 e <2 andfådd+ >2 e / <3 and+ >3 // andningsproblem	<i>none &lt;1 breath+ &gt; er &lt;2 breathless+ &gt;2 er / &lt;3 breath+ &gt;3 // breathing problems</i>
	@ <1 cutoff: andfåddhet >1 @ <2 cutoff: andfåddhet >2 @ <3 cutoff: andningsproblem >3	@ <1 cutoff: breathlessness >1 @ <2 cutoff: breathlessness >2 @ <3 cutoff: breathlessness/breathing problem>3
\$P:	< nej >	< no >
	@ < ingressive >	

Example 2. "Breathing problem" (HuD4)

Apparently experiencing both difficulties with recall and pronunciation (displayed by the OCM words *e* (er), the pause // and the self-interrupted words *anf+*(breath+), *andfådd+* (breathless+) and *and+*(breath+)), the physician chooses another word instead of the target one, i.e. *andningproblem* (breathing problem) instead of *andfåddhet* (breathlessness).

Self interrupted words are common in the interactions, often being related to language problems, as in the following example, where the physician recommends re-education to a patient: *kanske kan du få omsorgs+ omskolning* (maybe you can get care+ (omsorgs+)). In other cases, a physician just leaves a self-interrupted word without follow-up, e.g. *de kan också organisera möte med arbetsförmedlingen också med arbetsgivare om*

*det är nödv+* (they can also organize the meeting with employment agency also with employer if it is nec+).

### 3.1.3. Substitution of a word which has been judged inappropriate by moving to a related concept

In the following example the physician becomes dissatisfied with his choice of word and changes it in a more appropriate direction.

Speaker	Transcription	Translation into English
\$D	ska vi göra bak // går de bra	<i>shall we do backwards // is it fine</i>
\$P:	ja	<i>yes</i>
\$D:	<b>säkert</b> // ( bakåt ) är du <b>säkert</b>	<i>sure // (backwards) are you sure</i>
\$P:	ja	<i>yes</i>
\$D:	<b>stabil</b> menar jag	<i>stable i mean</i>
\$P:	nej / går de bara < framåt // > e de bra	<i>no / it goes just &lt; forwards // &gt; is it fine</i>

@ < laughter: P, D >

Example 3. "Stability or certainty" (GerD12)

This example shows a physician using the wrong word, presumably making a semantically associated error, confusing the word *säkert* (certain) with the word *stabil* (stable). One can see that the patient provides an answer confirming his certainty about being able to bend backwards. The physician then corrects himself and inquires whether the patient experiences stability in the body, the patient provides a joking answer, commenting that as long as things move forward it is OK. One can observe that the physician's wrong choice of a word first leads the patient to provide inadequate information, concerning whether he is certain or not and to make a joke to ease the situation.

### 3.1.4. Paraphrasing and abandoning a sought for word

Example 4 illustrates how a word search may be abandoned if the interlocutor's next utterance indicates that the meaning has been conveyed successfully, without the word having been found.

Speaker	Transcription	Translation into English
\$D:	<1 a >1 och de är en e e symtom symptom som e e <2 >2 man kan tänka att om e du e svettig <b>du kanske</b> <b>inte så sta:rk e kanske du</b> <b>e lite lite e:</b>	<1 yeah >1 and this is a er er symptom symptom that er er <2 >2 one can think about if er you are sweaty <b>you maybe are not that</b> <b>stro:ng er maybe you are a little</b> <b>little er:</b>
@ <1 head movement: nods >1 @ <2 sigh >2		
\$P:	ja fick en e m: <1 <2 b >2 vitaminspruta >1	i got a eh mm <1 <2 b >2 vitamin injection>1

Example 4. "Weak" (HuD3)

The Hungarian physician is apparently experiencing problems in explaining the symptom characteristics to the patient and lacking the word *svag* (weak), paraphrases it using *inte så stark* (not that strong) which, in this case, causes no problems with understanding. The interaction continues, no signs of patient's problems with understanding are observed. However, in another case, a lack of understanding is observed, when the physician discusses a low blood count with the patient, advising her that a transfusion is necessary. In example 5, the physician's explanation is not explicit enough and information is omitted which seems to be necessary in order for the patient to understand what is being said.

Speaker	Transcription	Translation into English
\$D:	ja de e lite lågt men e / a <b>de finns flera</b>	yeah is is a little bit low but er / yeah <b>there are many</b>
\$P:	va	what
\$D:	< <b>påsar med</b> >	< <b>bags with</b> >
@ < laughing >		
\$P:	de gör de va	there are aren't they

Example 5. "Bags with blood" (HuD2)

Here, the word search does not result in a paraphrase but in a lack of expressed words. The patient does not understand what the physician means by *flera* (many) and requests explanation. The physician completes the utterance by adding *påsar med* (bags with) meaning bags with blood for transfusion.

### 3.1.5. Substitutioning a word from a related language

In Example 6, the physician solves a word by using the similarity between Swedish and German in order to provide the word he needs. Another way of solving word-finding problems exhibited by the non-Swedish physicians consists in “borrowing” a word from their native languages and/or English. This often presupposes a relative similarity of the physicians’ native language to Swedish, as is the case with, for example, German. Consider the example below:

Speaker	Transcription	Translation into English
\$D:	okej nu skall vi titta / kan du ta byxor bort	<i>okay now we look / can you take trousers away</i>
\$P:	ja	<i>yeah</i>
	och lägga dej på // vad heter det <1(p)ritscha>1 // <2 (p)ritscha >2 // <3 e m >3	<i>and lie down on // what’s it called &lt;1 (p)ritscha &gt;1 // &lt;2 (p)ritscha &gt;2 // &lt;3 er m &gt;3</i>
@ <1 other language: German >1		
@ <2 other language: German >2		
@ <3 sigh >3		
\$D:	på rygg	<i>on the back</i>
\$P:	på rygg ja	<i>on the back yeah</i>

Example 6. “Pritscha- britsen” (GerD12)

Inviting the patient to lie on the examination bed, the physician uses the German noun *Pritsche* that corresponds to Swedish *britsen* (a plank bed). Being aware of an apparent similarity between Swedish and German, the physician uses the German word hoping that the patient will guess what is meant. Due to the fact that this example is taken from the part of the recording when physical examination occurs, it is not possible to know if the physician uses gestures, e.g. deictic gestures. Other examples involve using English words like *voices* and *relaxa* instead of Swedish *röster* and *slappna av*, etc.

### 3.1.6. Using medical terminology when ordinary language terms are lacking

The above-mentioned examples could occur in any social activity. A case that is more specific to medical consultation is the use of medical terminology by the non-Swedish physicians in case of word-finding problems. Using medical terminology in interaction is both helpful and problematic. Being a universal part of medical education, Latin terms are

widely used in medical literature and can be understood by health care personnel, but not necessarily by patients (especially non-chronical ones), who are often unfamiliar with ordinary colloquial names for diseases, symptoms, etc.

Consider the example below, an excerpt from an interaction between a German physician and a patient who had undergone back surgery:

Speaker	Transcription	Translation into English
\$D:	e: // e / (...) trombos oj oj oj oj oj // ja men varför <1 >1 (fusion vad var det) ostabilt 2 < (3< <b>spoldiroristes</b> >3) >2 //	yeah // eh / (...) thrombosis oh oh oh oh oh // yes but why <1 >1 (fusion what was that) unstable 2 < 3 < ( <b>spoldiroristes</b> ) >3 //
	@ < 1 gaze stop: looking down in the papers and reading >1 @ < 2 hand movement: waving illustrating instability >2	
	@ <3 SO: spondylolistes >3	@ <3 SO: spondylolisthesis >3 <sup>3</sup>
\$P:	<1 ja hänger inte me >1 <2 // >2	<1 i don't follow >1 <2 // >2
	@ <1 head movement: shake >1 @ <2 laughter >2	
\$D:	<1 gör du <2 inte >2 >1	<1 you do <2 not >2 >1
	@ <1 laughter: P >1 @ <2 gaze: looking in the papers >2	

Example 7. "Spoldiroristes" (GerD12)

As we can see, the physician's use of a medical term, the name of the disease, together with its poor pronunciation, causes lack of understanding in the interaction. One can also observe that the physician uses a hand gesture, apparently for the patient to distinguish what the physician means.

### 3.1.7. Use of deictic and iconic gestures to supplement verbal information

The example mentioned above, apart from illustrating the problems with understanding medical terms by the patient, shows the importance of using body movements as an aid in solving word-finding problems.

<sup>3</sup> Spondylolisthesis is a condition in which one vertebra slips on another, causing low back pain (Dawson 2002).

<b>Speaker</b>	<b>Transcription</b>	<b>Translation into English</b>
\$D:	[1 < du har ]1 opererat här >	[1 you < had ]1 surgery here >
<b>@ &lt; hand gesture: right hand on back &gt;</b>		
\$P:	< opererat ryggen ja >	< back surgery yeah >
<b>@ &lt; hand gesture: right hand on back &gt;</b>		
\$D:	ja	yeah
\$P:	fjärde femte	fourth fifth
\$D:	varför //	why //
\$P:	ja [2 de var väl ]2 ostabilt	yeah [2 it was unstable i suppose ]2
\$D:	[2 va var de ]2	[2 what was it ]2
\$D:	< ostabilt >	< unstable >
<b>@ &lt; head movement: nod &gt;</b>		
\$P:	ja	yeah
\$D:	< okej >	< okay >
<b>@ &lt; head movement: nod &gt;</b>		
\$P:	de e ju stelopererat [3 (...)]3 ja	the joints are fused [3 (...)]3 yeah
\$D:	[3 < det menar jag > ]3	[3 < that what i mean > ]3
<b>@ &lt; head movement: nod &gt;</b>		
\$D:	ostabilt <1 det <2 slider >2 så // främre >1 // de heter <3 (spoldirolistes) >3	unstable <1 it <2 flies >2 like this // front >1 // it is called <3 (spoldirolistes) >3
<b>@ &lt;1 hand gesture: right hand in the air doing a sliding gesture &gt;1</b>		
<b>@ &lt;2 SO: glider &gt;2</b>		<b>@ &lt;2 SO: glides &gt;2</b>
<b>@ &lt;3 SO: spondylolistes, hand gesture: pointing at P with right hand &gt;3</b>		<b>@ &lt;3 SO: spondylolisthesis, hand gesture: pointing at P with right hand &gt;3</b>
\$P:	< m >	< m >
<b>@ &lt; head movement: nod &gt;</b>		
\$P:	okej < // >	okay < // >
<b>@ &lt;laughter: D &gt;, &lt;facial gesture: P smiles &gt;</b>		

Example 8. "Spoldiroristes" (GerD12)

The gestures used, i.e. both physician and patient putting their hands on their backs more or less at the same time, the physician's gesture showing the instability of the spine by performing a sliding gesture as well as the patient nodding, that in a way indicates active listening, are all ways to handle the lack of understanding.

#### a. Use of deictic gesture to supplement verbal information

Body language and body contact are important parts of communication in health care between health care providers and patients, e.g. using deictic gestures in order to point at the part of the body where a problem occurs is a common way for a patient to show where his/her pain is localized as well

as for the physician to make clear to the patient what part of the body is concerned. Consider the example below:

Speaker	Transcription	Translation into English
\$D:	och e känsel // < har du nån // känsel på nedre [1 <b>extremiteter</b> ]1 > ingenting [2 alls ]2	<i>and er sensibility // &lt; have you some // sensibility on the lower [1 <b>extremities</b> ]1 &gt; nothing at [2 all ]2</i>
<i>@ &lt; hand gesture: D points at the lower part of P's body &gt;</i>		
\$P:	[1 nej ]1	<i>[1 no ]1</i>
\$P:	[2 lite ]2 beröring e	<i>[2 slight ]2 touching er</i>
\$D:	lite [3 beröring ]3	<i>slight [3 touching ]3</i>
\$P:	[3 emellanåt ]3 då	<i>[3 occasionally ]3 so</i>

Example 9. "Extremities" (IraD6)

Using a deictic gesture by the physician is helpful for the patient to understand what part of the body the physician is talking about. Furthermore, apart from making the message more specific, the gesture used has a clarifying function. The term *extremiteter* (extremities) is an uncommon word in spoken Swedish, and its usage without an accompanying deictic gesture might have been problematic for the patient to understand.

The gestures are also used for triggering the recalling of words in interaction:

Speaker	Transcription	Translation into English
\$D:	har du <1 <2 <b>brygg+</b> >2 [1 e]1 >1	<i>have you &lt;1 &lt;2 <b>brid+</b> &gt;2 [1 er ]1 &gt;1</i>
<i>@ &lt;1 hand gesture: right hand moving up towards her mouth &gt;1 @ &lt;2 cutoff: brygga &gt;2</i>		
\$P:	[1 a ]1 de har ja >	<i>[1 yes ]1 i have</i>
\$D:	<b>brygga</b> ja	<i><b>bridge</b> yeah</i>

Example 10. "Bridge" (HuD2)

The deictic gesture used by the physician both helps the patient to understand what is meant as well as helps the physician to find the word *brygga* (bridge).

### **b. Use of manual iconic gesture to supplement verbal information**

Apart from deictic gestures, iconic gestures are used as well. In the example below (the excerpt from the same interaction as example 9 above),

the iconic gesture is used by the Iranian female physician to illustrate the movement of a wheelchair asking the patient about his mobility at work:

Speaker	Transcription	Translation into English
\$D:	m < // > e1 jobbet är anpassat	m < // > er the job is adjusted
\$P:	a	yeah
\$D:	< ja du kan >	< yeah you can >
<b>@ &lt; hand gesture: showing a wheelchair movement with her hand &gt;</b>		
\$D:	ah	yeah

Example 11. “Wheelchair” (IraD6)

By illustrating the movement of a wheelchair, the physician attempts to get the necessary information from the patient, which the patient can then provide. The example shows how a gesture can completely replace a vocal message.

Apart from in task-focused exchanges where the participants discuss health-related issues, gestures are also used in situations when the non-Swedish physicians attempt to create a more personal relationship with their patients by providing psychological support and, having informal conversation, showing apprehension and understanding. As mentioned by the participants in interviews and questionnaires, this type of interaction is often more problematic than the medical type, requiring both linguistic and cultural competence (Berbyuk, forthcoming). In the example below, the physician notices the patient’s stress about the forthcoming physical examination and his bewilderment with the physician’s language, and therefore attempts to console him:

Speaker	Transcription	Translation into English
\$D:	mhm // okej // < ja då ska jag undersöka dej > liten	<i>mhm // okay // &lt; yeah well i will examine you &gt; a little</i>
@ < gaze: looking down in the papers >		
\$P:	m	<i>m</i>
\$D:	å0 du är färdig <1 // <2 de var inte så >2 farlig <3 <b>du är så</b> >3 <4 /// >4 frukta de <5 inte >5 <6 vi <7 <b>bit+</b> >7 <b>biter</b> inte >6 // <8 vi sprutar dej >8 inte >1	<i>and you are ready &lt;1 // &lt;2 it was not that &gt;2 dangerous &lt;3 <b>you are so</b> &gt;3 &lt;4 /// &gt; be afraid of it &lt;5 not &gt;5 &lt;6 we don't &lt;7 <b>bit+</b> &gt;7 <b>bite</b> &gt;6 // &lt;8 we don't inject you &gt;8 &gt;1</i>
@ <1 laughing >1		
@ <2 hand gesture: pointing at P with right hand >2		
@ <3 hand gesture: both hands waving >3		
@ <4 gaze: looking down in the papers >4		
@ <5 hand gesture: both hands waving >5		
@ <6 hand gesture: illustrating biting with left hand >6		
@ <7 cutoff: biter >7		
@ <8 hand gesture: illustrating an injection with a syringe >8		
\$P:	de blir säkert	<i>surely</i>
\$D:	vi bara pratar och // och undersöker liten och // försöker att hjälpa dej	<i>we just talk and // and examine you little and // try to help you</i>

Example 12. "Scared patient" (GerD12)

This example reflects the difficulties experienced by the foreign physician in a case where it is necessary to console the patient. This is also discussed in other studies on foreign physician-native patient interaction (Fiscella et al., 1997). The cutoff *bi+ biter* (bi+bites) as well as the long pauses reflect the low tempo of the physician's speech and his language difficulties. Iconic (functional) and deictic gestures help the physician to illustrate what is meant. Functioning as support for verbal expression, the gestures subsequently facilitate a better understanding in interaction. The example above might also reflect the Lexical Retrieval Hypothesis (Rauscher et al. (1996) that claim that gestures help to activate the lexical retrieval process, i.e. the "biting" iconic functional gesture retrieves the word *biter* (bites).

### c. Use of holistic iconic gesture to supplement information

Not only hand movements but also more complete involvement of the whole body is used by non-Swedish physicians to illustrate what is meant as in the example below:

Speaker	Transcription	Translation into English
\$D:	< du kan gå också normal >	< you can walk also normal >
	@ < head movement: nod >, < hand gesture: right hand circling in the air >	
\$P:	ja	yeah
\$D:	< inte att du <b>plötslit å</b> >	< not that you <b>suddenly and</b> >
	@ < body movement: showing how to stumble and fall >	
\$P:	jo om ja sitter still	yes if i sit still

Example 13. “Stumble and fall”

Not knowing how to explain stumbling and falling in Swedish and attempting to show to the patient what he means, the physician uses the iconic body movement to show the process of stumbling and falling that might result from the stiffness of the body.

### 3.2. *Strategies used by non-Swedish physicians for handling word finding difficulties in perception/understanding*

#### 3.2.1. The physician displays lack of understanding of a term and is given an explanation by a patient

As we can see, in case of word-finding problems, the non-Swedish physicians attempt to recall the words they need, use medical terminology, native language as well as English as recourses in combination with gestures. In addition, physicians get help from their communicative partners, patients, patients’ relatives or health personnel (if present). As mentioned above, Swedish patients interacting with non-Swedish physicians are reported by the latter to be helpful and tolerant (Allwood, Berbyuk & Edebäck, 2004). This can be partially explained by such traces in Swedish culture as tolerance, conflict avoidance, and fear of confrontation (ibid, Lewis, 2004). In the interaction, the patient is shown providing an explanation when the physician does not understand the word as in the following case:

Speaker	Transcription	Translation into English
\$P:	opererat för nageltrång också	<i>surgery for an ingrowing toenail also</i>
\$D:	< opererat >	< <i>surgery</i> >
@ < <i>inquiring</i> >		
\$P:	<b>stortårna</b>	<b><i>big toes</i></b>
\$D:	aha // var de nageln som går i < okej >	<i>oh // was it the nail that goes in &lt; okay &gt;</i>
@ < <i>quiet</i> >		
\$P:	ja	yes

Example 14. “Ingrowing toenail” (IraD5)

The patient noticing the physician’s problem with understanding the word *nageltrång* (ingrowing toenail) provides a hint *stortårna* (big toe) as well as a confirmation to the physician’s inquiry.

### 3.2.2. The physician displays lack of competence concerning choice of a term and a patient supplies correct term

#### *a. The patient supplies a word after an explicit question*

In the example below, the physician is unsure about how to write in the patient’s journal concerning a recent pregnancy and a newborn baby. The patient supplies word after physician’s explicit question:

Speaker	Transcription	Translation into English
\$D:	< har fått // hur säger man > barn flicka nej vad säger man	< <i>have got // how do you say</i> > <i>baby girl (nanny) no how do you say</i>
@ < <i>hand gesture: D is writing</i> >		
\$P:	<b>en flicka</b>	<b><i>a girl</i></b>
\$D:	flicka	<i>girl</i>
\$P:	ja	yes

Example 15. “Baby girl” (GerD12)

The physician first uses the word “barn flicka,” that literally means “child girl,” but is also a common term for “nanny.” The physician is unsure if the word chosen is correct and asks the patient for help, which the latter provides.

***b. The patient supplies a word as a correction of what the physician has said***

An interesting example of a patient correcting a physician is provided below, an excerpt from the interaction between an Iranian oculist and his Swedish male patient. After the operation, the physician asks the patient about his eyesight and the patient reports that the left eye functions better for short distance and the right – for long distance:

<b>Speaker</b>	<b>Transcription</b>	<b>Translation into English</b>
§P:	<1 de här funkar på >1 nära avstånd bäst [1 inte på ]1 långt avstånd <2 de funkar på långt avstånd <3 bäst >3 men inte på nära >2 / <4 så att dom >4	<1 this eye functions >1 best at short distance [1 not at ]1 long distance <2 this functions at long distance <3 best >3 but not at short >2 / <4 so that they >4
@	<1 hand gesture: points at left eye >1 <2 hand gesture: points at right eye >2 <3 head movement: nods >3 <4 hand gesture: P and D move both hands forth and back >4	
§D:	[1 jaha ]1	[1 aha ]1
§D:	<b>kombinerar</b>	<b>combine</b>
§P:	<1 ja dom <b>kompletterar</b> varandra väldigt <2 bra >2 >1	<1 yeah they <b>complement</b> each other very <2 well >2 >1
@	<1 hand gesture: D puts on some glasses on P >1 <2 giggling >2	
§D:	<1 <2 okej >2 <3 // >3 <4 förlåt >4 // >1 <5 om du tittar på tavlan >5 där borta	<1 <2 okay >2 <3 // >3 <4 sorry >4 // >1 <5 if you look at the board >5 right there
@	<1 hand gesture continued: D puts on some glasses on P >1 <2 quiet >2 <3 laughter: P >3 <4 gaze: D looks at the board >4 <5 gaze: D looks at the board >5	

*Example 16. “Combine or complement” (IraD9)*

The physician attempts to complete the patient’s utterance saying the word *kombinerar* (combine). In the subsequent utterance, the patient implicitly corrects the physician saying that the eyes *kompletterar* (complement) each other very well. The physician’s confusion can be noticed in his saying *förlåt* (sorry) preceded by a long pause.

***c. The patient supplies a word the physician can not retrieve***

Using Swedish as a foreign language often results, as mentioned above, in the physician keeping a slow tempo in interaction. In some cases, when the

patients notice the language problems of the physicians, they tend to complete the physicians' utterances, like in the example below:

Speaker	Transcription	Translation into English
\$D:	a du har // en < // >	yeah you have // a < // >
@ < laughter: D, P >		
\$P:	< en urinvägsinfektion som heter duga ja // >	< a really bad urinary infection // >

@ < laughter: D, P >

Example 17. "Urinary infection" (IraD7)

The pauses in the physician's utterance are interpreted by the patient as uncertainty with language competence resulting in the patient completing the utterance.

### 3.2.3. An accompanying person supplies the physician with a needed term

Apart from patients, patients' relatives (if present) help physicians and patients understand each other. It is often the case with elderly patients, who might have hearing problems that make it additionally difficult to understand the physician's accent (Allwood, Berbyuk & Edebäck, 2004). In the example below, the patient's daughter helps the patient, her father, and the physician:

Speaker	Transcription	Translation into English
\$D:	e:1 <1 >1 förlåt mej men ja måste e fråga dej e om e två saker till / e1 brukar du <2 dricka alkohol eller inte >2	er: < 1 >1 i beg you pardon but i must er: ask you er about two more things / er: do you usually <2 drink alcohol or not >2
@ <1 inhalation sound >1		
@ <2 head movement: nods >2		
\$P:	ja använder inte alkohol	i don't use alcohol
\$D:	okej < e du <b>nykter</b> >	okay < are you <b>sober</b> >
@ < head movement: nod >		
\$P:	< HM >	< HM >
@ < body movement: leans towards D to hear better >		
\$D:	< NYKTER >	< SOBER >
@ < head movement: nods >		
\$C:	< <b>nykterist ja</b> >	< teetotaller yeah >
@ < gaze: P looking at C >		
\$D:	[1 <b>nykterist ja</b> ]1 <b>nykterist</b> [2 okej ]2 < ja >	[1 teetotaller yeah ]1 teetotaler [2 okay ]2 < yeah >

@ < head movement: nods >

\$P:	[1 nykterist ]1	[1 teetotaller ]1
\$P:	[2 ö: ]2	[2 e:r ]2
\$P:	sen nittonhundrafemtitre	since nineteen fifty three
\$D:	< okej >	< okay >
@ < head movement: nods >		
\$P:	har ja inte använt sprit	i have not used alcohol
\$D:	oj	wow
\$P:	annat än i medicinskt bruk	in other way than in medical use
\$D:	<1 a visst visst visst >1	< 1 yeah sure sure sure >1
@ <1 head movement: nods >1		

*Example 18. "Sober" (HuD3)*

The physician's use of the wrong word form creates bewilderment and lack of understanding from the patient's side, which might think that the physician asks if he is *nykter* (sober) at the moment of consultation. The patient's relative (participant \$C) understands what is meant and provides the correct form of the word. It is plausible that the patient's gaze directed at the relative at that moment could be interpreted as a request for help.

In this example, an interesting observation can be made about Swedish culture. One of the subjects that the interview and the questionnaire respondents among health care personnel and physicians consider to be sensitive in communication with Swedish patients is a problem relating to alcohol consumption (ibid). The physician's careful way of asking the question about alcohol by introducing it with "pardon me" shows his awareness of the sensitivity of the topic. When the lack of understanding occurs, the physician becomes apparently stressed about it. It can subsequently be noticed in his feedback "yeah sure sure sure" as well as a supportive head nodding, after the patient's narrative about being a teetotaler.

#### 4. Summary and Discussion

The article has presented several examples of word-finding problems exhibited by non-Swedish physicians in communication with Swedish patients and strategies used for handling them. Some of the problems and strategies discussed are mispronunciation, paraphrasing, choosing another word, choosing the wrong word, etc. Leaving a word out is another observed strategy. Furthermore, there is the use of medical terminology and word borrowing from other languages, e.g. from the physician's native language or from English, to solve the problem. Body movement seems to be helpful for triggering a sought word (Lexical Retrieval Hypothesis).

Body movements (often in the form of manual gestures) can also have a clarifying function being used to support a verbal message. They might also be used as substitutes for a verbal message. In some cases, the interlocutors of non-Swedish physicians, i.e. the patients or their relatives help the physician to find a word or an explanation of an unknown word with or without the physician's request. They might also complete utterances and correct incorrect words. In diagram 1 below, we give an overview of the ways of handling word-finding problems that have been observed in the data.

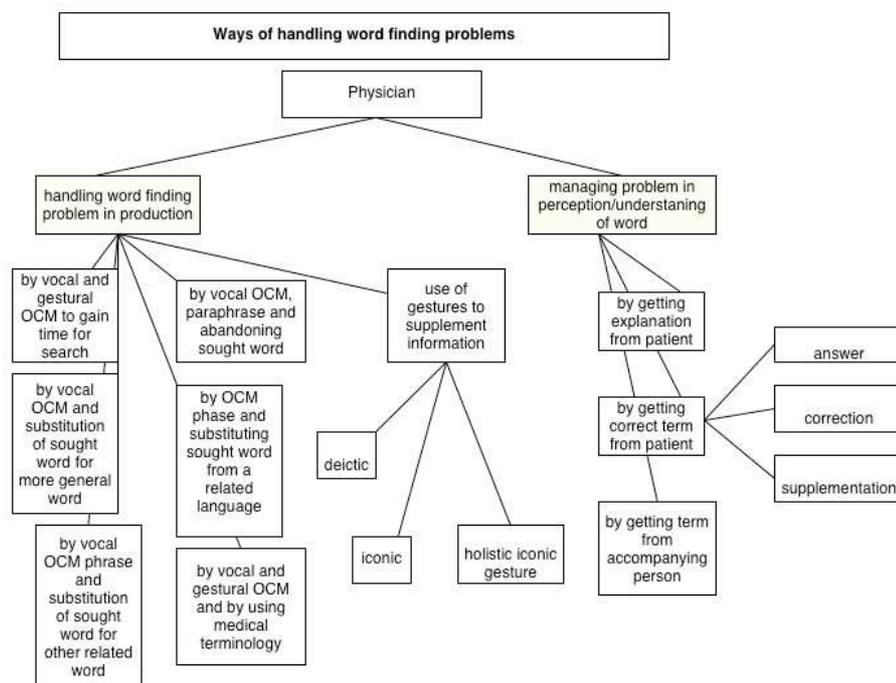


Diagram 1. Ways of handling word finding problems

Some quantitative data on the different ways of handling word finding problems in 33 medical consultations is presented in Table 3 below.

Table 3. Frequency of different ways of handling word finding difficulties

#	Ways of handling word finding difficulties	Number of occurrences
<b>The physician handles word finding problem in production</b>		
1.	by vocal and gestural OCM	23
2.	by vocal OCM and substitution of sought word for more general word	2
3.	by vocal OCM phrase and substitution of sought word for related word	2
4.	by vocal OCM, paraphrase and abandoning sought word	8
5.	by OCM phrase by substituting sought word from a related language	5
6.	by vocal and gestural OCM and by using medical terminology	3
7.	use of gestures to supplement information	
<i>a</i>	<i>deictic</i>	4
<i>b</i>	<i>particular iconic</i>	8
<i>c</i>	<i>holistic iconic</i>	1
<b>The physician handles problem in perception/understanding of word</b>		
8.	by getting explanation from patient	1
9.	by getting correct term from patient	2
<i>a</i>	<i>answer</i>	1
<i>b</i>	<i>correction</i>	1
<i>c</i>	<i>supplementation</i>	2
10.	by getting term from accompanying person	1
<b>Total:</b>		<b>64</b>

As we can see, in the majority of cases the physicians successfully recall the words they need themselves. The most common solution is to recall the word using vocal and gestural OCM (23 occurrences), to use gestures to supplement information (13 occurrences total) as well as to use vocal OCM, paraphrase and abandoning sought word (8 occurrences). Other strategies such as handling the lexical problem through the use of an OCM phrase, by using a word from a related language (5 occurrences), by vocal OCM and substitution of sought word for a more general word (2 occurrences) or a related word (2 occurrences) are represented as well. The physicians also use medical terminology in case of word finding problems (3 occurrences). The physicians rarely get help from their patients and patients' relatives, i.e. only in nine (9) occurrences out of 64, the patient or the patient's relative are directly involved and help the physician.

We, thus, see that the non-Swedish physicians use of Swedish as a foreign language can be a negative factor for interaction. An increased uncertainty in interaction, when both physician and patient are unsure if they understand each other correctly, might result in stress and frustration from both sides. Our study also shows that foreign physicians might be anxious about the "fear of patient bias," i.e. nervousness about a patients' underestimation of their medical training and competence received outside

of Sweden, i.e. that the patients are fearing inferior medical care (Allwood, Berbyuk, Edebäck 2004). Language problems are an additional factor that increases the physician's anxiety. The physician's lack of language competence can influence the patient's perception of the physician's interpersonal skills, i.e. the physician's problems with informal conversation might result in patients not feeling comfortable and secure. Especially, in sensitive situations, word-finding problems can have a negative impact on interaction.

The examples of interactions between the Swedish patients and non-Swedish physicians show that the latter often solve the problems themselves. In a few cases, the patients' help can be observed. That this does not happen more often might be the result of activity influence, i.e. the physician's dominant role might keep the patient from correcting physician's language.

The involvement of the patients in the cases observed above can be explained by the necessity to obtain the information they need during consultation as well as by the short power distance between physician and patient in Sweden reflected in a relatively low Power Distance Index in Hofstede's taxonomy of cultural patterns (Hofstede 2001). In spite of the fact that the patients' help, as reported by the non-Swedish physicians in the interviews, is often seen as positive factor, one should not underestimate the fact that the majority of the non-Swedish physicians involved in the study are used to a larger power distance than the one typical for Sweden. The help of the patients might result in the physicians' feeling of losing face, that might probably not often be expressed, but nevertheless be an experienced feeling.

The examples of the word finding problems and their handling also indicate the multimodality of the communication between non-Swedish physicians and their Swedish patients. The physician's use of gestures necessitates the patient's involvement and attention in order to understand what the physician means. Thus, the physicians' "weakness" in terms of language can result in increased involvement and participation of patients in interaction - something that is often viewed as a positive factor.

## References

- Allwood, J., Ahlsén, E., Lund, J., and Sundqvist, J. (in this volume) Multimodality in own communication management.
- Allwood, J., Berbyuk, N., & Edebäck, C. (2004). Obruten tanke (Unbroken Thought). *Invandrare & Minoriteter (Immigrants & Minorities)*, 5-6, pp. 22-26.
- Berbyuk, N., Allwood, J., & Edebäck, C. (2004). Being a non-Swedish physician in Sweden: A comparison of the views on work related communication of non-Swedish physicians and Swedish health care personnel. In: Allwood, J. & Dorrietz, B. (eds). *Intercultural Communication at Work. Selected papers from the 10th NIC Symposium on Intercultural Communication*, Department of Linguistics, Gothenburg University. Also available in: *Journal of Intercultural Communication* (7) <http://www.immi.se/intercultural>
- Berbyuk, N. (2005). Intercultural communication in health care systems: Non-Swedish physicians in Sweden. In: Gunnarsson, B-L. (ed). *The immigrant and the workplace*. FUMS, Institutionen för nordiska språk, Uppsala universitet. pp.47-65.
- Berbyuk, N. "Communication and interaction in multicultural Swedish health care" (preliminary title) PhD thesis (forthcoming).
- Davidson, B. (2001). Questions in cross-linguistic medical encounters; the role of the hospital interpreter. *Anthropological Quarterly*, Vol. 74 Issue 4, 170-79.
- Dawson, Edgar, D., 2002: 'Spondylolisthesis. What is it?' Retrieved 10 November 2004, from the World Wide Web: <http://www.spineuniverse.com/displayarticle.php/article1430.html>
- Demeurisse, G. (1999). Communication with the aphasic patient. *Rev Med Brux*, 20(4), A268-270.
- Ekström, U., Oskarsson, A. (2004). *Slutrapport projekt utländska läkare. Främjande av etnisk mångfald i arbetslivet*. Europeiska Socialfonden Mål 3, Göteborgs Stad, Integrationsverket, Länsarbetsnämnden Västra Götaland, Sahlgrenska Akademin, Socialstyrelsen, Västra Götalandsregionen.
- Erickson, F., Rittenberg, W. (1987). Topic control and person control: A thorny problem for foreign physicians in interaction with American patients. *Discourse Processes*, 401-415.

- Fernandez, A., Schillinger, D., Grumbach, K., Rosenthal, A., Stewart, A. L., Wang, F., et al. (2004). Physician language ability and cultural competence. An exploratory study of communication with spanish-speaking patients. *J Gen Intern Med*, 19(2), 167-174.
- Fiscella, K., Roman-Diaz, M., Lue, B. H., Botelho, R., & Frankel, R. (1997). 'Being a foreigner, I may be punished if I make a small mistake': Assessing transcultural experiences in caring for patients. *Fam Pract*, 14(2), 112-116.
- Educational Commission for Foreign Medical Graduates (ECFM) (1976). *Handbook for foreign medical graduates*.
- Hofstede, G. (2001). *Culture's consequences: comparing values, behaviors, institutions, and organizations across nations*. Thousand Oaks, Sage.
- Lewis, R. D. (2004). *When cultures collide. Managing successfully across cultures*. London: Nicholas Brealey Publishing.
- Lindberg, L. (2005). Socialstyrelsens statistik över legitimerade läkare och sjuksköterskor med svensk resp. Utländsk utbildning 2004. In N. Berbyuk (Ed.) (PM, Förhandlingsdelegationen): *Sveriges Kommuner och Landsting*, Avd. för Lärande och Arbetsmarknad.
- McEwen, E., & Anton-Culver, H. (1988). The medical communication of deaf patients. *J Fam Pract*, 26(3), 289-291.
- McGrath, B. P. (2004). Integration of overseas-trained doctors into the Australian medical workforce. *Med J Aust*, 181(11-12), 640-642.
- Mejia, A. (2004). Migration of physicians and nurses: A world wide picture. 1978. *Bull World Health Organ*, 82(8), 626-630.
- Miller, E. A., Laugesen, M., Lee, S. Y., & Mick, S. S. (1998). Emigration of New Zealand and Australian physicians to the United States and the international flow of medical personnel. *Health Policy*, 43(3), 253-270.
- Nivre et al. (2004). Göteborg transcription standard. V. 6.4. Department of Linguistics, Göteborg University.
- Rauscher, F.H., Krauss, R.M., & Chen, Y. (1996). Gesture, speech and lexical access: the role of lexical movements in speech production. *Psychological Science*, 7, 226-230.
- Steward, D. E. (2003). The internal medicine workforce, international medical graduates, and medical school departments of medicine. *Am J Med*, 115(1), 80-84.

## Biography

**Jens Allwood** is since 1986 professor of Linguistics at the department of Linguistics at Göteborg University. He is also director of the interdisciplinary cognitive science center SSKKII at the same university. His research primarily includes work in semantics and pragmatics. He is investigating spoken language interaction from several perspectives, e.g. corpus linguistics, computer modelling of dialog, sociolinguistics and psycholinguistics as well as intercultural communication. Presently he is heading projects concerned with the semantics of spoken language phenomena, multimodal communication, cultural variation in communication and the influence of social activity on spoken language.

**Natasha Berbyuk** is a PhD student at the Department of Linguistics and SSKKII at Göteborg University. Her research interests include intercultural communication, multimodal communication, pragmatics, corpus linguistics, and gender studies. Currently she is involved in a multidisciplinary research project about intercultural communication in the Swedish health care system.

## Author's address

*Jens Allwood  
Göteborg University, Department of Linguistics  
Box 200  
S-405 30 Göteborg  
Sweden  
phone: +46 31 773 1867/1923  
e-mail: jens@ling.gu.se*

*Nataliya Berbyuk, MA  
PhD student  
Dept of Linguistics  
Box 200  
405 30 Göteborg, Sweden  
phone: +46 31 773 5213  
fax: +46 31 773 4853  
e-mail: natasha@ling.gu.se  
<http://www.ling.gu.se/~natasha>*

# THE MUMIN ANNOTATION SCHEME FOR FEEDBACK, TURN MANAGEMENT AND SEQUENCING

*Jens Allwood (1), Loredana Cerrato (2), Kristiina Jokinen (3), Costanza Navarretta (4), and Patrizia Paggio (4).*

(1) University of Göteborg, (2) TMH/CTT, KTH, Sweden  
(3) University of Helsinki, (4) CST, University of Copenhagen

## Abstract

*This paper deals with the MUMIN multimodal annotation scheme (Allwood et al 2004), which was developed for the study of gestures and facial displays in interpersonal communication, with particular regard to the role played by multimodal expressions for feedback, turn management and sequencing. The scheme has been applied to the analysis of multimodal behaviour in short video clips in Swedish, Finnish and Danish. Preliminary results obtained in this study show that the categories defined in the scheme are reliable, and that the scheme as a whole constitutes a useful analysis tool in the study of multimodal communication behaviour.*

**Keywords:** Multimodal corpora, annotation, non-verbal expressions for feedback.

## 1. The MUMIN annotation scheme

The creation of annotated multimodal corpora is being recognised by a growing number of researchers, initiatives and organisations<sup>1</sup> as a prerequisite for the creation of more natural human-computer interfaces

---

<sup>1</sup> A long list of projects, initiatives and organisations that have addressed the issue is provided in Martin *et al* (2004).

based on models of human behaviour. However, there is still a lack of agreement as to what a general multimodal annotation scheme should look like, how it should be implemented, applied and evaluated. In this paper, we discuss the multimodal annotation scheme that has resulted from the collaborative effort of a group of researchers from the Nordic Network on Multimodal Interfaces MUMIN ([www.cst.dk/mumin](http://www.cst.dk/mumin)) and its application to the annotation of multimodal communication in video clips in Swedish, Finnish and Danish.

The construction of a multimodal corpus often reflects the specific requirements of an application and thus constitutes an attempt at modelling either input or output multimodal behaviour. An example of the former may be trying to foresee how the user combines voice and pen input in the scenario targeted by the system; an example of the latter to model how eyebrow movements and vocal expressions should be coordinated in a talking head. The MUMIN coding scheme, on the contrary, is not based on a set of system requirements, but is rather intended as a general instrument for the study of hand gestures and facial displays in interpersonal communication, in particular the role played by multimodal expressions for feedback, turn management and sequencing. It builds on previous studies of feedback strategies in human conversations (Clark & Schaefer 1989, Allwood *et al* 1992), and on recent work where vocal feedback has been categorised in behavioural or functional terms (Allwood 2001, Allwood & Cerrato 2003, Cerrato 2004).

Two kinds of annotation are considered. The first is modality-specific, and concerns the expression types, the second concerns multimodal communication. For each gesture<sup>2</sup> taken into consideration, a relation with the corresponding speech expression (if any) is also annotated. Note that in a dialogue, a gesture by one person may relate to speech by another. The main focus of the coding scheme is the annotation of feedback, turn-management and sequencing functions of multimodal expressions, as well as the way in which expressions belonging to different modalities are combined.

Focusing on these functions has several consequences for the way in which the coding scheme is constructed. First of all, the annotator is expected to *select* gestures to be annotated *only* if they play an observable communicative function. This means that not all gestures need be

---

<sup>2</sup> We use “gesture” as a general term for non-verbal expressions, in our case hand gestures and facial displays.

annotated, and that quite a number of them in fact will not be. For example, mechanical recurrent blinking of the eyes due to dryness will not be annotated because it does not have a communicative function. Another consequence of the focus we have chosen is that the attributes that have been defined to annotate the shape or dynamics of a gesture are not very detailed, because they only seek to capture features that are significant when studying interpersonal communication. While this is a reasonable limitation in a functional study of communication behaviour, the resulting annotation will not provide the necessary details regarding the shape and timing of gestures for applications where a precise morphological definition is essential, for instance as a basis for the design of a talking head. However, the annotation of gesture shape and dynamics can be extended for specific purposes without changing the functional level of the annotation, which is useful also in such applications, since it provides valuable information on when and why certain types of non-verbal behaviour should be generated.

In what follows we will first present the categories defined in the coding scheme, we will then describe the coding procedure and the materials used in our experiments, report the results obtained in two different case studies, and finally provide a general conclusion on the usefulness and potential applications of the scheme.

## **2. Annotation categories**

The specific annotation categories and corresponding tags that make up the coding scheme are given in Allwood *et al* (2004). In what follows, we will describe them briefly starting with the functional categories.

### ***2.1 Categories of feedback, turn management and sequencing***

The main purpose of the annotation is to capture the way in which facial displays and hand gestures, possibly in combination with verbal expressions, contribute to the general communicative phenomena of *feedback (give or elicit)*, *turn management* and *sequencing*. These three functions constitute the backbone of the scheme, and are intended to guide the selection of the gestures to be annotated. In defining the features for the annotation of feedback, turn management and sequencing, we have profited from an extensive number of references in which these phenomena are

treated from the point of view of verbal expressions. We believe the features in the coding scheme are applicable to the annotation of non-verbal and multimodal expressions for which they have been designed, and the preliminary results described in this paper confirm our belief. However, these results will have to be validated by applying the scheme to more practical coding tasks.

The production of feedback is a pervasive phenomenon in human communication. Participants in a conversation continuously exchange feedback as a way of providing signals about the success of their interaction. They give feedback to show their interlocutor that they are willing and able to continue the communication and that they are listening, paying attention, understanding or not understanding, agreeing or disagreeing with the message which is being conveyed. They elicit feedback to know how the interlocutor is reacting in terms of attention, understanding and agreement with what they are saying. While giving or eliciting feedback to the message that is being conveyed, both speaker and listener can show emotions and attitudes, for instance they can agree enthusiastically, or signal lack of acceptance and disappointment.

Both feedback giving and eliciting are annotated by means of the same three sets of attributes, called *Basic*, *Acceptance*, and *Attitudinal emotions/attitudes*. *Basic* features define the relevant gestures or facial displays in terms of whether they express or elicit:

- Continuation/contact and perception (CP), where the dialogue participants acknowledge contact and perception of each other.
- Continuation/contact, perception and understanding (CPU), where they also show explicit signs of understanding or not understanding of the message conveyed.

The two categories of basic feedback are intended to capture what Clark and Schaefer (1989) call *acknowledgement*, which describes a number of strategies used by dialogue participants to signal that a contribution has been understood well enough to allow the conversation to proceed.

*Acceptance*, which is a boolean feature, indicates that the subject has not only perceived and understood the message, but also shows or elicits signs of either agreeing with its content or rejecting it, e.g. by different head movements. Acceptance is treated as a separate dimension, different from understanding, also in coding schemes for dialogue annotation. For

instance, the DAMSL coding scheme distinguishes between *understanding* (“Huh”, “What?”, “I see”) and *agreement* (“Yes”, “No”, “Sounds good”).

Finally, feedback annotation can rely on a list of *emotions* and *attitudes* that can co-occur with one of the basic feedback features and with an acceptance feature. It includes the six basic emotions described and used in many studies (Ekman 1999, Cowi 2000 and Beskow *et al* 2004) plus others that we consider interesting for feedback, but for which there is less general agreement and less reliability. It is intended as an open and rather tentative list. Table 1 shows the feedback giving features: those for feedback eliciting are practically identical.

Table 1. Feedback giving annotation features

Function attribute		Function value
<b>FEEDBACK GIVE</b>	Basic	Contact/continuation Perception Understanding (CPU) Contact/continuation Perception (CP)
	Acceptance	Accept Non-accept
	Additional Emotion/Attitude	Happy, Sad, Surprised, Disgusted, Angry, Frightened, Certain, Uncertain, Interested, Uninterested, Disappointed, Satisfied, Other

If feedback is the machinery that crucially supports the success of the interaction in interpersonal communication, the flow of the interaction is also dependent on the turn management system. Optimal turn management has the effect of minimising overlapping speech and pauses in the conversation. Turn management is coded by the three general features *Turn gain*, *Turn end* and *Turn hold*. An additional dimension concerns whether the turn changes in agreement between the two speakers or not. Thus, a gain in turn can either be classified as a *Turn take* if the speaker takes a turn that was not offered, possibly by interrupting, or a *Turn accept* if the speaker accepts a turn that is being offered. Similarly, the end of a turn can also be achieved in different ways: we can have a *Turn yield* if the speaker releases the turn under pressure, a *Turn elicit* if the speaker offers the turn to the interlocutor, or a *Turn complete* if the speaker signals that they are about to complete their turn while at the same time implying that the dialogue has come to an end. The various features are shown in Table 2.

Table 2. Turn management annotation features

Function attribute		Function value
<b>TURN MANAGEMENT</b>	Turn-gain	Turn-take Turn-accept
	Turn-end	Turn-yield Turn-elicit Turn-complete
	Turn-hold	Turn-hold

Finally, sequencing is a dimension that concerns the organisation of a dialogue in meaningful sequences. The notion of sequence is intended to capture what in other frameworks has been described as sub-dialogues: it is a sequence of speech acts, and it may extend over several turns. A digression, however, may also constitute an independent sequence, which in this case would be included in a turn. In other words, sequencing is orthogonal to the turn system, and constitutes a different way of structuring the dialogue, based on content rather than speaker's turn. Sequencing is described by means of three features. *Opening sequence* indicates that a new speech act sequence is starting, for example in conjunction with a gesture that accompanies the phrase "by the way...". *Continue sequence* indicates that the current speech act sequence is ongoing, for example when a gesture is associated with enumerative phrases such as "the first... the second... the third...". *Closing sequence* indicates that the current speech act sequence is closed, which may be shown by a head turn or another gesture while uttering a phrase like "that's it, that's all".

Under normal circumstances, in face-to-face communication feedback, turn management and sequencing all involve use of multimodal expressions, and are therefore central phenomena in the context of a study of multimodal communication. Note also that these features are not mutually exclusive. For instance, turn management is partly done by feedback. You can accept a turn by giving feedback and you can yield a turn by eliciting information from the other party. Similarly, a feedback expression can indicate understanding and acceptance, or understanding and refusal at the same time. Within each feature, however, only one value is allowed. For example, a feedback giving expression in this coding scheme cannot be assigned accept and non-accept values at the same time.

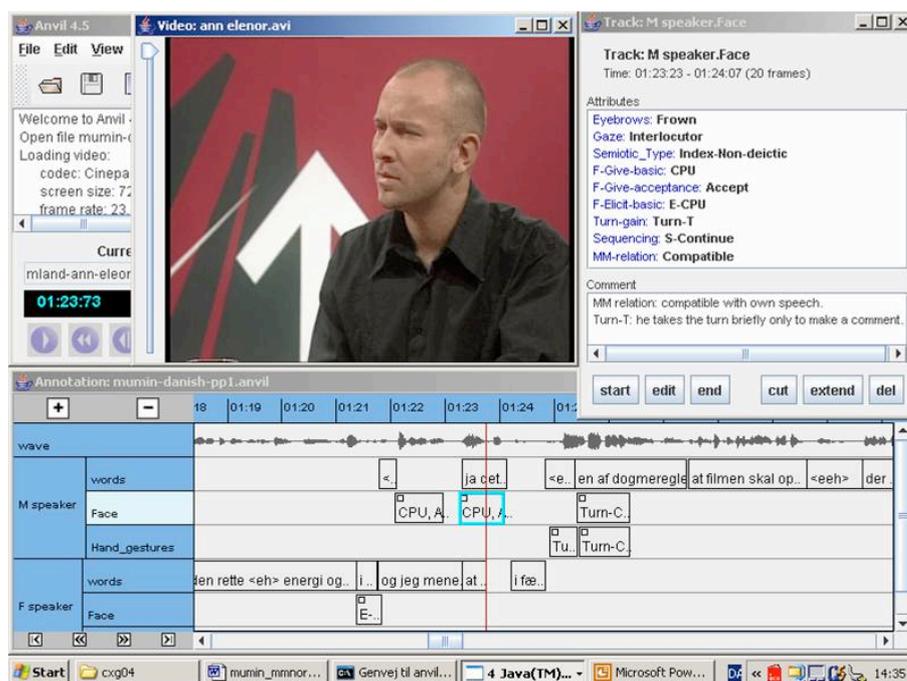


Figure 1. A multifunctional facial display: turn management and feedback

An example of a multifunctional facial display is shown in Figure 1: the speaker frowns and briefly takes the turn while agreeing with the interlocutor by uttering the words: “ja, det synes jeg” (Yes, I think so). By the same multimodal expression (facial display combined with speech utterance) the speaker also elicits feedback from the interlocutor and encourages her to continue the current sequence.

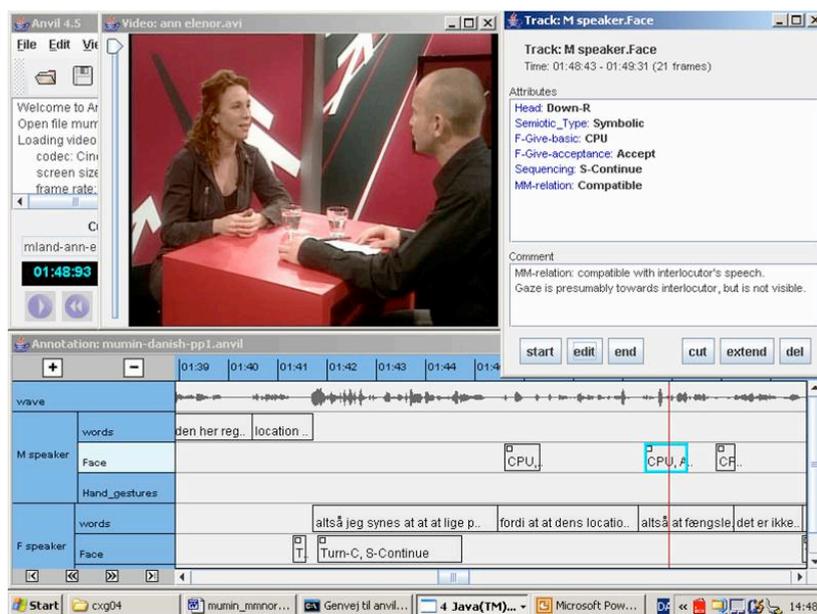


Figure 2. Basic feedback and acceptance by facial expressions

Figure 2 shows a frame of a sequence in which the same speaker nods repeatedly while the interlocutor is speaking, without, however, saying anything. The gesture, which is unfortunately not visible in the single frame, has been annotated as signalling basic feedback and acceptance, at the same time as encouraging the interlocutor to continue the sequence as in the previous example. Concerning the multimodal relation, this gesture is compatible with the interlocutor's speech, while the previous one was related to and compatible with the speaker's own utterance.

## 2.2 *Facial displays and hand gestures*

In addition to the functional categories described in the preceding section, facial displays and hand gestures are also annotated with respect to the shape and dynamics of the movement characterising the gesture. Since a fine-grained characterisation of these aspects is beyond the scope of the coding scheme, the categories we propose are not very detailed. However, they should be specific enough to be able to distinguish and characterise the various non-verbal expressions that play a role in feedback, turn management and sequencing. In particular, they are concerned with the movement dimension of facial displays and hand gestures, and should be understood as dynamic features that refer to a movement as a whole or a protracted state, rather than punctual categories referring to different stages of a movement. The duration of the movement or state is not indicated as an explicit attribute in the coding scheme, but we expect the concrete implementation to indicate start and end point of the gesture, and to ensure synchronisation between the various modality tracks. We also do not consider internal gesture segmentation since it does not seem very relevant for the analysis of communicative functions we are pursuing. However, nothing hinders annotators from extending the scheme in the direction of a more precise characterisation of the dynamics of gestures.

The term *facial displays* refers, according to Cassell (2000), to timed changes in eyebrow position, expressions of the mouth, movement of the head and of the eyes. The coding scheme includes features describing *General face* expressions such as *Smile* or *Scowl*, features of *Eyebrow movements* such as *Frown* or *Raise*, features referring to *Eye movement* such as *Close-both*, or *Extra-open*, features for *Gaze direction*, for movements of the *Mouth* and position of the *Lips*. Finally, a number of features refer to movements of the *Head*. The total number of different features for facial displays is 36.

The annotation of the shape and trajectory of hand gesture is much simplified with respect to other coding schemes, e.g. the scheme used at the McNeill Lab (Duncan 2004) which was our starting point. Features are defined concerning the two dimensions of *Handedness* and *Trajectory*, so that we distinguish between single-handed and double-handed gestures, and among a number of different simple trajectories analogous to what is done for gaze movement. The total number of features is seven. This is of course far from adequate for the physical descriptions of hand gestures that can be quite complex, and can be extended in several ways for different purposes and applications.

In addition to the features relating to shape and dynamics of non-verbal expressions, semiotic categories have also been defined common to both facial displays and hand gestures building on Pierce's semiotic types. They are *Indexical Deictic* and *Non-deictic*, *Iconic* and *Symbolic*.

### **2.3 Multimodal features**

Facial displays and gestures can be synchronized with spoken language and with each other at different levels: at the phoneme, word, phrase or long utterance level. In this coding scheme, the word is the smallest speech segment we expect annotators to annotate multimodal relations. We also assume that different codings can have different time spans. For instance, a cross-modal relation can be defined between a speech segment and a slightly subsequent gesture.

Our multimodal tags are quite simple, and not as numerous as those proposed e.g. by Poggi and Magno Caldognetto (1996). We make a basic distinction between two signs being *dependent* on or *independent* from each other. If they are dependent, they will either be *compatible* or *incompatible*. For two signs to be compatible, they must either complement or reinforce each other, while incompatibility arises if they express different contents, as it often happens in ironic contexts.

## **3. Annotation procedure and material**

The coding procedure was iteratively defined in the MUMIN workshops and steering group meetings. Furthermore, the MUMIN annotators were

given a tutorial on how to annotate by means of the three coding tools ANVIL (Kipp 2001), MultiTool (Gunnarsson 2002) and NITE (Bernsen et al 2002).

Examples of annotations created with the MUMIN coding scheme, and of ANVIL specification files building on this coding scheme, can be inspected at the MUMIN site at [www.cst.dk/mumin](http://www.cst.dk/mumin). The annotated material consists of:

- One minute clip from an interview of the actress Ann Eleanora Jørgensen by Per Juul Carlsen from the Danish DR-TV (Danmarks Radio)
- One minute interview of the finance minister Antti Kalliomäki from the Finnish Aamu-TV (Morning-TV). The video is provided by the courtesy of the CSC (Centre of Scientific Computing).
- One minute clip from the Swedish movie “Show me love”, consisting of an emotional dialog between father and daughter.

Since all of the videos are protected by copyright, they cannot be made publicly available, but examples will be accessible from the MUMIN site.

#### **4. First case study: the Danish annotation**

In the Danish case study two independent annotators with limited annotator experience annotated facial displays and hand gestures in the Danish video clip by means of the ANVIL platform. They started by annotating the non-verbal expressions of one of the interlocutors together to familiarise themselves with the coding scheme. Then they did the annotation task for the other dialogue participant independently in order to evaluate the reliability of the coding scheme.

The annotation has been evaluated based on the strategy described by Carletta *et al* (2004). First of all, a method for aligning the annotations of the coders had to be established: it was decided to accept a difference in time coding of under one fourth of a second per segmentation. In other words, if both coders annotated a gesture within the same time span apart from a possible difference in start and/or end of under  $\frac{1}{4}$  of a second, it was assumed that the two segments described the same expression. In all the cases where both coders annotated the same gesture, there was agreement of segmentation, with the exception of one case in which one coder

recorded one facial display as a unit, while the other split the same display into two (i.e. the two segments in one annotation covered temporally the same time span of one segment in the second annotation).

The first coder annotated 37 facial displays. The second one annotated 33. Of these 29 were annotated by both coders. One was coded by one coder as one segment, while it was split up into two segments by the second coder, as explained previously. The agreement in recognition of facial displays is thus 0.83 (0.86 considering the two split segments as one unit). Concerning hand gestures, the first coder annotated 6 of them, the second 4. Of these only two were in common (0.4 agreement for hand gesture recognition).

The reliability of gesture classification has been measured by means of the kappa-coefficient (Siegel and Castellan 1988). Kappa is calculated as follows:

$$K = (P(A) - P(E)) / (1 - P(E))$$

where  $P(A)$  is the proportion of times the coders agree and  $P(E)$  is the proportion of times one can expect them to agree by chance.  $P(E)$  varies depending on the number of available values that can be assigned to a single feature. For instance, if the annotators can choose between two values,  $P(E)$  will be 0.50. If the values from which to choose are 4,  $P(E)$  will be 0.25 and so on. The value of Kappa is 1 in case of total agreement and zero in case of total disagreement. Generally, a value above 0.6 is considered satisfactory. Below we show the kappa-score obtained for each feature in the facial displays recognised by both coders (29 facials). Table 3 reports the values obtained in the annotation of the shape of the facial display.

Table 4 the values for the feedback features, and Table 5 those obtained for the annotation of turn management, sequencing and multimodal relation. In the first row we indicate the names of the features, in the second row the  $P(A)$  for the values assigned to each feature, in the third row the corresponding  $P(E)$ , and finally in the fourth row we give the kappa-score for each feature.

Table 3. Kappa-score for classification of movement and semiotic type

	<b>General Face</b>	<b>Eye-brows</b>	<b>Eyes</b>	<b>Gaze</b>	<b>Mouth-openness</b>	<b>Mouth-lips</b>	<b>Head</b>	<b>Semiotic type</b>
P(A)	.93	.93	.9	.62	.97	.97	.65	.86
P(E)	.20	.25	.17	.17	.33	.20	.07	.20
Kappa	.91	.91	.88	.54	.96	.96	.62	.83

Table 4. Kappa-score for classification of feedback giving and eliciting

	<b>F-Give-basic</b>	<b>F-Give-acceptance</b>	<b>F-Give-emotion/attitude</b>	<b>F-Elicit-basic</b>	<b>F-Elicit-acceptance</b>	<b>F-Elicit-emotion/attitude</b>
P(A)	.79	.86	.86	.93	1	.93
P(E)	.33	.25	.08	.33	.25	.08
Kappa	.68	.81	.84	.9	1	.92

Table 5. Kappa-score for classification of turn management, sequencing and MM-relation

	<b>Turn-gain</b>	<b>Turn-end</b>	<b>Turn-hold</b>	<b>Sequencing</b>	<b>MM-relation</b>
P(A)	.89	.93	.96	.69	.82
P(E)	.33	.33	.05	.25	.25
Kappa	.83	.89	.92	.59	.76

The kappa-score for the classification of hand gestures was 1 for all features (total agreement). However, it is not possible to draw any conclusion about the encoding of hand gestures, because the data are too limited. Regarding the encodings of facial features, on the other hand, the study allows us to make a few observations. In general, the kappa-score is quite good for all the features, except those for *Gaze* and *Sequencing*.

The reason for the low agreement on gaze features was partly due to the fact that one coder encoded gaze relative to the head position (head up, no gaze), while the other coder chose to annotate the gaze instead of the head

when the head movement was little (no head movement, gaze up). Furthermore, the two coders used different strategies for gaze. In some cases they coded “gaze:side” with the comment “away from the interlocutor”, in some cases “gaze:other” with the comment “away from the interlocutor”. Thus, the interaction of head movement and gaze is an issue that the manual does not seem to treat satisfactorily.

The reason why the encoding of sequencing was problematic, thus resulting in a relatively low kappa-score (0.59), needs further analysis. The disagreement between the coders concerns especially the feature “sequencing:S-continue”, which they have chosen to use in different cases. To understand the problem, however, we need to conduct additional experiments.

The kappa-scores obtained on the annotation of the various features give us indications of a good reliability for most of the categories used. However, it does not tell us whether the coding scheme has the appropriate coverage. The material used in the Danish case study is of course very limited, so it is not a surprise that many of the available categories were not used (for instance, a very narrow range of expressions are relevant). However, it is worth noting that one of the basic feedback features, *F-elicited-acceptance*, was never used (thus the kappa-score concerns the default value “none”). To see whether this is an idiosyncratic fact of this particular dialogue or rather evidence of the fact that the feature is empirically inadequate, we need of course to look at more conversations. Concerning lack of necessary categories, on the other hand, it is obvious already from this limited study that body posture, which is not included in the scheme, is important for feedback: both coders have noted in their comments that a relevant movement of the torso should have been annotated. Therefore, body posture categories should be added to the scheme.

## **5. Second case study: the Swedish annotation**

The Swedish video clip consists of a one-minute dialogue excerpted from the Swedish film “Show me love”. The scene is a quite emotional conversation between two actors who interpret father and daughter. The actors are mostly taken in close ups of their faces. The actor who speaks is not always in focus, so in a couple of cases it has not been possible to see which facial display the actor was showing while uttering a feedback expression. Since the focus is on the actors’ faces, the hand movements

were rarely in the picture, which made it impossible to observe the possible hand gestures related to feedback, turn management and sequencing.

Only one expert annotator annotated the film scene, so it was not possible to carry out a formal evaluation of the reliability of the coding scheme.

A total of 12 facial displays related to feedback and 12 facial displays related to turn assignment were labeled. No sequencing facial displays were identified in this clip. Table 6 shows the number of annotated facial displays related to feedback giving and eliciting as well as turn management. Facial displays consisted of eye brow raises, smiles, gaze directions and head movements such as nods, shakes and tilts.

Table 6. Number of annotated feedback giving and eliciting turn management tokens

Turn-end	10
F-Give-emotion/attitude	7
F-Elicit-acceptance	2
F-Give-acceptance	1
F-Elicit-basic	1
F-Elicit-emotion/attitude	1
Turn-gain	1
Turn-hold	1
F-Give-basic	0

Since the video-clip is extracted from a film, all the conversational moves are pre-defined and for this reason only few turn-gain and turn-hold facial displays seem to occur. Given the emotional scene, it is not surprising that most of the feedback phenomena annotated have been labelled as F-Give-emotion/attitude.

In this clip there are two examples of the category F-Elicit-acceptance, which does not occur at all in the Danish material. One example is when the father, who has given his daughter a music CD as a birthday present, asks her if it was the correct one (i.e. the one she had desired). While asking this the father looks at his daughter and raises his eyebrows so as to request a positive acceptance feedback, which in fact comes in the form of a smile and a *yes thank you* from the daughter's side. This points to the fact

that the category is useful, and that its absence from the Danish data is due to the different communicative situation.

## 6. Conclusion

The MUMIN annotation scheme constitutes our first attempt at defining a scheme for the annotation of feedback, turn management and sequencing multimodal behaviour in human communication. From the results obtained on a few practical annotation cases, the categories defined in the scheme seem reliable although there was some insecurity about the encoding of some of the features, such as sequencing. Some of the attributes were never used in the present experiment, but we have too few annotations to conclude whether any of them are unnecessary. Other categories, on the other hand, should be added, particularly for the annotation of body posture, which is not part of this version of the coding scheme.

In general, we believe the availability of such a scheme is an important step towards creating annotated multimodal resources for the study of these phenomena in real face-to-face interaction, and for investigating many different aspects of human communication of interest not only to linguists and cognitive scientists but also to the human-machine interaction community. Examples of issues that can be investigated empirically by looking at annotated data are the extent to which gestural feedback co-occurs with verbal expressions; in what way different non-vocal feedback gestures combine; whether specific gestures are typically associated with a specific function; how multimodal feedback, turn management and sequencing strategies differ in different situations and cultural settings.

## References

- Allwood, J., Nivre, J., & Ahlsén, E. (1992). On the Semantics and Pragmatics of Linguistic Feedback. *Journal of Semantics* 9, 1–26.
- Allwood, J. (2001). *Dialog Coding – Function and Grammar*. Gothenburg Papers in Theoretical Linguistics, 85. Department of Linguistics, Gothenburg University.
- Allwood J., & Cerrato L (2003). A study of gestural feedback expressions. In Paggio *et al* (Eds) *Proceedings of the First Nordic Symposium on Multimodal Communication*, Copenhagen.

- Allwood, J., Cerrato, L., Dybkær, L., Jokinen, K., Navarretta, C., & Paggio, P. (2004). *The MUMIN multimodal coding scheme*. Technical report available at [www.cst.dk/mumin/stockholmws.html](http://www.cst.dk/mumin/stockholmws.html).
- Bernsen, N. O., Dybkær, L., & Kolodnytsky, M. (2002). *THE NITE WORKBENCH - A Tool for Annotation of Natural Interactivity and Multimodal Data*. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'2002)*, Las Palmas, May 2002.
- Beskow J., Cerrato L., Granström B., House D., Nordstrand M., & Svanfeldt G. (2004). The Swedish PF-Star Multimodal Corpora. *LREC Workshop on Multimodal Corpora: Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces*, Lisboa 25 May 2004.
- Carletta J., Isard A., Isard S., Kowto J.C., Doherty-Sneddon G., & Anderson A. H (1997). The Reliability of a Dialogue Structure Coding Scheme. In *Computational Linguistics* 23(1), 13–31.
- Cassell, J. (2000). Nudge Nudge Wink Wink: Elements of Face-to-Face Conversation for Embodied Conversational Agents. In Cassell, J. et al. (Eds.), *Embodied Conversational Agents*, 1–27. Cambridge, MA: MIT Press.
- Cerrato, L. (2004). A coding scheme for the annotation of feedback phenomena in conversational speech. In *Proceedings of the LREC Workshop on Multimodal Corpora: Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces*, Lisboa, 25 May 2004.
- Clark H., & Schaefer E. (1989). Contributing to Discourse. In *Cognitive Science* 13, 259–94.
- Core, M., & J. Allen (1997). Coding Dialogs with the DAMSL Annotation Scheme. Presented at *AAAI Fall Symposium on Communicative Action in Humans and Machines*, Boston, MA, November 1997.  
<ftp://ftp.cs.rochester.edu/pub/papers/ai/97.Core-Allen.AAAI2.ps.gz>
- Cowie R. (2000). Describing the emotional states expressed in speech, in *Proc. of ISCA Workshop on Speech and Emotion*, Belfast 2000, pp. 11–19.
- Duncan, Susan (2004). *McNeill Lab Coding Methods*. Available from <http://mcneilllab.uchicago.edu/topics/proc.html> (last accessed 26/4/2004).
- Ekman P. (1999) Basic emotions. In T. Dagleish and T. Power (Eds) *The Handbook of Cognition and Emotion* NY: J. Wiley, pp.45–60.
- Gunnarsson, M. (2002). *User Manual for MultiTool*. Available from [/www.ling.gu.se/~mgunnar/multitool/MT-manual.pdf](http://www.ling.gu.se/~mgunnar/multitool/MT-manual.pdf).

- Kipp, M. (2001). Anvil – A Generic Annotation Tool for Multimodal Dialogue. In *Proceedings of the 7<sup>th</sup> European Conference on Speech Communication and Technology (Eurospeech)*, pp. 1367–1370. Aalborg.
- Martin, J. C., Den Os, E., Kühnlein, P., Boves, L., Paggio, P., & Catizone, R. (2004). *Proceedings of the LREC workshop on Multimodal Corpora: Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces*, held in conjunction with LREC04, Lisbon, May 2004.
- Poggi, I., & Magno Caldognetto, E. (1996). A score for the analysis of gestures in multimodal communication. In: *Proceedings of the Workshop on the Integration of Gesture and Language in Speech*. Applied Science and Engineering Laboratories. L. Messing, Newark and Wilmington, Del, 235–244.
- Siegel S., & Castellan N.J.jr (1988). *Nonparametric Statistics for the Behavioral Sciences*, second edition. McGraw-Hill.

## Biographies

**Jens Allwood** has been since 1986 professor of Linguistics at the department of Linguistics at Göteborg University. He is also director of the inter-disciplinary cognitive science center SSKKII at the same university. His research primarily includes work in semantics and pragmatics. He is investigating spoken language interaction from several perspectives, e.g. corpus linguistics, computer modelling of dialog, sociolinguistics and psycholinguistics as well as intercultural communication. Presently he is heading projects concerned with the semantics of spoken language phenomena, multimodal communication, cultural variation in communication and the influence of social activity on spoken language

**Loredana Cerrato** is a PhD student affiliated at the Department of Speech Music and Hearing and the Centre for Speech Technology of the Royal Institute of Technology (KTH) in Stockholm. Her research focuses on the analysis of non-verbal behaviour in unimodal and multimodal materials from human as well as human-machine interaction. Before starting her doctorate at KTH she worked at Telia Promotor Infovox in Stockholm (now Acapela group), developing speech synthesis systems.

**Kristiina Jokinen** is Docent of Language Technology at the University of Helsinki and currently a Visiting Fellow of Clare Hall at the University of Cambridge. Her research concerns AI-based spoken dialogue management, adaptive interactive systems, multimodal interfaces, rational agents, and natural cooperative communication. She developed the Constructive Dialogue Model approach for interaction management in dialogue systems and has directed several national and international research projects on spoken dialogue systems and adaptive user modelling.

**Costanza Navarretta** is senior researcher at the Centre of Language Technology of the University of Copenhagen. She has worked in several areas of computational linguistics: reference resolution, discourse and dialogue structure, grammar formalisation and implementation, methodologies for building ontologies from texts, multimodal communication, computational lexicography, evaluation of NLP resources. She has participated in numerous Danish and European NLP-related projects and has been teaching at the IT University in Copenhagen since 2002.

**Patrizia Paggio** is senior researcher at the Centre for Language Technology of the University of Copenhagen. Her research has dealt with issues in a number of different areas of natural language processing including machine translation; evaluation of language technology tools; constraint-based approaches to syntax, semantics and pragmatics; ontology-based querying and multimodal interaction. Currently, her work focuses on the treatment of information structure in Danish. In the last three years she has coordinated the Nordic network for multimodal interfaces MUMIN.

**Authors' addresses:**

*Jens Allwood*  
*Dept of Linguistics*  
*Göteborg University*  
*Box 200*  
*405 30 Göteborg*  
*phone: +46 31 773 1876*  
*e-mail: jens@ling.gu.se*

*Loredana Cerrato*  
*KTH*  
*TMH/KTH*  
*Lindstedsvägen 24*  
*10044 Stockholm*  
*Sweden*  
*phone: +46(8)790 7856*  
*e-mail: loce@speech.kth.se*

*Kriistina Jokinen*  
*University of Helsinki*  
*Pl 94*  
*45100 Kouvola*  
*Finland*  
*phone: +358 5 8252 234*  
*e-mail: kristiina.jokinen@helsinki.fi*

*Patrizia Paggio*  
*Center for Sprogteknologi*  
*Njalsgade 80*  
*2300-DK Copenhagen S*  
*Denmark*  
*phone: +45 3532 9072*  
*e-mail patrizia@cst.dk*

*Constanza Navarretta*  
*Center for Sprogteknologi*  
*Njalsgade 80*  
*2300-DK Copenhagen S*  
*Denmark*  
*phone: +45 3532 9065*  
*e-mail: costanza@cst.dk*



# MULTI MODAL INTERACTION IN AN AUTOMATIC POOL TRAINER

*R. Atladottir, J. Gay, K.L. Jensen, R.B. Jensen, I. Lontis,  
L.B. Larsen, S. Larsen*  
Aalborg University, Denmark

## **Abstract**

*For testing user experience and usability issues an automatic pool trainer has been developed which utilizes multi modal user interaction through computer vision, speech recognition, and agent technologies. This paper presents a non-intrusive system, developed with user tasks in focus, for a standard non-modified pool table, relying on gesture and speech recognition supplementing each other in navigating the intuitive game menu. A camera is used for recognizing the gestures and feedback is given through sounds, speech synthesis and animations projected onto the table. The ease of use of the system has been proven through a number of user tests involving a total of more than 200 users. Further development of the system is discussed with focus on user feedback and exploration of other game schemes.*

## **1. Introduction**

This paper presents the development of IntelliPool, an automatic system for training the game of pool. The IntelliPool system has a twofold purpose. Firstly, it serves as a research platform for multi modal user interaction using computer vision and speech recognition, agent technologies, etc. Secondly, a computer aided learning system for the game of pool is developed as a specific application and represents the focus of this paper.

The Chameleon project carried out at Aalborg University in 1997-1999 defined the open architecture of 'IntelliMedia WorkBench', an integrated system of hardware and software modules 0. The workbench served as the original platform for the pool training system, which has been continuously

developed through a number of different versions [1], [3], [4], [5], [7], [8], [9]

The system evolution consists of improvement in the technique of the multi modal user interaction (i.e. the display, voice and movement interaction and the agent substituting the trainer), as well as improvements in the individual processing modules. More information about the system is available at the IntelliPool homepage (<http://cpk.auc.dk/SMC/pooltrainer>).

### ***1.1 Target Pool***

In 1993, the professional pool player Kim Davenport developed Target Pool 0, a widely used scheme for pool training. IntelliPool uses this system, offering the user an enhanced interaction environment. Target Pool consists of a collection of self-study courses that include exercises with different levels of difficulties and themes.

An exercise implements the practice of a technique required by a particular shot. Specific instructions are given for the initial ball placement, the trajectories and the “Target”, i.e. the desired location of the cue ball after the shot. This is presented graphically. A short instruction text accompanies each exercise to describe how to perform the shot.

The user has at hand a booklet describing the exercises, a score-board, and a thin cloth with a printed target to be placed on the pool table. Most of the exercises use only the cue ball and one other ball (the object ball). The concept behind an exercise is to sink the object ball into a designated pocket while the cue ball must stop as close to the center of the Target as possible.

### ***1.2 IntelliPool***

The intention of the IntelliPool system is to replace many of the tasks mentioned above by an interactive multimedia system. The main objectives are to relieve the user of the manual tasks of setting up and evaluating the exercise, as well as the book keeping tasks, thus enabling him to concentrate fully on the game.

The main body of this paper is outlined as follows: First the user interaction paradigm is presented, followed by a description of the system design. Then system performance and general user experience is discussed and finally some conclusive remarks about the study will be made.

## **2. User Interaction and Task**

This section describes the design of the user interaction.

### ***2.1 Constraints and Choices of Modalities***

Throughout the design, an overall goal has been to establish the interaction according to the user's needs and requirements. In particular, we wanted to avoid the addition of new devices that the user must wear or use, such as touch screens, keypads or buttons embedded in the pool table, PDAs, shutter glasses, etc. Instead we have opted for a configuration, where the control computer and interface devices are completely hidden to the user.

The two key tasks of IntelliPool are to instruct the user where to place the balls and evaluate the shot after the balls have stopped moving. This requires the ability to detect the position of the balls, and determine whether they are moving. This task can be done in several ways, but due to the non-intrusiveness (it must be possible to use any standard pool table and equipment) of the approach, the choice of computer vision is an obvious candidate.

Cameras are ubiquitous and can be placed well above the table. Likewise, speech recognition does not require the user to be at a specific location and hence can be used as input modality e.g. for choosing exercises.

### ***2.2 Multimodal UI Design***

Information must be presented to the user following the requirements set out above. Obvious candidates are the auditory and visual modalities. Sound is ubiquitous and can easily be conveyed to the user from loudspeakers placed in the surroundings.

Two types of auditory output are used, speech and non-speech. While speech is "rich" in term of the information it can convey, jingles and earcons can quickly inform the user of events and capture his attention.

Several options exist for visual output, such as using a monitor, a projector screen, touch screen or a wearable display. The latter has been adopted in an augmented reality system for support of pool players, namely the Stochasticks Project by Tony Jebara at Columbia University 0. We chose a different approach, which we believe allows the user a greater degree of

freedom: To project a visual display directly on the surface of the pool table using a powerful projector mounted above the pool table.

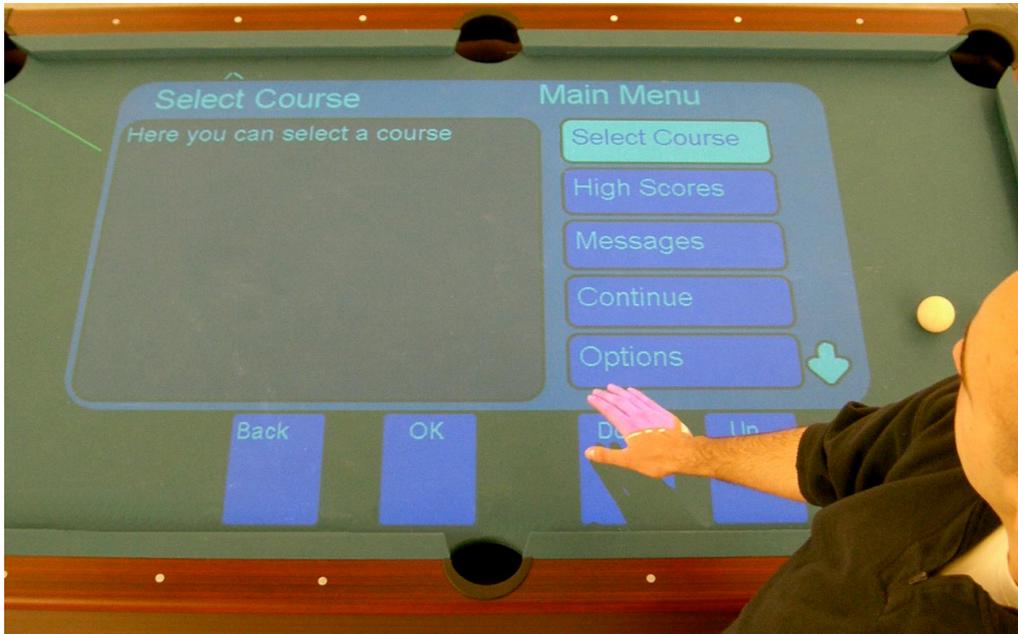


Figure 1. The menu displayed directly on the surface of the pool table.

This solved the very important problem of instructing the user in a natural and efficient manner how to set up the balls for a shot and where to aim. Any other solution would require the user to estimate the correct position from a set of instructions, either given orally or presented graphically in a display. In either case, the user will have no direct indication whether he has placed the balls correctly.

Generally there are two ways to interact with the current version of the system. One is via a virtual menu and the other is via speech and an interface agent.

### 2.3 Virtual Menu

The menu interaction is based on a virtual menu projected on the pool table (figure 1). This requires the user to stand on the side of the table while interacting. The users activate the menu to start the system and choose the appropriate actions. The menu is activated by covering two corner diamonds on the pool table. In the bottom of the menu there are four virtual buttons that are used to navigate and select menu items (figure 1). The virtual buttons are activated when the user places his hand over them for a certain amount of time.

The menu has a classical hierarchical tree structure with no shortcuts. One of the advantages of using a menu is that most users are familiar with interacting with menus from the use of computers. Also the menu shows all the available actions and the structure guides the user in his choices.

## ***2.4 Speech and Agent***

The user can also operate the system through spoken commands and receive feedback and guidance from the interface agent.

The interaction is based on a dialogue between the user and the agent. The user gives spoken commands which the system recognizes and reacts upon those commands via feedback from the agent and relevant actions in the system. The agent character is displayed on the pool table so the user's focus can be on the pool table the whole time. The agent's basic actions are to speak to the user and use gestures and animations to guide him. When the agent speaks the text is also shown in speech bubbles so the user can read the text if he prefers (*figure 2*) .

One of the advantages with speech interfaces is that the user is free to move around as he pleases during his interaction. Also the user's hands are free and he can hold the cue at all times. The system is in nature a tutoring system and people are used to having a human instructor to teach them. The personification of the user interface with the agent is thus thought to make the interaction more natural for the user.

## **3. Design & Implementation of Modules**

This section will describe the internal structure of the system by accounting for the design and implementation of the different modules of the system. The overall structure is depicted in *figure 3*.

### ***3.1 Vision Module***

The vision module, called Image Processor, is responsible for analyzing the images captured by the camera and notifying the rest of the system when events occur during game play. These events are tightly coupled to the flow of the exercise driven pool game. Typical events are: Detecting when the balls have been placed correctly before a shot, detecting when a

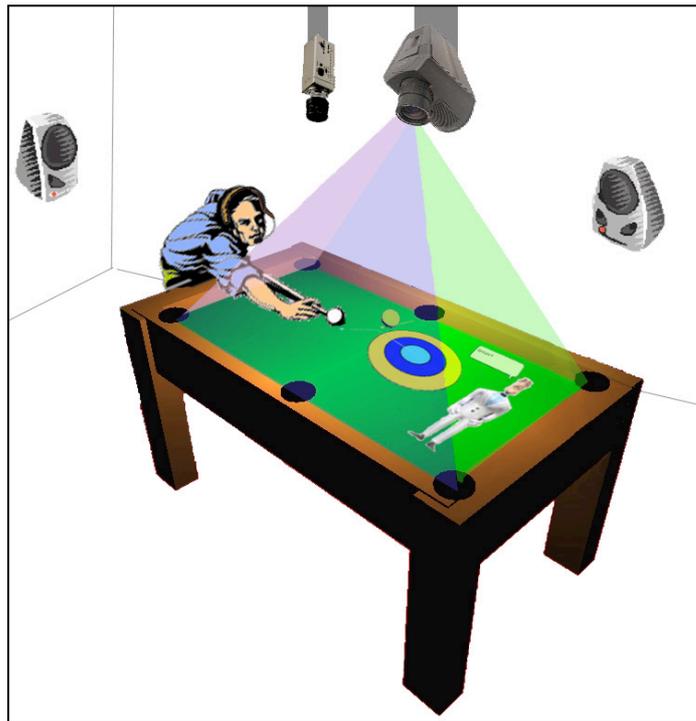
shot starts and ends, as well as detection of the balls' final position in order to assess the quality of the shot and award points to the user as feedback.

In addition to analyzing the images during game play the Image Processor module also serves an important role when both activating and navigating the menu (*figure 1*). Whenever the user holds a hand over one of the virtual buttons the Image Processor will notify the Kernel of the event that a button has been pressed.

All of the image processing is handled in real time ensuring the user a solid experience where no unnecessary waiting is introduced due to buffered images awaiting processing.

### 3.2 *Helios Module*

Helios is the image output server of the system. Its main task is to dynamically draw up the exercises and the system menus as dictated by the state of the game. The exercises are drawn as colored circles and lines and a bull's eye target. The circles indicate where the user must place the balls and the lines show the expected paths of the balls during the shot and thus marks which pockets should be hit. The target area shows where the cue ball should end up after the shot. See figure 2 for an example of this.



*Figure 2. Conceptual picture of the pool table.*

The menu is a conventional hierarchical menu allowing the user to choose courses and exercises and toggle various game options. An example of the menu is shown in *figure 1*

The Helios server uses a hierarchy of scalable graphical objects from which complex compound objects such as complete menus 0. Helios also supports

a number of advanced graphical display features, such as timed animations for displaying the scores.

### ***3.3 Speech Module***

To be able to recognize and react on the user's commands a speech engine is used (Microsoft English Speech Recognition Engine). A grammar is defined for the recognizer and includes the available commands in the form of rules for words and phrases to be recognized. The recognizer then only listens for these commands. The grammar vocabulary defined for IntelliPool is based on words and phrases applied by test users. It was apparent that users tended to use short command-like phrases rather than natural speech.

Finally the recognized commands should trigger appropriate actions in the system. The actions can for example include starting an exercise or some feedback from the interface agent. These actions are implemented in other modules of the system and to trigger them the Kernel has to send these modules messages. Therefore the speech module notifies the kernel which decides the appropriate response which usually involves sending a message to relevant output module(s).

### ***3.4 Agent Module***

The behavior of James the interface agent is controlled by the agent server. The technology used is Microsoft agent; a set of software services that supports the presentation of interactive animated characters: gesture, motion, text display (using cartoon speech balloon), text to speech synthesis and prerecorded audio files spoken by the agent. The settings of the agent are controlled in the agent server: choice of the character (James), type of voice, speed of elocution, character size and display of the speech balloon. All the possible sentences (text prompts) that the agent can say are stored in a database. The Kernel module sends to the agent server an input specifying the prompt to use, the appropriate gesture and the position of the agent on the pool table. The agent module is context sensitive, for example more information is included when a certain function is used for the first time. It is also possible to turn of the sound, hide or disable the agent.

### 3.5 Kernel Module and Communication

The Kernel is the central module of the system keeping track of the overall state and making sure this is also reflected in the subsystems. This is carried out by sending messages from the Kernel to the subsystems to setup and update each subsystem with data when it is relevant. To determine whether a certain state change is relevant to the different subsystems, each subsystem subscribe for all the system events that it wants to be notified of. This approach makes it straight forward to add new modalities to the system, because they can subscribe to any relevant events. The Kernel also receives notifications from the subsystems when the status is updated in a subsystem, to determine if this should cause the overall system to go to a new state.

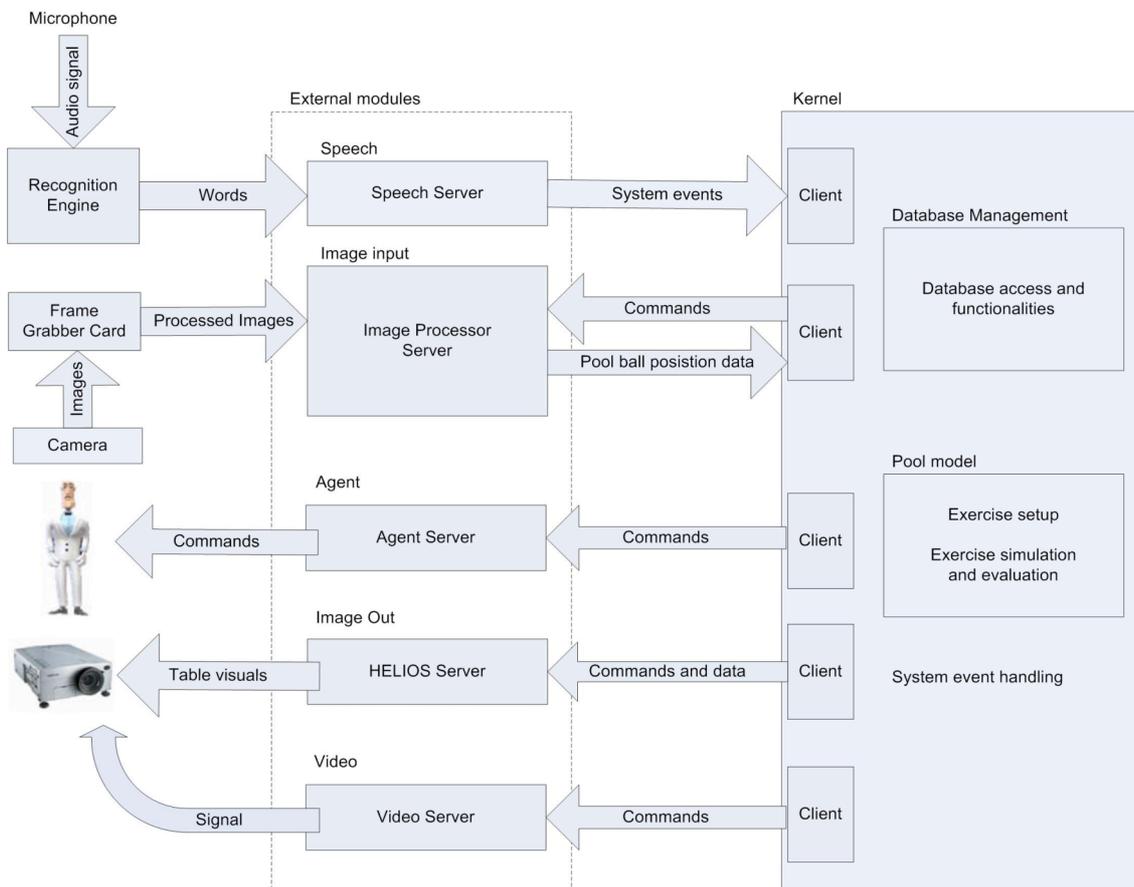


Figure 3. Communication between the kernel and the system modules.

Internally the Kernel has a state machine module which as the name implies is responsible for keeping track of the system state, and for firing the right events. It also takes care of error handling in case an error occurs, e.g. an illegal notification is received from a subsystem or no subsystem

replies to an event. Game specific communication is handled by specific game handler modules, which makes it easy to add more games than the Target Pool currently supported. Finally the Kernel has models of the menus and the Target Pool game, and thus any changes are handled here. Data regarding the menu setup, Target Pool exercises, scores and user preferences, etc, are acquired and stored in a database, making it easy to e.g. change the menu or add new exercises.

#### **4. Performance and User Experience**

During the development of IntelliPool numerous formal and informal user tests have been performed, and the current design is a reflection of the experiences collected through these. Most notably, an extensive user test was carried out during the DNF (Danish Natural sciences Festival) in September 2004, where more than 200 users tested and trained their pool skills with the IntelliPool over a 4 day period (Speech and Agent interaction were not used in this test).

The main accomplishment of the IntelliPool system is the seamlessness and non-intrusive way of which the users interact with it. Most users never realized or quickly forgot that they were in fact using a quite complex piece of engineering. The projector, camera and loudspeakers are sufficiently out of the way for the users not to recognize them at first. In fact many testers actually searched the table for real buttons or touch sensors.

##### ***4.1 Experiences from User Tests***

The notion of using the surface of the pool table as a display has proven highly successful. Users have found it very intuitive and easy to understand the purpose of the markings. In particular, this allowed the users to keep the attention focused directly on the table during the exercises. All users and especially children easily and quickly learned to master both setting up the exercises and navigating the menu system with their hands.

Some usability tests were performed on the system. Although they only included a limited amount of users several usability issues were identified. Some positive comments were also made and people seemed to be generally interested in the system and enjoyed using it.

## **4.2 *Virtual Menu***

Some users had some initial problem understanding the virtual menu concept. Their first instinct was to operate the menu by pushing with their hand on the appropriate action in the menu list rather than using the virtual buttons to navigate and select. This is understandable since conventional menus function in that way. When pressing the virtual buttons the users also need to place the hand long enough for the system to recognize it; but still not for too long since this would trigger it a second time. This was a problem for some users and they sometimes ended up selecting something they did not intent to. There were also some problems due to the image recognition where the system failed to recognize when a button was pressed. But this is understandable since this system is still merely a prototype. All in all the users were however fast to learn how to use the system and could then easily find their way through the conventional menu hierarchy.

## **4.3 *Speech and Agent***

Speaking to the system seamed awkward to some users since they were not used to speech interfaces. The users were often unsure of what to say to the system and they tended to use shorter keywords rather than natural language. Some users also had the tendency to think in terms of menu navigation when using the system, wanting to go back to some original state to perform a new action.

The speech interface also puts more constraints on the users' memory and some users did not know what their options were in the system. To compensate for this the users could ask the system about their possible options and the agent would also make some suggestions. Sometimes the users' commands were not recognized either due to missing keywords in the grammar or to errors in the speech recognition itself. The speech recognition problems can somewhat be explained by that the developers and the test users were not native English speakers. Some users noted that the agent voice was not natural or human. This could be improved with a prerecorded human voice.

It probably takes a longer time for the user to learn to use the speech interface but once he becomes experienced the speech interface can be faster to use since the user is not restricted to the menu hierarchy.

Also the users noted that they found the system fun to use and they seemed to like the agent.

#### ***4.4 Identified Problems***

From the technical viewpoint the system inherently suffers from several problems with regard to the stability of the physical environment in which it is set to run. Particularly the vision and speech modalities are sensitive to highly dynamic room lighting and high auditory noise levels respectively. Even though sophisticated computer vision techniques and a commercial speech recognition engine is used, some realistic constraints must thus be met by the environment for a smooth user experience. The robustness of the system was proved at the DNF where at times up to 30 children crowded the pool table simultaneous without provoking any errors in the execution. This ensures a genuine non-intrusive user experience.

### **5. Discussion & Conclusions**

As described the IntelliPool system employs a number of advanced user interface techniques to achieve an intuitive and efficient interaction style. Besides more formal usability test the system has been tried by more than 200 users over a four day period at the DNF, and it is our impression that users very quickly grasp the underlying concepts and immediately start to interact with the system without the need for assistance.

While being robust in a controlled environment the stability of the vision and speech modalities are sensitive to changes in room lightning and auditory noise levels. Thus currently the user is required to wear a noise reduction microphone and use a push-to-talk button.

### **6. Further Work**

In a previous version a module giving feedback about the errors the user might have made during a shot has been implemented. This module should be integrated into the new architecture. Also, the system allows for the addition of other game types than the Target Pool, and it would be interesting to examine the possibility of other kind of entertaining games. In time, the system has the potential to be developed into a commercial system, requiring robustness and general maturing.

## Acknowledgments

The authors wish to thank all those who have enthusiastically contributed to the IntelliPool. A full list of contributors can furthermore be viewed at the IntelliPool homepage: <http://cpk.auc.dk/SMC/pooltrainer/>.

## References

- Bondensen, P., Poulsen, P., & Lykkegaard, M. "Smart-Pool – A multi modal pool training system", Aalborg University, June 1999
- Brøndsted, T., Dalsgaard, P., Larsen, L. B., Manthey, M., McKevitt, P., Moeslund, T., Olesen, K. "A platform for developing Intelligent Multi Media Applications", *Technical Report R-98 1004*, May 1998, CPK, Aalborg University.
- Buch, J. et al.: "Intelligent Multimedia Based Pool Trainer". *Report*, IMM June 1998
- Chaib, D., Pannet, Y., Tseveen-Ochir, A.: "User error diagnosis in the Automatic Pool-trainer system", *Report*, IMM January 2004, Aalborg University
- Duval, B., Gay, J., & Atladottir, R. "Multimodal Interaction in the Intelligent Pool Trainer", *Report*, IMM January 2005, Aalborg University
- Davenport, K. "Target Pool", Target Pool Productions, P.O Box 219, Marysville, Michigan, 48040, 1992.
- Jensen, K. L., Jensen, R. B., Kun, I., & Larsen, S. "IntelliPool<sub>3</sub>", *Report*, IMM June 2004, Aalborg University
- Jensen, M. & Vodzi, W. K.. "Adding speech and a software agent to the Intelligent Pool Trainer", *Thesis Report*, Aalborg University, June 2002.
- Larsen, L. B., Jensen, P. M., Kammersgaard, K., & Kromann L. "The Automated Pool Trainer - A multi modal system for learning the game of Pool", *Intelligent Multi Media, Computing and Communications: Technologies and Applications of the Future*. John Wiley and Sons ISBN 0-471-20435-8, June, 2001, pp.90-96
- The Stochastic Project:  
<http://www1.cs.columbia.edu/~jebara/stochastic.html>

## Biography

Research: HCI and Usability issues within Spoken and multimodal Dialogue Systems. Teaching: Coordinator of the Intelligent MultiMedia Masters programme. Courses in Spoken Dialogue systems and usability engineering. Expertise: Communication and informatics, user interaction, design of HCI, speech technology, usability, digital processing of speech and audio signals, multi media technology.

## Authors' addresses

*Lars Bo Larsen*  
*Associate Professor, PhD.*  
*Speech and Multimedia Communications*  
*Dept. of Communication Technology*  
*Aalborg University*  
*Niels Jernes Vej 12, DK-9220*  
*Aalborg*  
*Denmark*  
*phone: +45 9635 8635*  
*e-mail: bl@kom.aau.dk*  
*fax: +45 9815 1583*

*Rosa Atladottir, Jeremie Gay, Kasper Løvborg Jensen, Rene Balle Jensen,*  
*Søren Larsen, Ildiko Lontis,*  
*Speech and Multimedia Communications*  
*Dept. of Communication Technology*  
*Aalborg University*  
*Niels Jernes Vej 12, DK-9220*  
*Aalborg*  
*Denmark*  
*phone: +45 9635 8650*



# THE PHILOSOPHY BEHIND A (DANISH) VOICE-CONTROLLED INTERFACE TO INTERNET BROWSING FOR MOTOR-HANDICAPPED

*Tom Brøndsted*

Dept of Communication Technology  
Aalborg University, Denmark

## **Abstract**

*The public-funded project "Indtal" ("Speak-it") has succeeded in developing a Danish voice-controlled utility for internet browsing targeting motor-handicapped users having difficulties using a standard keyboard and/or a standard mouse. The system underlies a number of a priori defined design criteria: learnability and memorability rather than naturalness, minimal need for maintenance after release, support for "all" web standards (not just HTML conforming to certain "recommendations"), independency of the language on the websites being browsed, allowance for multimodal control along with the unimodal oral mode, etc. These criteria have led to a primarily message-driven system interacting with an existing browser on the end users' systems.*

**Keywords:** Alternative web-browsing, voice-controlled applications, e-inclusion, ubiquitous speech processing, accessibility for disabled persons.

## **1. The Project Indtal**

The project Indtal ("Speak-it") has aimed at developing a Danish voice-controlled tool for internet browsing targeting motor-handicapped users having difficulties using a standard keyboard and mouse. The project was funded by the *Danish National IT and Telecom Agency* under the *Ministry of Science* and ran from primo January 2004 to ultimo February 2005. The

project partners were the *Department of Communication Technology* at Aalborg University being in charge of the actual speech recognition technology and the software company *Efaktum* in Hjørring implementing the back-end communication between the recogniser and the browser engine of the system. Further the project involved two non-technical partners, "*Specialskolen for Voksne, Vendsyssel*" and "*Teknologicentret for Handicappede, Nordjyllands Amt*", institutions located in North Jutland and specialized in compensating courses and consultancy for adults with special needs, including adults with physical disabilities. These non-technical partners have been in charge of project management and contact with an advisory-board of potential end-users, and the maintenance of a website [www.indtal.dk](http://www.indtal.dk) where disabled users can download the browser for free.

This paper consists of two parts: 1) The first part outlines five central design criteria characterizing the Indtal-browser as opposed to other alternative browsers addressing disabled users. 2) The second part describes the recognition front-end developed for the system. The first part is to a large extent a summary of a paper submitted to Interspeech 2005.

## **2. Design & Implementation Criteria**

The Indtal-browser differs from other “alternative web browsers” by a number of design and implementation criteria. We define an alternative web browser as a browser offering an alternative to either standard visual output rendering or to standard keyboard and mouse input control – or both. A number of such alternative browsers are listed at W3C’s website hosting the Web Accessibility Initiative (WAI) (W3C 2005) and similar websites devoted to disabled users. Roughly, we distinguish *two major groups* both of which can deploy speech recognition and/or speech synthesis (Table 1):

Table 1: Major groups of “alternative browsers”

	<b>GROUP 1</b>	<b>GROUP 2</b>
<b>TARGET GROUP</b>	Visually impaired users	Any user preferring hands-free (+ eyes-free) browsing
<b>INPUT</b>	A set of function keys (+ an equivalent set of spoken commands), <i>no</i> control of the mouse cursor, mouse clicks etc.	A set of spoken commands + dynamically generated commands for activating links on the current page
<b>OUTPUT</b>	Structured output sent to a Braille display or a speech synthesizer	Unaltered visual rendering enriched with so-called Saycons (+ speech synthesis)
<b>PARSING OF WEB-CONTENT</b>	“Deep” parsing of the web pages being browsed	No “deep” parsing of the web pages being browsed

Examples of group 1 are Braillesurf (Hadjadj et al 1999, Schwarz et al. 2005), BrookesTalk (Zajicek et al. 2000), Emacspeak, and Homer (Mihelic et al 2002). Examples of group 2 are Conversa Voice Surfer (formally Conversa Web) (Robin et al. 1998), HFB (HandsFree Browser by EduMedia) and add-ons shipped with certain versions of IBM’s ViaVoice and Dragon Natural-Speaking.

The Indtal-browser belongs to the latter group though it has an *explicit focus on end-users with mobile disabilities*. Generally web-content is highly visually oriented and it makes in our view no sense to attempt to support eyes-free browsing unless the needs of visually impaired users are addressed explicitly. *Combining* hands-free and eyes-free facilities for web browsing hardly makes sense at all. Users with both mobile and visual handicaps are extremely few and their most severe everyday problems do not encompass access to the web.

Apart from the stricter focus on the end-users, the Indtal-browser has been implemented to meet the following five criteria:

*Criterion 1: Minimizing Future Maintenance Requirements*

Many alternative browsers developed during the last decade have been quietly withdrawn leaving no trace except for the broken links on the referring sites like the WAI site hosted by W3C. One possible explanation for the apparent short lifetime of such systems is that they are developed in the framework of research projects or they are (like Indtal) the result of a – once-and-for-all funding leaving no resources for subsequent maintenance.

To minimize the requirements for future maintenance, Indtal has chosen a mainly message-driven approach. The system runs on Win32 systems, is dependent only on the Microsoft C-runtime library, and uses Microsoft Internet Explorer as its browser. The alternative possibility of building the system as a modification of open-source browsers like the GTK Web browser Dillo ([www.dillo.org](http://www.dillo.org)) or Mozilla ([mozilla.org](http://mozilla.org)) was ruled out.

### *Criterion 2: Allowance for other Input Devices*

The message-driven approach described above further has the advantage that speech control can coexist with other “third party devices” generating keyboard and mouse messages to the operative system and the browser engine.

The contact with potential end-users during the design and implementation phase has shown that many of them to some extent are capable of operating a standard PC with some additional equipment, typically short hand-mounted “sticks” to use with standard keyboards, head trackers and eye trackers to generate mouse messages and operate on-screen keyboards, specialized “joysticks” tailored for the end-user who may be able to control the neck, a few fingers, etc. These devices can coexist with the oral control of Indtal.

### *Criterion 3: Support for “all” Web Standards*

The Indtal browser aims at a non-normative approach to the format and structure of the web pages being browsed. Many alternative web browsers only (or mainly) support HTML, typically with further restrictions regarding the fulfilment of certain “recommendations” (e.g. the WAI recommendations of W3C, W3C 2005).

For the alternative browsers of group 1 (explicitly addressing visually impaired users) this restriction is unavoidable since they have to employ a deeper “understanding” (than group 2) of the web pages being browsed. For instance, by allowing visually impaired users to skim the content by outputting only headlines or links, the browser relies on the web content being well-formed and in compliance with the WAI-recommendations or similar.

Alternative browsers of group 2 only encounter similar problems to the extent that they attempt to incorporate also some of the functionality specific to group 1. Otherwise the normative approach to the content being browsed must be considered an *unnecessary limitation*.

With one exception, the Indtal browser does not attempt to alter or “translate” the standard visual output rendering of web content. The exception is the visual enumeration of links (the HTML <a> elements) that at the users’ request are displayed in the browser window (indicating how to activate the links by spoken commands: e.g. “*go to link number twenty four*”). Hence the parsing of web-content in Indtal is minimal and extremely robust.

Further, to allow users also to access web-content implemented in *non-HTML* (e.g. activating mouse-over events like pull-down menus implemented in ECMA-scripts), Indtal also deploys a voice-controlled mouse. The system depicts the mouse cursor as the center of a compass with rulers in eight corners (north, north-east, east, etc.). Each ruler depicts a point and numbered value for every 100th pixel helping the user moving the cursor by commands like “*go north-east two hundred and ten*” etc. The cursor can be positioned anywhere on the screen with just two commands, though users (including trained ones) usually need a few more!

*Criterion 4: Independence of the language on the web pages being browsed*  
Danish constitutes a small language community, also on the web! We assume that Danes are more likely to view web pages in non-native languages than e.g. English users. Hence, it would be perceived as a severe limitation if the Indtal-browser could only access web pages composed in Danish.

The alternative browsers belonging to group 1 (cf. table 1) have to employ a deep “understanding” of the web-pages being browsed and often language-dependent parsing techniques are used. E.g. intelligent summarizing of (long) documents presuppose language-dependent techniques. Further, if the textual content of web-pages is sent to a speech synthesizer, the language dependency is increased.

Alternative browsers belonging to group 2 need not employ techniques dependent on the language of the web-page being browsed. Many of them are language-dependent either because they implement some functionality otherwise specific only to group 1 or because they support the so-called Saycons™ technology.

The Saycons™ technology implies that links can be activated by dynamically generated voice commands, e.g. that the standard sub sections

found in web versions of newspapers can be accessed by commands like *go to "Sports", "Domestic News", "International News", etc.* This increases the (apparent) naturalness of the application. However, the problems involved are: 1) links (text within the <a>-element and the alt-value of pictures within the <a>-element) must be unique, easy to pronounce, and acoustically discriminative. 2) The links must be composed in the supported language (otherwise the automatic transcription to phonemes will not work).

Due to these problems, the Indtal-browser does not support the Saycons™ technology. The numbering of links described above is the only functionality allowing the user to activate a link by a single command. As a result, the lexicalized vocabulary used in Indtal is *closed*. This allows for training of whole word acoustic models that are more robust than flexible (vocabulary-independent) models modeling e.g. generalized triphones.

#### *Criterion 5: Memorability and Learnability rather than Naturalness*

The motivation for using speech recognition and speech synthesis technology in human-computer interaction is often given in terms of “naturalness” and similar:

- *“Voice browsing is more natural and convenient than point-and-click browsing”* (Conversa Web)
- *“Voice is the most natural and effective way we communicate. In the years to come it will greatly facilitate how we interact with technology”* (Opera Software ASA)

Hugh Cameron (Cameron 2000) represents a much more sceptical view when it comes to the use of speech technology in HCI:

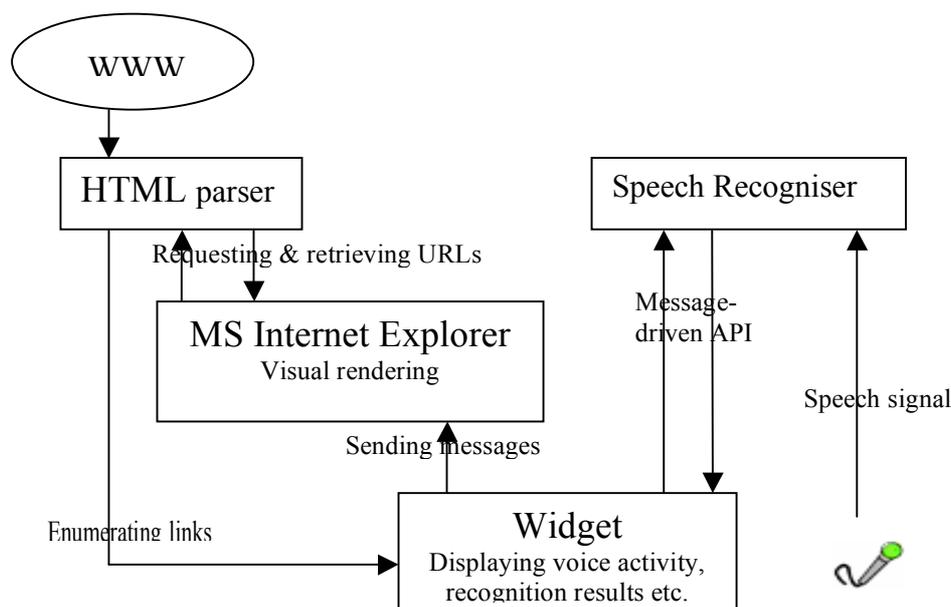
*“When will people use speech to communicate with machines?”*

- *When they are offered no choice.*
- *When it corresponds to the privacy of their surroundings and the task at hand.*
- *When it’s quicker than any alternative”*

The Indtal-browser explicitly addresses users with mobile disabilities. Hence, the use of speech recognition technology is justified even by Cameron’s far more critical criteria. However, the same criteria cannot justify the use of speech synthesis.

### 3. Front-end Speech Recogniser

The overall architecture of the Indtal-browser system is depicted in figure 1. The front-end consists of the speech recogniser communicating with the actual application (visual for the user only in the form of a bar-like widget) which sends the appropriate messages to the browser-engine of the system (MS Internet Explorer). The HTML-parser which is a part of the back-end application enumerates links and can on users' request force the browser to display the link numbers.



*Fig 1. Overall Architecture*

The speech recogniser of the system resides in a dynamic link library, is application-independent, and to some extent based on the SpeechDat(II) reference recogniser (Lindberg et al. 2000). It has been re-implemented to support real-audio input, hardware-detection and mixer-settings (recording level), endpoint-detection, language-selection based on the PC's local, pronunciation dictionaries in the SAMPA format, multiple grammars that can be activated and de-activated, and an application-independent API using the established call-back or "listener" paradigm.

To be used in the desktop-environment with acoustic models trained on data recorded over the fixed-line telephony-network (e.g. standard SpeechDat(II)-models), a down-sampling filter simulating telephone bandwidth with standard COMBO-characteristics can be applied.

The recogniser uses core modules from HVite of the HTK toolkit (Young et al. 1997): Front-end feature processing (Mel-scaled cepstral coefficients), internal representation of acoustic models (currently generalized triphones with state tying) and grammar lattices. Further, the actual Viterby decoding algorithm is based on HVite.

The API has lent its basic abstractions from MS Sapi 4.0 and JSAPI where the typical steps for an application to access a speech recogniser include:

- 1) creating the recogniser for a specific language and submitting a callback-handle (that retrieves information about speech activity etc.),
- 2) creating one or more “rule grammars” and adding a “listener” to each of them (the listener retrieves recognition results including the name of the grammar accepting input, an  $n$ -best list where each item consists of a sequence of “tokens” with time-information and score)
- 3) enabling and disabling grammars, committing changes, resuming recognition etc.

The recognition module has been implemented entirely in C and C++, however due to hardware detection and support for direct microphone input etc. it only compiles and runs on WIN32 systems. The API has been implemented as an ANSI C application interface allowing access also from other programming languages (e.g. the back-end application of Indtal is implemented in Delphi).

On top of the ANSI C API a further API based on JNI (Java’s Native Interface) has been implemented, and a number of interfaces specified for the standard extension package *javax.speech.recogniser* have been implemented in Java. Thus the recogniser (with the JAVA extension called “JHvite”!) to a large extent is JSAPI compliant.

The *application and language independence* of the recogniser has proven its worth in a couple of student projects. A further demonstration of the independence, a simple multilingual voice-controlled calculator using the dynamic link library with the recogniser shipped with Indtal, can be downloaded from *kom.aau.dk/~tb/indtal/*. The calculator is controlled by speaking natural numbers (e.g. *twenty five point five*) and arithmetical operations (*plus, minus, multiplied by, divided by*). To control the calculator using e.g. English or German commands, one simply has to add

a sub-directory with the corresponding local-name (*en* or *de*) and copy the standard SpeechDat(II) files to that location: 1) models (*tied\_32\_2.mmf*), 2) list file (*tied.lis*), 3) SAMPA pronunciation dictionary (*lexicon.tbl*), 4) mapping file if required (*phone.map*).

Due to ownership-problems, Indtal can only provide users with Danish SpeechDat(II) files (located in the *da*-subdirectory).

#### 4. Conclusion

The Indtal-browser has been evaluated by usability-experts at Aalborg University based on interviews with test persons belonging to the target-group (Jensen et al. 2005). The conclusion of this evaluation suggests various improvements to the system. It should come as no surprise that some users are not 100% happy about the accuracy of the speech recogniser. However, we know that mobile handicaps sometimes influence users' ability to articulate normally. A speech database recorded in the desktop-environment to be used for more robust whole-word models have been established during the project, but the released version uses the standard SpeechDat(II) database for triphone-modelling.

Ultimately, the success of the Indtal-browser must be measured based on still unanswered questions like:

- How many mobile-impaired people use Indtal in one year from now?
- Is Indtal still used, available for download, and maintained in e.g. four years from now?

For the Department of Communication Technology the Indtal-project has also been a welcome opportunity to establish some general Danish speech recognition resources that can be used in other projects and serve educational purposes.

## References

- Cameron, Hugh, (2000). "Speech at the interface". *Proceedings of the COST249 Workshop on Speech in Telephone Networks*, Ghent.
- Hadjadj, Djamel & Dominique Burger (1999). "BrailleSurf: An HTML Browser for visually handicapped people". In *Proc. of 14th conference on "Technology and Persons with Disabilities"*. Los Angeles.
- Jensen, Janne Jul, Lars Bo Larsen, Erik Aaskoven, Tom Brøndsted, Christian Gai Hjulmand, Børge Lindberg, Peter P. Pedersen (2005, April). *Bruger-evaluering af indtal.dk*, Internal Report, Aalborg University.
- Lindberg, Børge & Finn Tore Johansen (2), Narada Warakagoda (2), Gunnar Lehtinen (3), Zdravko Kacic, Andrej Zgank, Kjell Elenius, Giampiero Salvi. (2000). A Noise Robust Multilingual Reference Recogniser based on SpeechDat(II). *ICSLP 2000*.
- Mihelic, France & Nikola Pavesic, Simon Dobrisek, Jerneja Gros, Bostjan Vesnicer, Janez Zibert. Homer (2002). *A Small Self Voicing Web Browser for Blind People*. Laboratory of Artificial Perception, Systems and Cybernetics Faculty of Electrical Engineering, University of Ljubljana, Slovenia,
- Robin, Michael B. & Charles T. Hemphill. Considerations in Producing a Commercial Voice Browser, *W3C WS on "Voice Browsers"*. Massachusetts, 1998.
- Schwarz, Emmanuel, Gaële Hénault, Dominique Burger. *BrailleSurf 4*, [www.snv.jussieu.fr/inova/bs4/](http://www.snv.jussieu.fr/inova/bs4/), visited March 2005
- W3C: *Web Accessibility Initiative (WAI)*, [www.w3.org/WAI/](http://www.w3.org/WAI/) visited March 2005.
- Young, S., Valtchev, V., & Woodland, P. (1997, March). *The HTK book* (for HTK Version 2.1) Entropic Cambridge Research Laboratory,
- Zajicek, M. & I. Venetsanopoulos. (2000). Using Microsoft Active Accessibility in a Web Browser for the blind and visually impaired. *Proc. of the Annual International Conference "Technology and Persons with Disabilities"*, Los Angeles.

## Biography

**Tom Brøndsted** is associate professor affiliated at Department of Communication Technology, Aalborg University. He has published over 70 papers and articles mainly in the area of speech recognition and speech understanding, natural language parsing partly in a multimodal context, grammar formalisms, etc. He has implemented parsers and other software components used in various national and international research projects. Homepage <http://kom.aau.dk/~tb>

## Author's Address

*Tom Brøndsted  
Department of Communication Technology,  
Aalborg University  
Niels Jernes Vej 12,  
DK-9220 Aalborg Ø  
Denmark  
phone: office: A6-321, tel. +45 96 35 86 36  
e-mail: [tb@kom.aau.dk](mailto:tb@kom.aau.dk)*



# LINGUISTIC FUNCTIONS OF HEAD NODS

*Loredana Cerrato*

Department of Speech, Music and Hearing  
KTH, Stockholm, Sweden

## **Abstract**

*The aim of the present study is to investigate which communicative functions head nods can have in spoken Swedish.*

*By nod is here meant a vertical down-up movement of the head.*

*To classify the communicative functions of head nods 10 short video-recorded Swedish dialogues were analysed and labeled. The labels used are referred to the different communicative functions that the head nods carry out in the given context.*

*The results show that the most common function carried out by head nods is that of feedback. Beside feedback function, head nods can be produced to signal turn taking, focus and emphasis, to give affirmative responses and to show courtesy.*

*The visual information carried out by head nods in spoken communicative interactions is without doubt extremely important; therefore it should be exploited in the field of human-machine interfaces. This could be done by integrating head nods in the design and development of embodied conversational agents. Thanks to the production of head nods, embodied conversational agents might become more effective and appear more natural during their interactions with human beings.*

**Keywords:** Head nods, communicative feedback

## **1. Introduction**

In a previous study of head movements [Cerrato & Skhiri 2003] carried on four dialogues recorded with a similar set-up as in the present study, it was

shown that head nods are the most common head movements produced during semi-spontaneous conversations in Swedish, and their function is mostly that of giving continuation feedback. This preliminary analysis of head movements related to linguistic communicative functions also shows evidence that it is possible to identify a general pattern for each specific head movement (for instance for head nods, jerks or tilts) even if it is not possible to establish a one-to-one correspondence between a head movement and a specific communicative function. In fact different movements can be produced with the same communicative function (for instance a nod and a jerk can be used to give feedback continuation of contact) and the same movement (for instance a head nod) can be produced with different feedback functions, for instance to give continuation feedback or to show agreement.

McClave's [2000] analysis of dialogues of American English speakers illustrate that the head movements have predictable patterns and have semantic, discourse and communicative functions, as for instance convey propositional content, carry out semantic meanings beyond affirmation and negation and elicit and give feedback. In particular McClave underlines that most of the head nods produced to give feedback are a response to previous request of feedback, in terms of head nods.

The aim of the present study is to investigate in more detail head nods, in order to find out which specific communicative functions, beside feedback, they can carry out in spoken Swedish and provide a precise description of head nods in term of their shape and duration, which might be exploited for the implementation of more natural head nods in the design of embodied conversational agents.

The results of previous study of short feedback verbal expressions such as *yes*, *mm* and *ok* in languages such as English, Italian and Swedish [Jurafsky et al. 1998, Cerrato 2003] suggest that these short verbal expressions, with a feedback continuation function, show shorter duration and lower energy than other more complex verbal feedback expressions having more complex feedback functions such as agreement, expressive and so on. These short verbal expressions are intended to be unobtrusive and have the simple function of showing that the interlocutor is following the interaction and is not yet willing to express a judgment, show agreement, disagreement or to take the floor.

The analysis of the gestures related to verbal feedback expressions in four spontaneous conversations recorded in a travel agency in Sweden [Allwood & Cerrato 2003] showed that in 67% of cases accompanying gestures consisted of head nods, either single nods or multiple nods.

Starting from these previous results it is possible to hypothesize that minimal gestures are related to minimal feedback expressions having the feedback function of showing continuation of contact, while more extensive and complex gestures are produced together with more composite verbal feedback expression that carry out feedback function other than showing continuation of contact.

## **2 Materials and Methods**

### ***2.1 Materials***

10 short dialogues of the length of 1 minute each in a travel agency scenario were used to carry out the present study. These dialogues are part of the KTH-PF-Star Multimodal corpora [Beskow et al 2004]. The dialogues were recorded by means of an opto-electronic motion tracking system (QUALISYS MacReflex<sup>1</sup>). Thanks to reflective markers attached on the speaker's face the system is able to record, with 4 infrared cameras, the displacement and transition speed for each marker from a position to another, every 17 milliseconds. This way the dynamics of every facial expression is captured with high precision.

A total of 29 IR-sensitive markers were attached to the speaker's face, of which 4 markers were used as reference markers. The marker setup corresponds to MPEG-4 feature point (FP) configuration. Thanks to the reflective markers it is possible to record the 3D positions for each marker with sub-millimetre accuracy.

Starting from the four camera recordings, the 3D reconstruction is done automatically. Audio data is recorded at the same time on DAT tape. Video-recording are also performed by means of Sony DV video-cameras.

The 10 short dialogues were recorded in order to provide materials for the analysis of spontaneous communicative visual expressions. Two

---

<sup>1</sup> Qualysis: <http://www.qualysis.se> (March 05)

participants were instructed to interact with each other pretending to be in a travel agency. They were following a script with ten different situations which were thought to lead to different expressions of emotions and production of different communicative dialogic expressions. The focus of the 3D recording was on the participant with the markers on his face, whom from now on will be referred to as speaker A. However the other speaker (from now on speaker B) was also recorded, since two SONY DV digital video cameras were used.

The dialogues can be defined as semi-spontaneous since the two speakers had a short script describing the scenario and the task to perform, and they had to improvise the dialogue.

The two participants alternate in their roles, this way in 5 dialogues speaker A plays the role of travel agent, while speaker B plays the role of the customer, and in the other 5 dialogues the roles are switched.

All the dialogues have the length of 1 minute (and this is due to the fact that the recording system cannot record chunks longer than 60 seconds). The dialogues consist of about 10-16 contributions per interlocutor. The total number of labeled head nods is 92 for speaker A and 105 for speaker B.

## **2.2 Method**

Annotation, segmentation and measurement of the duration of head nods in the audio-visual material were carried out with the help of the software package “Wavesurfer” [Sjolander & Beskow 2000] provided with a video plug-in that allows seeing the video recordings in .mpeg format. Moreover on a panel synchronized to the video and to the waveform it is possible to display the chosen location dimension of the 3D data. The marker on the nose tip was used as reference for the detection of nods.

Temporal values were measured both from waveforms and, when possible, from the relative displacement of the marker on the nose tip.

A multi-layer coding scheme was adopted to annotate the type and function of the head nods, synchronized with the possible verbal expression produced at the same time.

The head nod type could be a single nod (S-Nod) or a repetitive, multiple nod (R-Nods). Several functions categories were defined *a priori*, based on previous observation of head nods and on literature references [Allwood & Cerrato 2003, McClave 2000, Knapp & Hall 2002]. Beside feedback functions, turn managing, sequencing, and courtesy function were defined and also giving affirmative responses and signaling focus and emphasis were considered as possible functions of head nods.

Table 1 shows a scheme of the detailed coding for the functions of the head nods and the relative labels used in the annotation.

Feedback is here intended in accordance to [Allwood et al. 1993] as a mechanism “which enables the participants of a conversation to exchange information about four basic communicative functions: *contact, perception, understanding* and *attitudinal reactions*”. The general category of feedback is further divided into sub-categories depending on the specific explicit function that the head nod carries out in the given context [Cerrato 2004].

Turn managing functions are an elaboration of those proposed by Duncan [1974]. Often feedback and turn managing function are interwoven, because a feedback expression can be used to show continuation of contact and no intention to take the floor, or can show continuation of contact and at the same time signal the intention to get the floor.

As concerning affirmative responses, the difference between a positive feedback and an affirmative response is quite subtle, however the criteria followed to assign the label of affirmative response was that of looking for a positive answer to a polar question.

Usually at the end of an interaction the interlocutors greet and thank each other by saying some courtesy words, which are often accompanied by the production of a single slow head nod, which has the function of showing politeness, courtesy.

The category filler refers to those head nods that might be produced in own communication management that is hesitations and self-corrections [Allwood 2001]

The category batonic refers to those cases in which head nods are produced to signal focus on words or constituents or to signal emphasis.

Table 1. Coding scheme for the functions of head nods relative labels

<b>FUNCTION CATEGORIES</b>	<b>LABELS</b>
<b>FEEDBACK</b>	
Give Continuation (I go on)	<b>FBCI</b>
Give Continuation (you go on)	<b>FBCY</b>
Give Acceptance	<b>FBA</b>
Give Expressive	<b>FBEX</b>
Give Refusal	<b>FBR</b>
Elicit Req. Conf	<b>FBREQ</b>
<b>TURN MANAGING</b>	
Turn offer	<b>TGO</b>
Turn submit	<b>TGS</b>
Turn taking (accept)	<b>TTA</b>
Turn taking (gain)	<b>TTG</b>
Turn Request	<b>TR</b>
Turn Maintain	<b>TM</b>
<b>BATONIC</b>	
To mark word or constituents with focus	<b>BFocus</b>
To signal emphasis	<b>BEmph</b>
<b>POLITENESS</b>	
Show reverence	<b>Pol</b>
<b>AFFIRMATIVE RESPONSE</b>	<b>RP</b>
<b>FILLER</b>	<b>SFill</b>

### *2.2.1 Distributional analysis*

In order to be able to code head nods and their functions it is necessary to identify them in the first place. To do so it is crucial to carefully analyse the video-recordings and take contextual information into account, which means interpreting and categorising head nods in terms of explicit reactions to the previous communicative act. For speaker A, the one with the markers on his face, it was possible to access the 3D data and have a more complete picture of the type of produced head movement and of their starting and ending point. For speaker B the identification relied only on the visual information available in the video recordings.

Table 2a and 2b show the distribution of head nod types and their function respectively for speaker A and B in those five dialogues in which Speaker

A plays the role of travel agent and speaker B that of customer. S-Nod means a single nod, while R-Nods means a repetitive nod.

Head nods produced with the function of feedback are easy to identify in the dialogues, since they are mostly produced contemporarily with the production of short verbal feedback expressions such as: *yes, mh, certainly* and similar.

For the head nods having other functions than feedback sometimes identification could be problematic, since they are produced together with different words or utterances and often simultaneously with other kind of movements (as for instance with a forward or backward movement of the whole trunk, with eyebrow movements, and other facial expressions) or as a continuum sequence with other movements, as for instance before or after a jerk (i.e. a fast backward movement of the head) or a tilt (i.e. a single movement of the head leaning on one side) and so on.

Each labeled head nod was assigned a function label.

Table 2a. Distribution of the type and function of the head nods produced by speaker A in the dialogues in which he plays the role of travel agent

Function	Type		Total
	S-Nod	R-Nods	
<b>FBA</b>	5	10	15
<b>FBCY</b>	7	2	9
<b>FBREQ</b>	3	3	6
<b>Pol</b>	6		6
<b>Bemph</b>	4	2	6
<b>Bfocus</b>	3	2	5
<b>FBR</b>		3	3
<b>FBCI</b>	1	1	2
Total			52

Table 2b. Distribution of the type and function of the head nods produced by speaker B in the dialogues in which he plays the role of customer

Function	Type		Total
	S-Nod	R-Nods	
<b>FBA</b>	5	9	14
<b>FBREQ</b>	2	8	10
<b>FBCY</b>	4	3	7
<b>Bfocus</b>	5	2	7
<b>Bemph</b>	4		4
<b>Pol</b>	4		4
<b>FBEX</b>		3	3
<b>TGO</b>	1		1
<b>FBR</b>	1		1
<b>RP</b>	1		1
<b>Total</b>			52

Table 3a and 3b show the distribution of head nod type and their function respectively for speaker A and B in those five dialogues in which Speaker A plays the role of customer and speaker B that of travel agent.

Table 3a. Distribution of the type and function of the head nods produced by speaker A in the role of customer

Function	Type		Total
	S-Nod	R-Nods	
<b>FBA</b>	3	7	10
<b>FBREQ</b>	1	9	10
<b>FBCY</b>	8		8
<b>Bfocus</b>	3		3
<b>FBCI</b>	1	2	3
<b>Bemph</b>	2		2
<b>FBR</b>	2		2
<b>Pol</b>	1		1
<b>FBEX</b>	0	1	1
<b>Total</b>			40

Table 3b. Distribution of the type and function of the head nods produced by speaker B in the role of travel agent

Function	Type		Total
	S-Nod	R-Nods	
<b>FBA</b>	6	8	14
<b>FBREQ</b>	3	8	11
<b>FBCY</b>	4	1	5
<b>FBCI</b>	2	1	3
<b>Bfocus</b>	7		7
<b>Bemph</b>	4		4
<b>Pol</b>	4		4
<b>TGO</b>	3		3
<b>SFiller</b>	1		1
<b>RP</b>	1		1
<b>Total</b>			53

In total 70% of the times head nods have been assigned a feedback function, in particular giving continuation feedback (FBCY), giving agreement (FBA), and request feedback (FBREQ). Beside feedback it is quite frequent the production of head nods to signal focus (Bfocus).

For Speaker A single nods have been assigned in 31% of cases the function of feedback continuation of contact (FBCY) and these single head nods, which often accompany short expressions such as *mm, ja {j}a*, have a mean duration of 0,44 msec.

27% of times the label assigned was that of feedback acceptance (FBA), 17% focus and emphasis (Bfocus, Bemph) and 7% of times that of showing courtesy, politeness (Pol). In this last case the head nod is slower, with an average duration 0,6 msec.

Speaker A in 5% of cases produces a single head nod with the function of giving a negative feedback (FBR), which means to refuse the information received either because of misunderstanding or disagreement. Usually negative feedback is realized with shakes rather than with a head nod, however in these cases the nod is always in a continuum with other gestures, among which a shake or a tilt.

Speaker A produces repetitive nods mostly as a request for feedback (FBREQ) and to give acceptance feedback (FBA).

For Speaker B, single nods have been assigned only 10% of the times the feedback continuation of contact label (FBCY), and in 27% of cases the

function assigned was that of feedback acceptance (FBA). Single nods with FBA function were, for speaker B, often accompanying short expressions such as *mm, ja, ok*.

In 20% of cases the label assigned to the single nods produced by speaker B was that of requesting feedback (FBREQ), in 17% of cases that of signaling focus (BFocus) and in 6% of cases the function was that of showing courtesy, politeness (Pol). Also speaker B produces repetitive nods mostly as a request for feedback (FBREQ) and to give acceptance feedback (FBA), but also to give feedback continuation of contact (FBCY).

Tables 4a and 4b show the average duration and standard deviation for single head nods according to different functions. These results show that the head nods (FBCY), without showing agreement, or acceptance and without showing the intention to get the floor are shorter compared to the single nods produced with other functions.

The longest single nods are those produced together with courtesy words (POL). Several examples of a single nod produced to accompany a “thank you” or a greeting word at the end of an interaction have been labelled in the 10 dialogues, in average these head nods that accompany courtesy words, have a duration of 0.71 msec. Given its characteristics this head nod can be interpreted as a “reduced form” of a courtesy bow.

Table 4a. Duration of the single nods per function for Speaker A

<b>Function</b>	<b>Duration in msec</b>	<b>Standdev</b>
POL	0,65	0,19
FBA	0,64	0,22
BFocus	0,58	0,2
FBCY	0,37	0,16

Table 4b. Duration of the single nods per function for Speaker B

<b>Function</b>	<b>Duration in msec</b>	<b>Standdev</b>
POL	0,78	0,14
BFocus	0,57	0,16
FBA	0,50	0,17
FBCY	0,43	0,1

### 3. Semantic analysis

In this session we are going to look at some examples extracted from the dialogues to understand when the head nod is produced and what is its function/meaning in the given context.

The first 3 examples are excerpted from dialogue 3003, in which speaker A plays the role of the customer and speaker B that of the travel agent. In this dialogue, which counts a total of 30 contributions, speaker A wishes to buy a packet-travel to Italy. He has previously been in contact with the travel agent and in the situation they meet in the agency to look at some possibilities and take a decision. In example 1 speaker B says to the travel agent: *du skulle titta up på några paketer för mig och min fru* (you should look for some packets for me and my wife). The word *paketer* (packets) carries the focus and is accompanied by a single head nod which has been labeled as Bfocus. At this request the travel agent (speaker B) replies *ja précis* (yes exactly), which has the function of a continuation feedback, which shows the intention of keeping the floor. Speaker B does not produce any head nod in this contribution, but he raises his eyebrows to show that he wishes to keep the floor, he in fact continues by illustrating the possible alternatives in the following contribution, however before he continues to speak, speaker A produces a short m-like expression, with the function of showing continuation feedback and no intention to take the floor. This expression is accompanied by a short single head nod of the duration of 0,22 msec. This kind of head nods are quite minimal: they have an average duration of 0,40 ms and they usually accompany short verbal feedback expressions such as *yes,mm, ok*. These short verbal expressions, with a continuation function (FBCY) show shorter duration and lower energy than other more complex → composite verbal feedback expressions having even more complex feedback functions (agreement, expressive). This is because they are intended to be unobtrusive and simply show that the interlocutor is following the interaction and is not yet willing to express a judgment or to take the floor.

Another example of a continuation feedback intended to show continuation of contact and no intention to take the floor is in the next contribution by speaker A, who says *ja precis* (yes exactly) while the speaker B is illustrating the possible alternatives for the travel, in this case however the feedback verbal expression is accompanied by repetitive head nods (2 for precision).

A 6.8500 9.5567 du skulle titta up på några **paketer (S-Nod/Bfocus)** för mig och min fru

B 9.6396 10.2948 ja *précis* (EbRaise/FBCI)

A 10.3250 10.5500 mm (**S-Nod/FBCY**)

M 10.369 13.221 det var lite olika alternative där med all-inclusive och ehm/

A 13.0250 14.1750 {j}a *precis* (**R-Nods/FBCY**)

B 14.2286 14.5628 {j}a// och det beror lite gran på vilken // vad ni kommer att välja då om ni vill åka till **Torino (S-Nod/BFocus)** eller om ni vill åka ner kanske södra delen mot Rom

*Example 1 from dialogue 3003*

The second excerpt from dialogue 3003 offers instances of other categories of feedback. In the first contribution speaker A, the customer, asks *ja men vad har du för alternativ då* (yes but what do you have for alternatives then) -referring to the packet travels- and the travel agent answers with a *ja* (yes) accompanied by a phonological lengthening and a short single head nod. The phonological lengthening signals that the travel agent wishes to keep the floor and go on speaking to fulfil the customer request, so he says *det finns den här tiodagarsresa då* (there is this 10-day packet) and he puts the focus on *tiodagarsresa* producing a short single head nod which is also a request for feedback. In fact speaker B, as a reply, produces a short m-like sound accompanied by a short single head nod that shows continuation of contact even if not yet agreement and acceptance, since the information given by speaker B is not complete yet. The rest of the information comes in the following contribution when the travel agent says that the packet costs SEK 8000 per person.

A 39.6443 43.0000 ja men vad har du för alternativ då

B 40.177 43.3120 jaa / /(**S-Nod/FBCI**)det finns den här **tiodagarsresa** då/ (**R-Nods/BFocus-FBREQ**)

A 45.6000 46.2250 mm (**S-Nod/FBCY**)

B 45.6280 47.2010 som / den **kostar** (**S-Nod/BFocus**) SEK 8000 per person

*Example 2 from dialogue 3003*

The third example from dialogue 3003 is an instance of a single head nod with a FBA function, which means acceptance, agreement. Speaker B, the customer, asks *vilken flygbolag är det?* (which flight company is it?) and the travel agent answers *{j}a det är SAS* (yes, it is SAS), emphasizing the word SAS which is simultaneously produced with a short single head nod. SAS is the Scandinavian Airlines, which in this context represents a kind of

warranty for a good and safe flight. To this the customer says *ok*, accompanied by a short single nod, which has the function of acceptance, agreement (FBA).

B 48.9250 50.8500 ah men det låter jätte bra, vilken flygbolag är det?

A 51.2019 51.9409 {j}a det är SAS (**S-Nod/BEmp**)

B 51.9000 52.2000 ok (**S-Nod/FBA**)

*Example 3 from dialogue 3003*

Example 4, is excerpted from another dialogue, 4005, and reports an instance of the feedback category FBEX, which means a feedback that explicitly shows an expressive/emotional reaction.

In this dialogue Speaker B is the customer, he is complaining about the fact that during a holiday in a holiday resort managed by the travel company he and his wife got stomach flu, as a consequence he is asking for a reimbursement. The travel agent tries to clarify that they cannot pay any refunds since they cannot be sure that the stomach flu was actually caused by the food and drinks consumed at the resort. As a reaction to this refusal by the travel agent the customer, quite annoyed, almost blackmails the travel agent by saying: *ok då får jag väl vända mig till konsumentombudsmannen då* (*then I will contact the Consumer Protection Organization*) producing repetitive head nods and with an annoyed tone of voice. These repetitive nods have been labeled as Expressive feedback and also request for feedback, since the customer is not only showing his annoyed attitude, but he is actually trying to get a reaction from the travel agent. To this the travel agent replies: *ehm ja och det är det är väl kanske det du får väl göra då* (*ehm ok maybe this is what you can do*) and also produces repetitive head nods. This is also a good example of how the production of head nods by one interlocutor triggers the production of head nods of the other interlocutor.

B 44.4181 48.2246 ok då får jag väl vända mig till konsumentombudsmannen då (**R-Nods/FBEX [annoyed], FBREQ**)

A 48.6404 52.3250 ehm ja oh det är det är väl kanske det du får väl göra då (**R-Nods/FBA**)

*Example 4 from dialogue 4005*

In the example 5, from dialogue 3002, the two interlocutors have concluded their interaction and the travel agent says that he is going to book the ticket they agreed on (*då bokar jag en sån*) to which the customer reacts with an

*ok tack så mycket (ok thank you so much)* and produces a head nod while uttering *tack (thanks)*, to which the travel agent replies with a *tack* accompanied by a single head nod. These two head nods have been labelled as politeness (Pol), since they have been produced to thank the interlocutor and show courtesy.

B	58.564	59.0765	då bokar jag en sân		
A	59.4250	59.8250	ok tack så mycket	<b>(S-Nod/Pol)</b>	0.76 msec
B	59.816	60.211	tack	<b>(S-Nod/Pol)</b>	0.67 msec

*Example 5 from dialogue 3002*

#### 4. Conclusions

The results of this study aiming at investigating the communicative function of head nods in Swedish dialogic speech shows that in 70% of the analysed cases the function of head nods is related to feedback. Beside feedback head nods are produced to signal focus and emphasis, show courtesy, signal turn taking and give affirmative responses.

The initial hypothesis that short, minimal head nods might be related to short verbal feedback expression carrying out the function of giving continuation feedback, seems to be proven only in the case of speaker A. In fact if it is true that minimal head nods are mainly produced to accompany short verbal expressions having the function of showing feedback continuation of contact, it is also true that for speaker B, minimal single head nods are produced also when the function of feedback is to show acceptance/agreement. Once again the results show that it is possible to identify a general pattern for head nods but it is not possible to establish a one-to-one correspondence between a head nod and a specific communicative function. However the results of the duration analysis shows that single head nods can have different duration which seem to be correlated to the different communicative functions. If these differences in the duration will be proven to be significant, they could represent a distinctive cue for the different communicative functions that head nods can carry out, cue which could be then exploited in the implementation of communicative head nods in talking heads used in human machine interfaces.

## References

- Allwood, J., Nivre, J., & Ahlsén, E. (1992), On the Semantics and Pragmatics of Linguistic Feedback in *Journal of Semantics*,
- Allwood, J., (2001). Dialog Coding - Function and Grammar: Göteborg Coding Schemas. *Gothenburg Papers in Theoretical Linguistics 85*, Dept of Linguistics, University of Göteborg, 1-67
- Allwood, J., Cerrato L, (2003), A study of gestural feedback expressions. The First Nordic Symposium on Multimodal Communication, Copenhagen First Nordic Symposium on Multimodal Communication, Paggio, P., Jokinen, K., Jönsson, A., (Eds), Copenhagen, 23-24 September 2003, 7-22
- Beskow, J., Cerrato, L, Granström B, House, D, Nordstrand, M., & Svanfeldt, G. (2004). The Swedish PF-Star Multimodal Corpora, *LREC Workshop on Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces*, Lisboa 25 May 2004, 34-37
- Cerrato, L., Skhiri, M. (2003), A method for the analysis and measurement of communicative head movements in human dialogues. *Proc of AVSP '03*, 251-256
- Cerrato, L. (2003) A comparative study of verbal feedback in Italian and Swedish map-task dialogues In Proceedings of the Nordic Symposium on the comparison of spoken languages, November 2003, Copenhagen 99-126
- Cerrato, L. (2004). A coding scheme for the annotation of feedback phenomena in conversational *LREC Workshop on Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces*, Lisboa 25 May 2004, 25-28
- Duncan, S. (1974) Some Signals and Rules for Taking Speaker Turns in conversations. In S. Weitz (ed.) *Nonverbal Communication*. New York:
- Jurafsky, D., Shriberg, E., Fox, B., & Curl, T. (1998). Lexical, Prosodic, and Syntactic Cues for Dialogue Acts. *Proceedings of ACL/COLING 98 Workshop on Discourse Relations and Discourse Markers*, 114-120, Montreal.
- Knapp, M., Hall, J. A. (2002). *Nonverbal Communication in Human Interaction* (FIFTH EDITION) 2002 Wadsworth.
- McClave, E. Z. (2000). Linguistic functions of head movements in the context of speech, in *Journal of Pragmatics*, 32, 855-878 Oxford University Press, 298-311
- Sjölander, K. & Beskow, J. (2000). "WaveSurfer - an Open Source Speech Tool", in *Proceedings of ICSLP 2000*, Beijing, China, 2000.

## **Biography**

**Loredana Cerrato** is a PhD student affiliated at the Department of Speech Music and Hearing and the Centre for Speech Technology of the Royal Institute of Technology (KTH) in Stockholm. Her research focuses on the analysis of non-verbal behaviour in unimodal and multimodal materials from human as well as human-machine interaction. Before starting her doctorate at KTH she worked at Telia Promotor Infovox in Stockholm (now Acapela group), developing speech synthesis systems.

## **Author's address**

*Loredana Cerrato  
Department of Speech, Music and Hearing  
TMH/KTH  
Lindstedtsv.24  
10044 Stockholm.  
e-mail: loce@speech.kth.se*

# A METHOD FOR THE DETECTION OF COMMUNICATIVE HEAD NODS IN EXPRESSIVE SPEECH

*Loredana Cerrato and Gunilla Svanfeldt*  
Department of Speech, Music and Hearing  
KTH, Stockholm, Sweden

## **Abstract**

*The aim of this study is to propose a method for automatic detection of head nods during the production of semi-spontaneous speech. This method also provides means for extracting certain characteristics of head nods, that may vary depending on placement, function and even underlying emotional expression.*

*The material used is part of the Swedish PF-Star corpora which were recorded by means of an optical motion capture system (Qualisys) able to successfully register articulatory movements as well as head movements and facial expressions. The material consists of short sentences as well as of dialogic speech produced by a Swedish actor.*

*The method for automatic head nods detection on the 3D data acquired with Qualisys is based on criteria for slope, amplitude and a minimum number of consecutive frames. The criteria are tuned on head nods that have been manually annotated. These parameters can be varied to detect different kinds of head movements and can also be combined with other parameters in order to detect facial gestures, such as eyebrow displacements.*

*For this study we focused in particular on the detection of head nods, since in earlier studies they have been found to be important visual cues in particular for signaling feedback and focus.*

*In order to evaluate the method a preliminary test was run on semi-spontaneous dialogic speech, which is also part of the Swedish PF-Star corpora and produced by the same actor who read the sentences. The results show that the parameters and the criteria that had been set on the basis of the training corpus are valid*

*also for the dialogic speech, even if more sophisticated parameters could be useful to achieve a more precise result.*

**Keywords:** Head nods, expressive speech

## 1. Introduction

Previous studies on head nods have mainly focussed on the analysis of the distribution and semantic function of head nods in conversational speech (McClave 2000, Allwood & Cerrato 2003) and even of the physical properties of head nods (Birdwhistell 1970, Cerrato & Skhiri 2003).

An attempt to quantify the extent of head movements was made by Birdwhistell (1970) who supposed that body movements, including head nods, are directly linked to linguistic structure and proposed a hierarchical system of units of movement in which lower-level units (kines) combined to form higher-level units (kinemes). A kine is an isolable feature of movement, while a kineme is defined as a group of movements having the same meaning in the American culture. In this scheme head nods were considered as distinct kinesic units. Birdwhistell estimated that a similar population of movers will make a full 15-degree nod in moments which can extend from about 0.5 sec to around 1.5 sec. The velocity, not the duration, is significant here. Birdwhistell defines also kinic variants (Hn) with a velocity range from about .8 degrees per frame (1/24 sec) to around 3 degrees per frame. Head movements outside these specifications belong to different units.

Investigating the American movement system Birdwhistell found it possible to isolate a series of ranges of variation which “modify” the kinesic structures and which have an analytic identity separate from these structures. These variations, termed the *motion qualifiers*, include:

*Intensity*: which delineates the degree of muscular tension involved in the production of a kine. Intensity is even sub-divided into five relative degrees of tension.

*Range*: which refers to the width or extent of movement involved in performance of a given kine. Range is also subdivided into five degrees.

*Velocity*: which refers to the temporal length (relative to the range) involved in the production of a kine. Velocity has only a three-degree scale.

Unfortunately it is unclear how the degree of nodding was measured. The angle is dependent on where the reference point is, for this reason it is difficult to compare other results to those proposed by Birdwhistell.

The possibility to use 3D data opened new opportunities to observe the characteristics of head movements. Cerrato & Skhiri (2003) proposed a method for the analysis and quantification of head movements in semi-spontaneous speech, by using the information recorded by means of the 3D recording. Their preliminary analysis of head movements related to linguistic communicative functions show evidence that it is possible to identify a general pattern for each specific head movement (for instance for head nods, jerks or tilts) even if it is not possible to establish a one-to-one correspondence between a head movement and a specific communicative function. In fact different movements can be produced with the same communicative function (for instance a nod and a jerk can be used to show feedback continuation of contact) and the same movement (for instance a head nod) can be produced with different functions, for instance to give feedback, to signal focus, to show courtesy and so on.

The production of head movements in speech has recently received attention not only in the field of linguistic studies, but also in the field of human-machine interactions. The visual information carried out by head movements, in particular by head nods in spoken communicative interactions, is without doubt extremely important, since it can carry out several communicative functions, such as showing attention, interest, disinterest (Harrison 1974), giving and eliciting feedback (Allwood & Cerrato 2003), signaling focus and emphasis and so on. Therefore it should be exploited in the field of human-machine interfaces. This could be done by integrating head nods in the design and development of embodied conversational agents. Thanks to the production of head nods embodied conversational agents might become more efficient and appear more natural during their interactions with human beings.

Graf et al. (2002) performed studies on head movements related to prosody, with the aim of improving the performance of a talking head. The movements were extracted from video recordings, and the material consisted of short read sentences and greetings. They found that patterns of head movements are strongly correlated with the prosodic structure of the text. However, the type of text highly influenced the head movements: when reading journal texts the head movements were less pronounced than during greeting utterances.

## 2. Materials and Method

### 2.1 Materials

Two different kinds of materials have been used for this study: on one hand read sentences uttered with controlled variable position of the focus, on the other hand semi-spontaneous dialogic speech. During the annotation of our corpus, head nods were found to be frequent in the utterances produced with a confirming expression, and therefore this material has been used for the training part of the study.

A total of 39 sentences uttered with a focused word, by a Swedish male speaker with a confirming expression were used for the tuning of the criteria for the selection of the parameters in our method. The short sentences were produced in a semantic neutral way, with a confirming expression and also with a varying focus, as in the following example:

Båten seglade förbi  
Båten seglade förbi  
Båten seglade förbi

10 short dialogues of the length of 1 minute each in a travel agency scenario were used to evaluate the head nods detector. The 10 short dialogues were recorded in order to provide materials for the analysis of spontaneous communicative visual expressions. Two participants were instructed to interact with each other, pretending to be in a travel agency. The focus of the 3D recording was on the participant with the markers on his face (the same actor who read the sentences). The two participants were following a script with ten different situations which were thought to lead to different expressions of emotions and production of several communicative dialogic expressions. The dialogues can be defined as semi-spontaneous since the two speakers had a short script describing the scenario and the task to perform, and they had to improvise the dialogue.

The sentences and the dialogues are part of the KTH-PF-Star Multimodal corpora (Beskow et al. 2004), and were recorded by means of an opto-electronic motion tracking system (Qualisys<sup>1</sup>). Using reflective markers attached on the speaker's face the system is able to record, with 4 infrared cameras, the position of each marker every 17 milliseconds. This way the dynamics of every facial expression is captured with high precision.

---

<sup>1</sup> Qualisys motion tracking system: <http://www.qualisys.se> (March 05)

A total of 29 infrared-sensitive markers were attached to the speaker's face, of which 4 markers were used as reference markers. The marker setup (as shown in figure 1) corresponds to MPEG-4 feature point (FP) configuration. Audio data was recorded on DAT-tape and visual data was recorded using a standard digital video camera and the optical motion tracking system.

Starting from the four camera recordings, the 3D reconstruction is done automatically. Audio data is recorded at the same time on DAT tape. Video-recordings are also performed by means of Sony DV video-cameras.

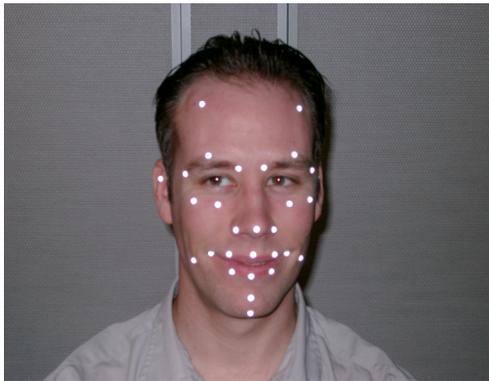


Figure 1. Position of the infrared markers.

### 2.1.1 Annotation of the material

The 15 sentences were manually annotated by two annotators who independently from each other identified the head nods and marked, in each sentence, the most prominent of the head nods. The annotation had a very high level of inter-agreement. However, since there exists no gold standard (the total number of presumptive nods is not known), the conventional algorithms for calculating inter-agreement were not applicable. Therefore it can be simply reported that both annotators found 139 nods each, of which 8 intervals of disagreement consisted of very vague nods.

The head nods in the 10 dialogues were identified and manually annotated by one annotator who assigned to each head nod a communicative function. Several functions categories were defined *a priori*, based on previous observation of head nods and on literature references. Functions include: feedback, turn managing, sequencing, and courtesy function plus giving affirmative responses and signaling focus and emphasis. Some of these categories have sub-categories (for a more detailed clarification of the functions and of the coding scheme see Cerrato 2005). However 70% of the head nods were assigned a feedback category.

Annotation of head nods in the audio-visual material were carried out with the help of the software package “Wavesurfer” (Sjölander & Beskow 2000) provided with a video plug-in, that allows the video files to be viewed in .mpeg format together with the 3D data.

## 2.2 Method

A very simple approach was chosen for head nods detection. The marker on the nose was used as reference for the detection of nods. It was thus assumed that the movements of this marker are representative for the movements of the head. Of course, other movements as well, such as changes in body posture, can cause the same displacement of the marker as a nod might do.

The chosen parameters were: (vertical) amplitude, length (in frames) and slope. Only the negative slopes were considered, since this is the most prominent feature of a nod. Therefore, the temporal length of a nod corresponds to approximately half the total length of the total cycle. In order to find suitable criteria for the parameters, only one parameter at a time was varied, and a record was kept concerning the “under/over” hit rate of the automatic process as compared to the manual marking of head nods done by human annotators.

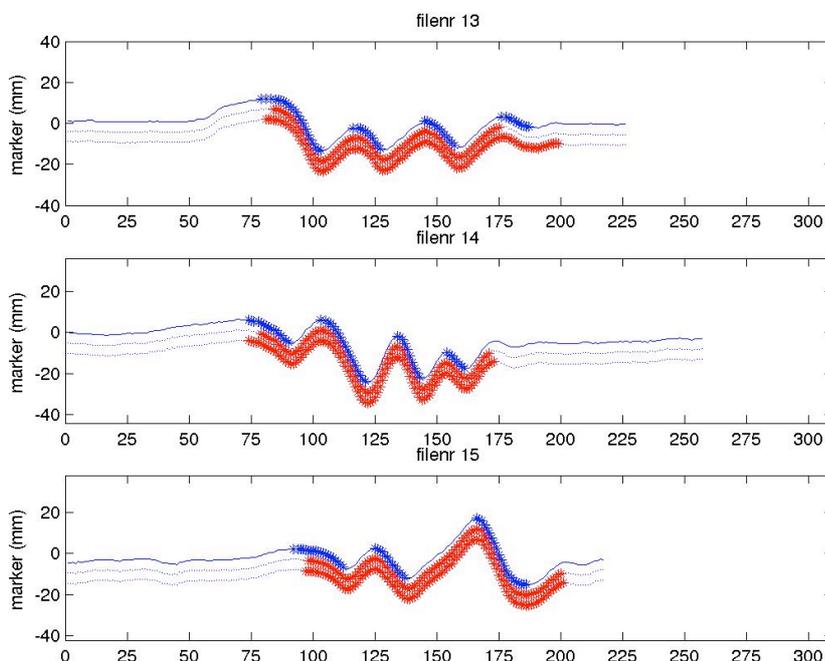


Figure 2. Head nods displayed as vertical location of the nose tip as function of time (in frames). The upper line is the actual location, where the bold line denotes the automatically found nods. The two lower thick lines indicate where the annotators have found nods.

Then the values that gave the best results were chosen. Since it was earlier noticed that in the sentences uttered with the confirming expression the actor produced lots of head nods, these sentences were chosen as training material. The minimum criteria values for the automatic detector that were reached were:

- minimum number of frames: 7 (=116ms)
- minimum amplitude: 4 mm
- minimum slope: -0.3 mm/frames (= -5mm/ms)

On 39 sentences with the confirming expression in total 139 nods were identified by the annotators, although some disagreement existed. Compared to one of the annotators, the script missed 4 and found 3 additional nods. Compared to the other annotator, 3 nods were missed out, and 2 additional were found by the automatic process.

The displacement of the marker on the nose tip during the sentence “Båten seglade förbi” (The boat sailed by) is showed in figure 2. The vertical movements are shown on the y-axis, and the frames are displayed on the x-axis. The upper line is the actual location of the marker for each frame, and the thicker (blue) line (“\*”) denotes where the automatic detector has found a nod. The lines beneath are the same trajectory, only displaced 5 (resp 10) mm for clarity reasons. The (red) bold lines on those correspond to the annotators marking of nods. This way the result can easily be compared.

### 3. Preliminary Evaluation

One skilled annotator had identified head nods and labeled their specific feedback function in the given context in the 10 dialogues. The automatic detection was carried out on the 10 dialogues and then the mismatches in the 2 annotations (manual vs. automatic) were analysed.

The evaluation of the method on the dialogues showed the complexity of spontaneous speech in comparison to the controlled sentences, on which the criteria for detection were based. In the dialogues there were a lot more head movements, and these were not as smooth and cyclic as in the sentences that were used for training. However, almost all annotated nods were found by the script. Since in general it seemed that the criteria were too low set for the dialogues, in comparison to the training corpus, one would assume that all nods should be found by the automatic detection. So we analysed the nods that were not found. When looking at the video, we

saw that some of these cases could (and probably would) in the given context be perceived as a nod, although the 3D-data show that there was no downwards movement.

To illustrate the difference in the use of head nods in the sentences as compared to the dialogues, see respectively figures 3a and 3b. In both figures the vertical displacement is shown on the y-axis, and the frames are displayed on the x-axis. The upper line is the actual location of the marker for each frame, and the thicker (blue) line (“\*”) denotes where the automatic detector has found a nod. In figure 3a, the lower (red) bold lines correspond to the annotators marking of nods. In figure 3b the (red) bold line is just one, since only one annotator manually identified nods in the dialogues.

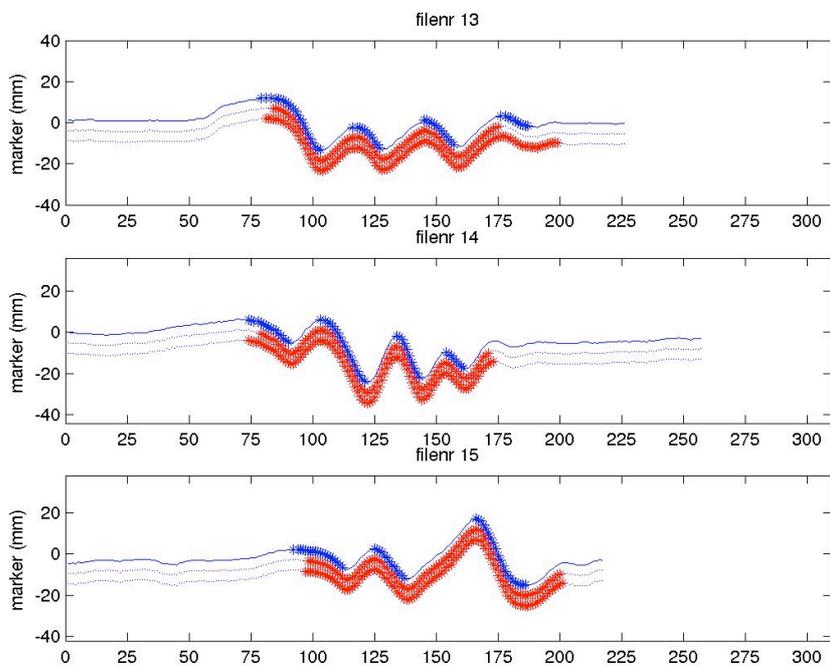
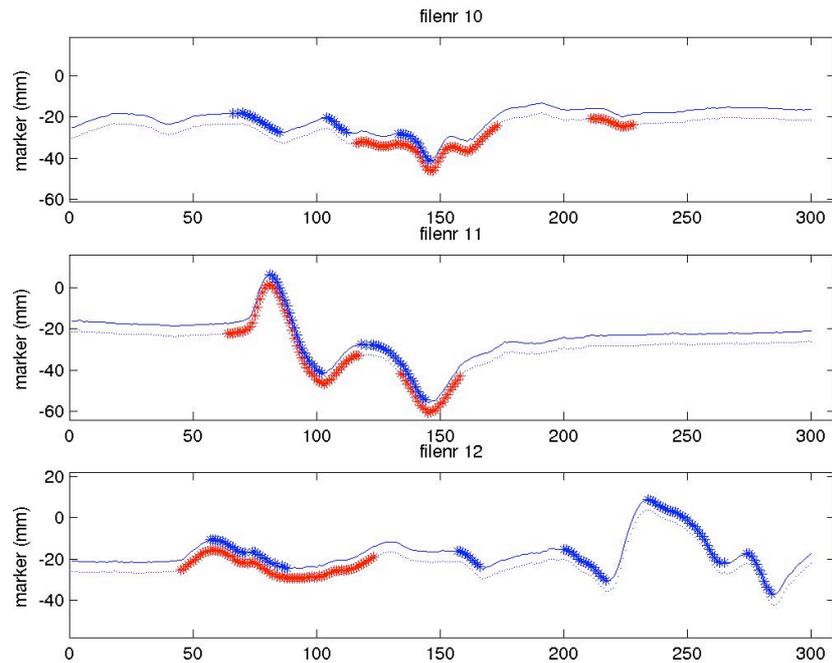


Figure 3a. Sentences where the head nods are easily detected.



*Figure 3b. Dialogic speech with complex head movements.*

As can be seen in figure 3b, several intervals were identified as nods by the automatic detection, although they were not marked as such by the manual annotation. One reason for the “over hit rate” of the detector might be that the criteria were too lax due to differences in controlled versus spontaneous speech. However the movements might not just be stronger; they may also be of a different character. What, to the automatic detector seems like a nod since only the vertical movement of the head is considered, might not have been manually annotated as such since additional sideway movements might have been present. To check whether there might be any truth in this explanation, another criterion was added to the detector. The criterion concerned maximum sideway displacement, this way in the second run of the automatic detector the sideways movements were restricted. The results of this sideway displacement restriction was that the number of “over hits” decreased by almost 48%, but some of the previous correctly annotated nods were also lost (22%). An example of the reduction of the “over hits” is reported in figs. 4a and 4b. So apparently, there are some other factors that still need to be considered for an automatic detection of nod, maybe the distinguishing factor is not in the movement itself, but in the linguistic-conversational context.

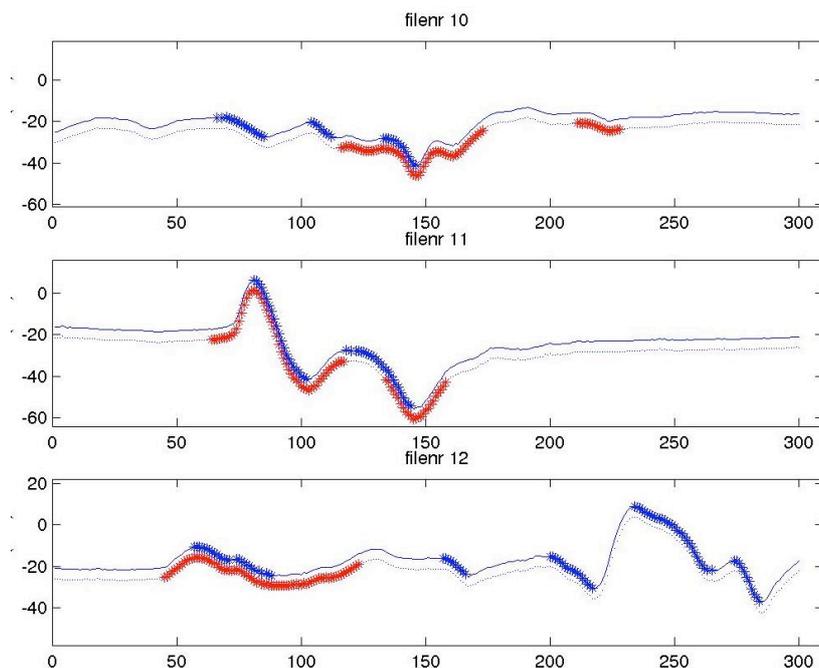


Figure 4a. Three segments of a dialogue, the detection of nods is done with the original parameters.

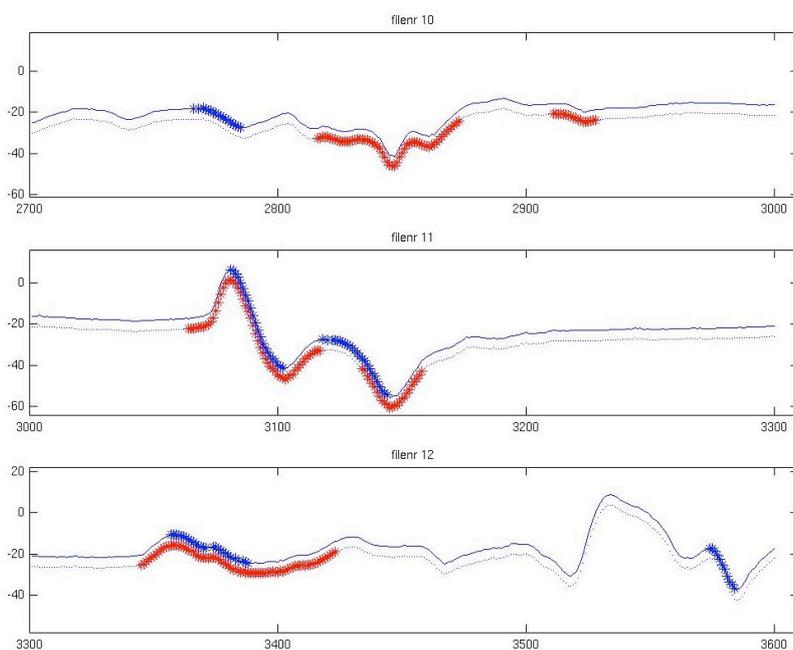


Figure 4b. Three segments of the same dialogue as in 4a, but with an additional criterion in the x-direction.

#### 4. Conclusions and Future Work

The results of the preliminary evaluation of the automatic detector of head nods showed the complexity of spontaneous speech in comparison to the controlled sentences, on which the criteria for detection were based upon. The automatic detector is able to identify head nods, but so far it does not succeed in recognizing different types of head nods related to different communicative functions.

In this preliminary study, only criteria for minimum values were used, but in order to separate different nod functions, it is likely that also maximum criteria should be used. Also, the parameters were treated one by one, and perhaps more precise results would be achieved if the combinations were to be considered and the criteria set according to that.

One feature that would need to be added is repetitiveness, since both annotators seem to have less constraints on, for example, amplitude when the nod is in a series of nods.

There is also the question of how much sideways movement that is to be allowed for a head movement to be considered a head nod. When sideway movements are restricted, some of the movements that have been considered to be a head nod by the annotators are not “let through” the automatic process.

#### References

- Allwood, J., & Cerrato, L. (2003). A study of gestural feedback expressions *First Nordic Symposium on Multimodal Communication*, Paggio P. Jokinen K. Jönsson A. (Eds), Copenhagen, 23-24 September 2003, 7-22.
- Beskow, J., & Cerrato, L., Granström, B., House, D., Nordstrand, M., & Svanfeldt, G. (2004). The Swedish PF-Star Multimodal Corpora in *LREC Workshop on Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces*, Lisbon 25 May 2004 34-37.
- Birdwhistell, R. L.(1970). *Kinesics and context*. Philadelphia, PA: University of Pennsylvania Press.

- Cerrato, L. (2005): Linguistic functions of head nods, In *Proc. of the 2<sup>nd</sup> Nordic Symposium on Multimodal Communication* (in press)
- Cerrato, L., & Skhiri, M. (2003). A method for the analysis and measurement of communicative head movements in human dialogues. *Proc of AVSP '03*, 251-256
- Graf, H. P., Cosatto, E., Strom, V., & Huang, F.J. (2002). Visual Prosody: Facial Movements Accompanying Speech. Proceedings of the fifth IEEE International Conference on Automatic Face and Gesture Recognition (FGR'02), Washington D.C.
- Harrison, P. R. (1974) *An introduction to non verbal communication* Prentice Hall New Jersey
- McClave, E. Z. (2000): Linguistic functions of head movements in the context of speech, in *Journal of Pragmatics* 32(2000) 855-878
- Sjölander, K., & Beskow, J. (2000). "WaveSurfer - an Open Source Speech Tool", in *Proceedings of ICSLP 2000*, Beijing, China, October 2000,

## Biography

**Loredana Cerrato** has a degree in foreign languages, literature and linguistics and is a PhD student affiliated at the Department of Speech Music and Hearing and the Centre for Speech Technology of the Royal Institute of Technology (KTH) in Stockholm and the Graduate School of Language Technology (GSCI). Her research focuses on the analysis of non-verbal behaviour in unimodal and multimodal materials from human as well as human-machine interaction. Before starting her doctorate at KTH she worked at Telia Promotor Infovox in Stockholm (now Acapela group), developing speech synthesis systems.

**Gunilla Svanfeldt** is a PhD student affiliated with the Department of Speech Music and Hearing and the Centre for Speech Technology of the Royal Institute of Technology (KTH) in Stockholm. Her research focuses on the analysis of visual correlates to emotional speech.

### Authors' addresses:

*Loredana Cerrato*  
*Department of Speech, Music and Hearing*  
*TMH/KTH*  
*Lindstedtsv.24*  
*10044 Stockholm.*  
*e-mail: loce@speech.kth.se*

*Gunilla Svanfeldt*  
*Department of Speech, Music and Hearing*  
*TMH/KTH*  
*Lindstedtsv.24*  
*10044 Stockholm.*  
*e-mail: gunillas@speech.kth.se*



# DESIGNING MULTIMODAL COMMUNICATION SYSTEM FOR FIREFIGHTERS

*Fang Chen*

Chalmers University of Technology  
Gothenburg, Sweden

## **Abstract**

*Presentation of information and communication between the members of a rescue team (firefighters) at the accident site to the rescue control centre is crucial for the success of the rescue operation. The present information and communication systems for firefighters are, by their own words, at the level of the “stone age”. Basically, there is no external visual information available for rescue members at the accident site. It is necessary to redesign their communication system in order to support their rescue work. This report presents a multimodal interaction design of the portable information communication system for firefighters to use in the rescue field. The user-centered design process is applied. It started by site visit and interviews with the rescue members to identify the design requirements and highlight system problems. Based on the requirements study, the multimodal interaction communication system is proposed for use by experienced firefighters. After using the cognitive walkthrough methods with the experienced firefighters, a prototype of the design will be tested by the firefighters. Many research questions regarding the multimodal interaction design continue to arise. This project is still in the early stage of its development. Here we will report the structure of the prototype and the possible research questions.*

**Keywords:** Multimodal interface, firefighter, communication

## 1. Introduction

For any natural and man-made disasters a quick, accurate damage assessment information, correct decision making and effective organization are necessary for speedy and effective rescue work. In most cases, communication for getting damage information and organizing the rescue work depends on telephone or mobile phone service (Fujiwara, 2004; Harnett, 2004). The heavy phone use by the general public causes sudden and severe congestion in the phone system.

In Sweden, as well as in many other countries, the communication for rescue performance on the disaster site is via analogical radio channels to avoid overloading the telephone system. A digital radio network system, the RAKEL (Radiokommunikation för Effektiv Ledning) project is being constructed right now. It includes the infrastructure's contraction, the network, the net-operators and the supporting system for the control centre and individual users. The project started in 2004 and is expected to be completed by 2010.

The aftermath of the rescue operation at the accident site is highly dependent on the intricate co-operation between different units in the task force and collaboration among different rescue members within the units. The information transfer and communication between the rescue members and to the control centre is crucial for the success of the rescue performance.

In Sweden, the information and communication system for firefighters are, in their own words, at the level of the "stone age". Besides voice communication via telephone/radio, there is no other possibility of getting necessary rescue information. The cooperation of the rescue members, and the coordination between different rescue units, such as different rescue teams, ambulance staffs and policemen, are based on limited radio communication and sometimes face-to face communication on the sites.

But times are changing and the requirements for the rescue work is changing, necessitating a new communication system that can support the changes. Following up the development of information technology, especially when the traditional analogical system is changing to the digital network system, it provides large opportunities for disaster communication

at very different levels. A new design of the communication system for the rescue operation is necessary.

This project is investigating how to design the multimodal interaction device for rescue members to meet the new requirements of the rescue work. The project is still in the earlier stages. We will report the result of the field study and the design proposal of the device.

## **2. Research methods**

This project takes the user-center-design (UCD) approach (Earthy, 2001; Maguire, 2001; Smith, 1997) for the design process. We will focus on the usability of the design to the end user - the firefighters. Therefore, we are involving experienced firefighters in our design process from the early stages of formulating the project until the final usability tests. The usability of the design is our main goal. We are taking a few steps for this project:

*Step one:* Before the initial design work, we set up a workshop with a few firefighters to go through the detail of how they perform their rescue work, how is their present information communication system works, what the existing problems are, and what the crucial requirements are for the design.

*Step two.* Based on the results from step one, we formulate the design structure of the new communication device. The design proposal would take the consideration of present technical feasibility, application context, usability requirements and different ergonomics factors.

*Step three:* We set up the second workshop with experienced firefighters again and discussed our design proposal. Cognitive walkthrough method was applied here. The detail of multimodal interface and the interaction with the new device is discussed.

*Step four and more...:* After the modified design proposal is verified, a testable prototype will be developed and this prototype will then be tested by the firefighters again. A few iterations will be carried out before the final prototype can be produced.

The total usability of the design is what we focus on. This means we will not only consider the usability of the interface, but the usability of the entire device. This will take consideration of their special working situation

and the environmental conditions of the application and the respective ergonomics issue of the design.

We are still in the earlier stages of this project. Here we will report the result from first workshop and the design proposal after the second workshop (from step one to three). The communication system design for the leader of the rescue work is our focus in this paper.

### **3. The communication situation at the present time**

When accidents occur, for example, if there is a fire in an apartment building, the initial report will come to the SOS centre. SOS will normally send the alarm to three different units, the fire station (rädningsverket), police station and ambulance service. From the fire station, normally, there are four fire engines, 208, 201, 202 and 203 with about 8 to 9 firefighters which is the rescue team that will be sent out first. Among them, the ISL (rescue leader, insattsledare) is the team leader.

There are a few radio channels that are used for different voice conversations at the accident site. Different members of the rescue team use different radio channels. ISL has three channels used simultaneously, together with one GSM mobile phone.

K 57: drive-front channel (framkörningskanal)

K 81: leading channel (ledningskanal)

K 80: rescue channel (rädningskanal)

Figure 1 shows the radio apparatus used by the firefighters. In Figure 1, the two radios on the left side (one red and one black with very simple interface) are the same radio for K80. The red one on the right side with a few more digit keys is the radio for K57 and K81. This radio is very powerful. It can communicate with different rescue units, such as police, ambulance, helicopters, etc.

K57 is the open radio channel used before the rescue work starts. It is normally attached inside the vehicle so its signal is very strong. As soon as the firefighters get out of the vehicle at the scene of the accident, they are required to switch on K80.



Figure 1. Different radio apparatus

ISL uses K57 to communicate with other firefighters to give very clear orders about how to carry out the rescue work. K81 is used only to communicate with the SOL (samband och ledning på brand stationen, control center in fire station), and with Bi (fire engineer, brandingenjör) when necessary. For ISL, K57 and K 81 is in one radio apparatus and he can press a button to switch between these two channels. K80 has to be in a separate radio apparatus. There are always two channels on all the time and ISL can hear everyone else talking who is on the same channels. If he needs to talk, he only has to push the button on the microphone. GSM mobile telephone is used only when personal communication is needed. Figure 2 shows how he is equipped.



Figure 2. The equipment for ISL.

If ISL needs to talk to the police (chief) or ambulance staffs, it is almost impossible with the present radio channels (K57 is available in ambulances, but not possible when the staff gets out of the ambulance), unless he knows the digital number of the police officer. They normally communicate face to face instead, when necessary. If the accident site is very big, it will take time to find each other.

Besides managing the rescue operation, ISL needs to make the documentation for the rescue processing. It is defined by law that he must document all the activities, for different purposes:

1. The security group in Commune needs to have detailed information for the purpose of improving the situation, for risk analysis.
2. Different insurance companies require details
3. For learning and training purposes

There are over 5 pages of rescue work report that the ISL has to fill in each time. A few items are very detailed, such as (in Swedish):

- 1) Ingrepp I annans rätt: Beslutsfattare, tidpunkt för beslut, vem och vad beslutet avser samt tidsomfattning, skäl till beslut, utförare av uppgiften.
- 2) Avslut av räddningsinsats: beslutsfattare, tid för avslut, Behov av bevakning, restvärde, kontaktperson ägare/nyttjanderättsinnehavare.
- 3) Läge vid framkomst.
- 4) Konsekvens av olyckan (människor/egendom/miljö): objekt, skada, hot, avsikt med insats, insats, prognos.
- 5) Preliminär bedömning av orsak till olyckan.

At present, if the ISL needs to document something in time so it will not be forgotten, he uses K81 to call SOL, so the person sitting at SOL will write down what he is saying. He will also take a digital camera to take some pictures for documentation purposes.

#### **4. Problems with existing communication system**

We can summarize a few problems in the existing communication system that we have found out in our study and our new design is intended to solve these problems:

1. There is no visual information relating to the disaster available on site for the firefighters.
2. Too many radio apparatus on hand.
3. High probability of using the wrong channel, or wrong apparatus for the specific communication.
4. No feedback about who is on the line.

5. When two radio channels are on at the same time, there is a high probability that important information may be missed.
6. It is not possible for the ISL to have direct communication with a specific rescue member in the team.
7. Everybody will hear all the communications. It has the risk of information overflow and may interrupt the rescue operation.
8. It is hard to know who is doing what and how their operation is progressing.

## **5. Design proposal**

The device that we are designing now will meet the following demands:

1. One device that can perform all the functions that ISL needs for his work.
2. It shall be easy and natural for ISL to interact with this device
3. It shall be highly useful in his working situation.

First of all, the device shall be light to carry. It shall be possible to fix it on the sleeve of his garment. The physical device shall look like a PDA and have the following functions: 1) different radio voice conversation; 2) availability to access the radio digital database; and 3) documentation. It can take both speech and manual input and graphic/text and TTS output. As ISL does not need to carry the heavy physical rescue equipment, it is possible for him to get the necessary information visually from the screen of PDA. By wearing a headset with a microphone close to his mouth it is possible for him to avoid the unnecessary negative effects from noise.

We discussed the details of the interface design with an experienced ISL during our second workshop by way of cognitive walkthrough. We principally agreed with the following multimodal interface design:

1) Speech commands for radio communication. We use speech command to switch the radio channels between different radio frequencies. We can also use speech commands to activate the radio or by pushing a button to talk, depending on the preference of the user. The ISL can easily “call” a specific person in the team by using special speech command, and give a special order (or have a discussion) with this person. This will avoid the information overflow in the old system. He can also give the order so the other person’s radio can be set to the right channel. The device will display

who is on the line, so he can get direct feedback from this person. We analysed the importance of different communications and noticed that certain channels (or certain persons in the team) have higher priority than others. So if this important channel has signals, then this channel can “break in” on other ongoing communications.

2) Accessing the database and navigating among different kinds of information. Speech commands are used as the main tool for this function, but a combination with manual input (such as pointers) should be better. Visual display of the information combined with some simple dialogue and TTS feedback are the main consideration. We did not discuss this part in detail, since its design will depend on how the database is designed.

3) Command for video documentation and speech dictation. The command for controlling the digital video camera can be very simple. As ISL needs to “write” the rescue performance report, he needs to give a speech command to activate the speech dictating system, so he can “write” his report by talking to the system.

## **6. Discussion**

One major factor for the success of the rescue operation is how well the commander, ISL, succeeds in synchronising the available rescue resources and coordinating the ongoing operations. He is the key person to master the dynamics of an emergency operation and the communication system design for ISL use is the main consideration in this project.

As soon as the rescue station gets the information from SOS about an accident (what happened and where it is), a team of fire engines with about 8 firefighters are immediately sent on their way to the scene of the accident. From that moment, all the information that the rescue work needs and all the communications are carried out via different radio channels and mobile phones. Some of the rescue members have to carry several radios for different communication purposes. It is very surprising to see that there is no information relating to the disaster visually available for the firefighters. There is no study showing how many errors are made by using different radio channels, how these errors, or the information overflow, affect their rescue operation.

We consider the new designed device shall be easy to carry, simply because ISL cannot just sit inside the car to control the rescue operation. With a portable device our main consideration is how can ISL easily and naturally interact with the device is. Multimodal user interface have better fulfilment. There is a study on multimodal user interface of speech process in MiPad's (Deng, 2002). It indicated the possibility and advantages of using such a system to combine phone functions with hand-computer functions. A careful selection of the commands is the key to the success of this project. We use different speech commands for different functions. The speech command controlled functions include setting different radio channels, switching on the radio to talk, navigating the database and even for documentation. The speech commands have to be very clear for the user and there is no overlapping of the commands and functions. Each specific speech command is designed only for one specific function. Manual performance of all the designed functions should always be available.

Error correction due to misrecognition of speech commands will be carefully considered in the design. Sturm (Sturm, 2005) studied form-filling and error correction. With a multimodal interaction design, such devices as re-speaking and an alternative list as error correction facilities, together with the soft-keyboard technique as additional error correction facilities turns out to be more effective error correction. In the present design, the misrecognition of speech commands, for most of the functions, will not pose a big problem. It seems to have a high error tolerance. For example, if the user called one channel, but the system recognized it as another channel, it is not critical and he can repeat the channel name again. As the proposed system has not yet been tested on site, we have very limited knowledge about the kind of errors that can occur and how they may affect the performance.

There is much work we need to do before a testable prototype can be designed. We need to carry the study on state-of-the-art of the possible technology, both hardware and software. Many research publications on multimodal interaction will help us with the details of the interface design. Still, a few iteration processing, usability tests on firefighters are important to reach the high usability of the design.

## Acknowledgement

I would take the opportunity to express my thanks to Mr. Leif Karlsson, for his understanding, support and cooperation in my research work. He represents the experienced “user” in this project and he has given me a lot of useful suggestions for the design. He is actually the handsome man in the picture (figure 2).

## References

- Deng, L., Wang, K., Acero, A., Hon, H-W., Droppo, J., Wang, Y-Y., Jacoby, D., Mahajan, M., Chelba, C., Huang, X.D. (2002). Distributed Speech Processing in MiPad's Multimodal User Interface. *IEEE trans. on Speech and Audio Processing*, 10(8), 605-619.
- Earthy, J., Jones, B.S., Bevan, N. (2001). The improvement of human-centred processes---facing the challenge and reaping the benefit of ISO 13407. *International Journal Human-Computer Studies*, 55, 553-585.
- Fujiwara, T., Watanabe, T. (2004). An ad hoc networking scheme in hybrid networks for emergency communications. *Ad Hoc Networks*.
- Harnett, B. M., Doarn, C.R., Zhao, X., Merrell, R.C. (2004). Redundant wireless communication technologies for real-time surveillance. *Telematics and Informatics*, 21, 375-386.
- Maguire, M. (2001). Methods to support human-centred design. *International Journal Human-Computer Studies*, 55, 587-634.
- Smith, A. (1997). *Human-Computer Factors: A study of Users and Information Systems*. The McGraw-Hill Companies.
- Sturm, J., Boves, L. (2005). Effective error recovery strategies for multimodal form-filling applications. *Speech Communication*, 45, 289-303.

## **Biography**

**Fang Chen** received her Ph.D. degree in 1997 from Linköping University of Technology in Sweden. She started to carry out different research works on human-computer interaction and speech and sound related interface design since then. In 2003, she received her Docent (Associate Professorship). She joined the interaction design group in Department of Computer Science and Engineering, Chalmers University of Technology in 2004.

## **Author's address**

*Fang Chen  
Interaction Design Group,  
Department of Computer Science and Engineering  
Chalmers University of Technology  
SE-412 96, Göteborg, Sweden  
phone: +64 31 7721076  
e-mail: [fanch@cs.chalmers.se](mailto:fanch@cs.chalmers.se)*



# GESTURE AND SPEECH MANIFESTATIONS OF PERSPECTIVE ON MEMORY OF EVENTS WITH VARYING DEGREE OF PARTICIPATION

*Pierre Gander*

SSKKII Center for interdisciplinary research on language, semantics, cognition, communication, information, interaction, Göteborg University, Sweden

## **Abstract**

*The present study investigates perspective on events as manifested in gesture and speech while people talk about past events. Perspective here refers to the mental spatial and psychological distance of the speaker in relation to some remembered event. Earlier studies suggest that the audience (i.e., the players) considers the agency in computer games as an extension of themselves and use the pronoun 'I' when talking about it. Eight participants experienced a computer game and four other conditions with varying degree of participation. Afterwards, interviews with participants talking about the events were recorded and transcribed. Iconic gesture was coded as observer or character viewpoint. Speech references to agents were coded (e.g., pronouns). Results show that the computer game gives a close perspective similar to personal experience, but also frequent usage of indefinite pronoun ('man'), which serves to express what should be done in the game. Iconic gesture does not seem to reveal perspective, but rather inherent spatial properties of actions.*

**Keywords:** Perspective, gesture, speech, memory, narrative, computer games

## 1. Introduction

The present paper reports on a study of how perspective is manifested in gesture and speech, when people remember and talk about events they have participated in<sup>1</sup>. More specifically, the question of concern is: What *perspective* do people adopt on actions and events from *participatory stories*, such as computer games? A participatory story is a physical system which allows the audience to choose between event sequences (that is, influence how the story should unfold) (Gander, 2005). The term *perspective* has a variety of meanings, but is here constrained to mean how a speaker mentally positions herself, both spatially and regarding psychological distance, in relation to some remembered event or action. Thus, perspective is here defined as a cognitive construct. This perspective can then be *manifested* in many ways, for instance, in gesture and speech.

At the heart of participatory stories is the concept of *agency*. The audience *carries out actions* when interacting with a participatory story. How is experiencing a story different when you carry out actions in it compared to when you do not? And how is it different to carry out actions in a participatory story compared to carrying out actions in the real world? When remembering events from one's own personal experience the perspective on the memory is that one participated in the event and one would use the pronoun *I*. In contrast, remembering events from a fictional text, such as a novel, would lead to a different perspective—that of an observer, looking at the event from the outside. Here, when talking about agency, one would use pronouns such as *he* or *she*, or nouns such as *woman*. Considering that participatory stories contain the elements of fiction as well as action, the question becomes what perspective people adopt on actions and events from participatory stories.

Other studies have shown that people often use pronoun *I* to talk about agency while playing computer games (Johansson, 2000; Linderoth, 2004). Wilhelmsson (2001) gives an explanation for how *I* can have multiple meanings in his conceptual framework for cognitive aspects of computer game activity, where he uses the work of Lakoff on the conceptual structure of the self. According to Lakoff (1996), the self is conceptualised as consisting of two parts: the Self, which is the part carrying out actions and associated with social roles and past actions, and the Subject, who is the locus of character experience, reasoning, and feeling. Wilhelmsson characterises the *Game Ego* as a Self that extends into the computer game.

---

<sup>1</sup> A more detailed account of this study is available in Gander (2005).

In the present study, perspective on memories of actions and events was investigated by using transcriptions of speech and gesture from interviews with eight participants as data, recorded after the participants had experienced events from five sources: *Game*: events from the textual adventure computer game *Anchorhead* which the participants played, *Story*: events from a Sherlock Holmes short story which the participants read, *Personal*: events that took place earlier in the participant's life, *Tasks*: events from three practical tasks carried out in a laboratory setting, and *Non-participatory*: events from a special, adapted non-participatory version of the computer game, printed on sheets of paper, which the participants read.

The perspectives used to talk about events from the participatory story (the Game condition) were then compared to the other four conditions. Two individual studies will now be presented which investigate how perspective is manifested, starting with gesture.

## 2. Perspective as Manifested in Gesture

McNeill (1992) studied retelling of narrative and showed that gesture can give important information about what speakers are thinking, for instance, concerning *perspective*. For McNeill, the perspective taken on a described event means the *distance* with which the speaker positions herself in relation to the event. Gesturally, perspective is manifested in iconic gestures, where two perspectives are possible: *character viewpoint* which shows an action from a character's viewpoint and signals closeness to the events described, and *observer viewpoint*, which expresses an outside viewpoint and signals distancing from the event described.

If the audience in participatory stories 'identifies' with the agent in the participatory story—if the player character is a 'Game Ego' by means of a 'projected Self', in Wilhelmsson's (2001) words—the audience would then experience, encode, and remember events from a first-person perspective. Thus, the prediction was that gesture would show a higher proportion of character than observer viewpoint in retellings of events from participatory stories, compared to the short story and the non-participatory version of the participatory story.

## 2.1 Method

The viewpoints of iconic gestures were transcribed and coded using a simplified version of a coding scheme from McNeill (1992) as *observer* or *character*.<sup>2</sup> For analysis of gesture, only cases of reference to past events were included, in other words, ‘the level of narrated events’ (McNeill, 1992).

## 2.2 Results and discussion<sup>3</sup>

The proportions of observer and character viewpoints of iconic gestures can be seen in Figure 1. The proportions were equal, except in the Tasks condition where character viewpoint was greater, lowering the proportion of observer viewpoint to a few percent.

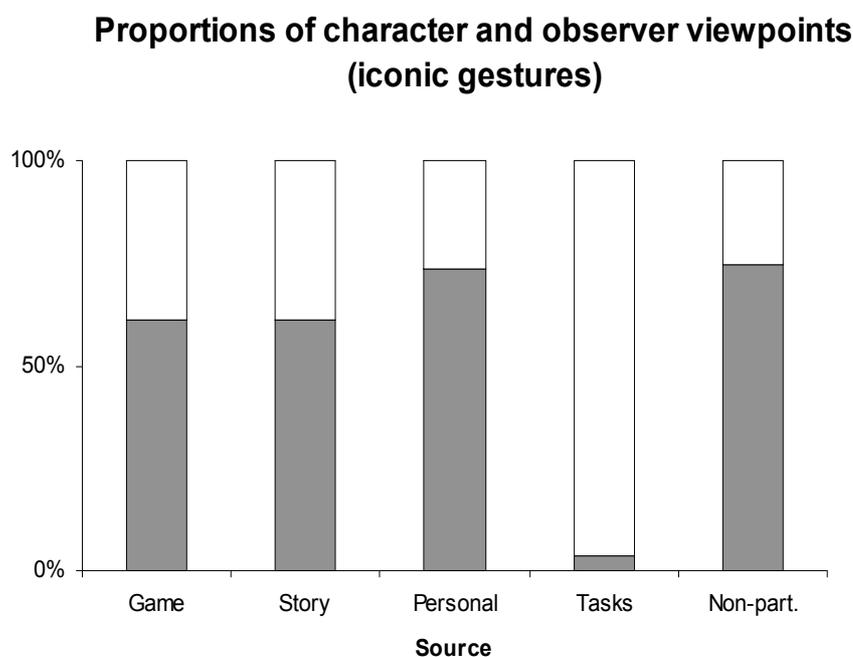


Figure 1. Proportions of character and observer viewpoints in iconic gestures.

A factor that may explain why there was a difference between the Personal and Tasks conditions—two conditions which would be expected to show similar results because both are personally experienced, real events—is that different amounts of time passed since the events took place. In research on

<sup>2</sup> Measures of inter-rater reliability of the codings are left out here for brevity, but can be found in Gander (2005).

<sup>3</sup> Statistical analyses are left out here for brevity, but are available in Gander (2005).

visual perspective in episodic memory, memory age has been shown to produce more memories with an outside perspective (Nigro & Neisser, 1983; McIsaac & Eich, 2002). However, it does not seem to fully account for the results, because the large difference found in the present study—character viewpoint outnumbering observer viewpoint by a factor of 30—has not been found in those studies.

How can the extreme result in the Tasks condition be explained? Maybe the observed proportions of observer and character viewpoints do not correlate with participation, but result from the nature of what needs to be communicated. The description of certain actions may demand an observer viewpoint while other actions may demand a character viewpoint. This may have nothing to do with who carried out the action: the participant or someone else. Particularly, taking an observer viewpoint may be necessary in order to gesturally show actions which involve space that lies beyond the reach of the speaker's arms. That way, a real space is compressed into a smaller space, which serves as a model to illustrate the actions which took place. Actions may be described using character viewpoint gestures as far as possible, and only if events involve space larger than that around the body, are gestures with observer viewpoint used. Reviewing the results from the five conditions shows that they are consistent with this explanation. The laboratory task condition encompassed events which could be well described gesturally using a character viewpoint, since they involved only a limited space around the body. The other conditions all in addition involved events that took place in a larger space, which may have forced the participants to use an observer viewpoint as well. This explanation requires us to see the computer game events not as something occurring in front of a computer on a desk, but as events in the game world.

Next, an analysis of perspective as manifested in speech is presented which will give more insight into what perspective participants adopt when talking about actions and events from the participatory story.

### **3. Perspective as Manifested in Gesture**

When talking about a past event, a speaker has a lexical choice of how to refer to the animate agents involved in the event. This choice of terms for the referents is one type of choice of perspective, and will here be called *agent perspective*. Agent perspective reveals the speaker's view of the events and gives clues to ongoing cognitive activity. Let us exemplify this by looking at the agent perspectives possible in Swedish when one is talk-

ing about past events. The first-person singular pronoun *jag* ('I') may be used, for instance *jag tog en promenad i morse* ('I went for a walk this morning'). This choice of pronoun indicates a closeness to and signals participation in the event talked about on the part of the speaker.

The second-person singular pronoun *du* ('you') signals lack of participation and distance on the part of the speaker, as in *du tog en promenad i morse* ('you went for a walk this morning').

The use of a third-person singular pronoun signals distance of the speaker from the event and does not show that the speaker participated in the event—for example, *han tog en promenad i morse* ('he went for a walk this morning'). The same distant perspective is expressed by using nouns, such as *mannen* ('the man') or *Michael*.

Another way of referring to the agent of some event is by using the indefinite pronoun *man*, as in *man kunde inte tro sina ögon* ('one couldn't believe one's eyes'). There is no single English equivalent to this Swedish pronoun (Andersson, 1972; Norell, 1995)—*man* can be translated into 'one', 'you', or 'they', or expressed using other constructions such as passive, depending on the meaning. Regardless of usage, *man* can be said to distance the speaker from what is being said.

Describing an event using the pronoun *vi* ('we')—which can be viewed as the plural version of *jag* ('I')—signals closeness to the events. It does however express a somewhat weaker participation in the events on the part of the speaker compared to *jag*.

The pronoun *dom* ('they') (including the forms *de* and *dem*) is the plural form to refer to the third person. It signals neither closeness to nor participation in the events talked about.

Which agent perspectives occur when participants talk about events from the five conditions? More specifically, how do people refer to the player character in the computer game—as 'I' or 'she' or in some other way? Based on Wilhelmsson's (2001) analysis of Game Ego, it was predicted that the participants referred to the player character in a way similar to personal experience, that is, using the pronoun *jag* ('I').

### **3.1 Method**

Parts of the method were similar to the method used to analyse gesture and a presentation can be found in that section above.

When coding agent perspective, the participants' utterances were searched for terms fulfilling three criteria; the term should i) *refer to an animate agent* (the speaker, a character in a story or a game, etc.), ii) *refer to someone who performs an action*, and iii) *occur on the level of narrated events* (in the case of the computer game: references to the game world and its objects, etc., in the case of the short story: references to the story world, and so on).

### 3.2 Results and Discussion

Comparing overall patterns of agent perspective, as shown in Figure 2, the results form two main groups: The Game, Personal and Tasks conditions as one group and the Story and Non-participatory conditions as a second group. Interestingly, talk about events from the participatory story and the non-participatory version of the same (these two had identical textual descriptions) showed completely different patterns of agent perspective, the former dominated by *jag* ('I') and *man* ('one'/'you'), while the latter dominated by the agent perspective group consisting of singular and plural third-person pronouns, and nouns.

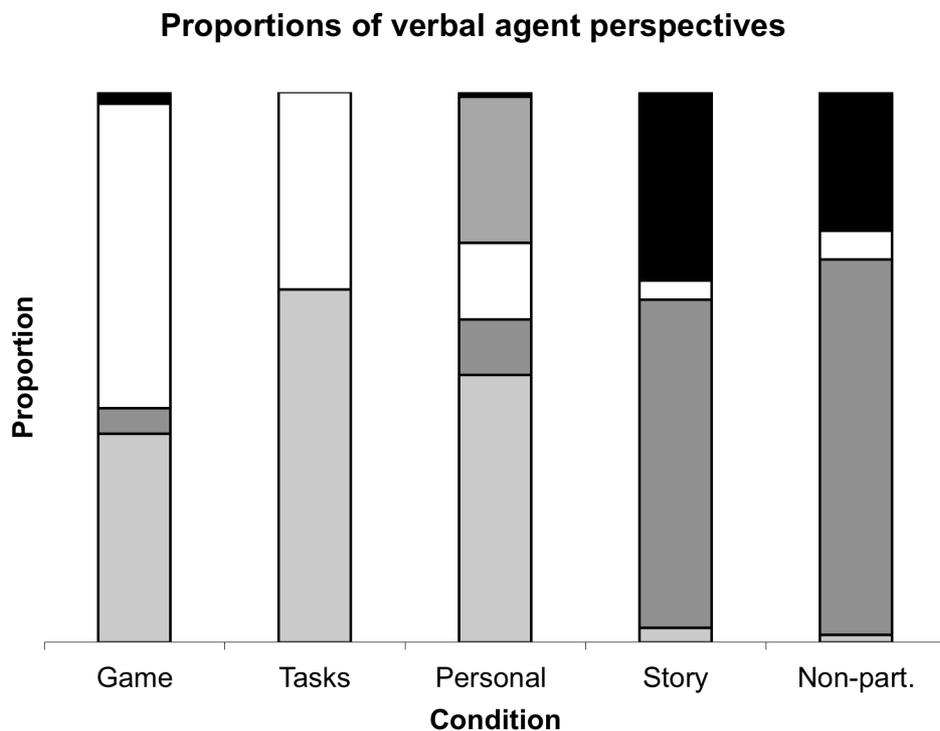


Figure 2. Proportions of agent perspective in talk about events from the five sources (average for all participants).

The interpretation of these results is that participants mainly took an outside perspective when talking about the events from the short story and the non-participatory version. In the other three conditions, the perspective was mainly an inside perspective, but the common use of the indefinite pronoun *man* ('one'/'you'), especially in the Game condition, added distance to the perspective.

The case of most theoretical concern in the results is when participants talk about events from the computer game. In contrast to the other conditions, where the perspective is more or less given, the events from the computer game present an intriguing case for studying how actions and events in the computer game are thought about by the participants.

Perhaps the most unexpected result concerning the use of perspective in the Game condition is the frequent occurrence of the indefinite pronoun *man* ('one'/'you') (55.4 percent, compared to 5.2 percent in the Non-participatory condition). What does *man* mean? In one sense, the meaning of *man* seems to be rather close to *jag* ('I') in that these two terms can sometimes be exchanged without changing much of the meaning. Sometimes *man* was used in order to express something one should do, considering the intention of the designer of the computer game. However, the event described need not have occurred. Thus, using *man* was a way of referring to actions which should be performed, even if they were not actually performed by the participant when playing the game. Used in this way, *man* refers to a group of people of which the speaker is not a part (cf. Andersson, 1972). There was sometimes referential ambiguity in the use of *man*, so that is unclear whether something happened to the person or the game character.

The two senses of the first-person singular pronoun *jag* ('I') noted by Johansson (2000)—the *participant* and the *player character*—were found also in the present study. The referent of *jag*—like the indefinite pronoun *man* ('one'/'you')—was ambiguous and referred sometimes to the participant as an individual and sometimes to the player character. There seem, however, to be ways of separating the two meanings of *jag*. In the use of *jag* to mean the individual, references are to (i) mental states such as thoughts, (ii) the physical surrounding in the playing situation, and (iii) managing and progressing in the game as an abstract object. In the use of the other meaning of *jag*, the player character, there are references to the game world and its objects, places, and characters. These characteristics can work as guidelines, but they do not completely eliminate the referential ambiguity of *jag*.

No participant referred to the player character with the third-person pronoun, e.g., *hon* ('she') or a noun, such as 'the woman'.

The results support Wilhelmsson's application of Lakoff's Self and Subject. Participants talked as if they were themselves present in the game world. Thus, participants' Subjects seem to have been projected into the player character's Self.

#### 4. Conclusion

The results of two analyses regarding which perspectives were adopted by participants on past actions and events revealed differences between the five conditions: a computer game, a special non-participatory version of the computer game, a short story, personal experience, and practical laboratory tasks. Considering first the analysis of perspective as expressed in gesture, the results revealed that in talking about the practical laboratory tasks, participants adopted mainly a close, inside perspective (an iconic gesture character viewpoint). In the other four conditions, both 'outside' and 'inside' perspectives were present (proportions of character viewpoint and observer viewpoint were about equal).

The analysis of how perspective was revealed in speech resulted in two groups. Talk about the computer game, personal experience, and practical laboratory tasks featured mostly a close perspective, while talk about the short story and the non-participatory version of the computer game featured mostly a distant perspective.

Do the results of the two studies of expressed perspective provide a unified picture? It is only for expression of perspective on actions and events from the computer game that gesture and speech reveal the same picture. In the Game condition, both character/close and observer/distant perspectives were adopted, and these occurred in similar proportions for both gesture and speech. For all the other four conditions, there were discrepancies in that a difference was found between character and observer perspectives for gesture, but no difference for perspective expressed in speech, or vice versa. One suggestion is that inherent spatial properties of actions rather than spatial or psychological closeness to the events may influence the choice of gesture viewpoint. Actions which need a large space in order to be depicted gesturally may lead to an observer viewpoint.

The analysis of perspective as manifested in speech give the most detailed results. These results show that the participatory story and its non-participatory

patory version elicited different perspectives. Participants adopted a perspective on events and actions from a participatory story similar to events and actions from personal experience. Thus, it seems that participation is the reason for adopting a personal perspective in the participatory story.

## References

- Andersson, L-G. (1972). *Man: ett pronomen*. Gothenburg Papers in Theoretical Linguistics 15. Göteborg: Department of Linguistics, Göteborg University, Sweden.
- Gander, P. (2005). *Participating in a story: Exploring audience cognition*. PhD dissertation, Department of Cognitive Science, Lund University, Sweden.
- Johansson, B. (2000). 'Kom och ät!' 'Jag ska bara dö först.' *Datorn i barns vardag*. Göteborg, Sweden: Etnologiska Föreningen i Västsverige.
- Lakoff, G. (1996). Sorry, I am not myself today: The metaphor system for conceptualizing the self. In G. Fauconnier & E. Sweetser (Eds.), *Spaces, worlds, and grammar* (pp. 91–123). Chicago and London: The University of Chicago Press.
- Linderoth, J. (2004). *Datorspelandets mening: Bortom idén om den interaktiva illusionen*. PhD dissertation, Department of Educational Sciences, Göteborg University, Sweden.
- McIsaac, H. K., & Eich, E. (2002). Vantage point in episodic memory. *Psychonomic Bulletin & Review*, 9, 146–150.
- McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: Chicago University Press.
- Nigro, G., & Neisser, U. (1983). Point of view in personal memories. *Cognitive Psychology*, 15, 467–482.
- Norell, P. (1995). 'Answers to that question were now being sought'; English translations of the Swedish indefinite pronoun man in fiction and non-fiction texts. In G. Melchers, & B. Warren (Eds.), *Studies in anglistics* (pp. 191–200). Stockholm: Almqvist & Wiksell International.
- Wilhelmsson, U. (2001). *Enacting the point of being: Computer games, interaction and film theory: affordances and constraints, metaphorical concepts and experientialist cognition observed through the environment in computer games*. PhD dissertation, Department of Film and Media Studies, Copenhagen University.

## **Biography**

**Pierre Gander** took his MSc degree at the University of Skövde, Sweden in 1996 and completed his PhD in cognitive science at Lund University, Sweden in 2005. The PhD dissertation, entitled ‘Participating in a story: Exploring audience cognition’ empirically investigated some cognitive aspects of the users of computer games. Having been mostly interested in the connection between language and cognition, he is now pursuing research on the mental representation of fictional information.

### **Author’s address:**

*Pierre Gander  
Göteborg University  
SSKKII  
Box 200  
S-405 30 Göteborg, Sweden  
phone: +46 31 773 2985  
e-mail: pierre.gander@ling.gu.se*



# THE MODAL DIFFERENCES IN CHILDREN'S COMMUNICATION

*Mia Heikkilä*

Uppsala University, Sweden

## **Abstract**

*This paper on multimodal communication is concerned with questions on how children use modes while communicating. Focus is on how the modes speech, gaze and gesture are realized and what functions the modes tend to have in two different centrally occurring activities at pre-school, pre-school class and at school<sup>1</sup>. The activities analyzed are 'play' and (teacher initiated) 'work', in total 15 children are participating in the analyzed activities. No teachers are included in this study. A research question of this PhD project is concerned with what functions the modes tend to have in relation to what is focused on in the different activities (see Kress, 2003). The fieldwork was carried out during three years, 2000 – 2003 and some of the children were followed during the complete period of the research. Ethno-graphic video registrations of activities in these three school settings are the basis for the detailed transcriptions and analysis of 19 sequences of play and work.*

*The first findings show that children use modes quite differently in these activities. In the 'play' activities the body movements and use of gesture are much greater than in 'work' activities where gaze and speech were the most foregrounded modes in the communication.*

---

<sup>1</sup> In Swedish those are called förskolan, förskoleklassen and Årskurs ett. In Anglo-Saxon countries the translation used doesn't necessarily make much sense, but since this is a Nordic conference I choose not to go further into explaining the differences between the settings.

**Keywords:** multimodality, communication, children, education, individuality, group-participation

## 1. Introduction<sup>2</sup>

An illustration and understanding of children's communication can be presented in quite different ways depending on focus, interest and area of research. The main interest and curiosity that lies behind this study is how we can understand and verbalize, discuss and enhance meaning making in schools. Since my disciplinary area is that of education meaning making and learning can be considered connected interest areas. I look at theories of multimodality and thereby analysis of communication as a fruitful way to find means to understand meaning making and children's representations of life at school.

To capture modal differences in children's communication an introduction of the study conducted is needed. In this text I will present the focus of study, how an ethnographic fieldwork was carried out and how the process of analysis and interpretation of the empirical material was done. After doing so, a shorter presentation into the field of social semiotic multimodality is made with emphasis on different areas within multimodal theory that are of special interest in this study. The study itself and some results will be presented at the end of the text. Within the confines of this paper I will not be able to include all parts of the results and I have therefore chosen to pick out some main results, which can be seen as more general ones.

I deliberately want to seek ways of presenting studies of human communication and interaction, which in the long run can facilitate new insights into the complexity of understanding communication. I therefore emphasize the artefacts used in the analyzed activities, which I present. My point in doing so is mainly that of exploring what happens to the perception of communication when attention is directed towards artefacts and the assumption that artefacts in fact can be seen as a mode of communication and therefore crucial to understand in order to understand communication itself.

---

<sup>2</sup> This paper is a first summary of a PhD project that has run over five years (2000-2005) and will be finished by the end of 2005.

## 2. The study

The study was/is conducted during 2000 – 2005. The study is part of a larger project carried out at the Department of Education in Uppsala and the PhD project can be considered an independent part of that. The larger project is called Pre-school and school in co-operation (PISCO, <http://www.ped.uu.se/fisk>) and was financed during 1999 through 2004, by the Swedish National Agency for Education. PISCO's main focus was to study the interaction between pre-schoolers, pre-school classers and first graders in order to analyse social inclusion and exclusion. Although partly different theoretical standpoints are taken in the two parallel projects the common interest is that of interaction and communication.

The main focus in this project is to study children's communication and meaning making through their use of different modes of communication - speech, body position, gaze and use of artefacts - and to analyze how these operate in different activities. The aim is to understand and find regularities in how children's communication occurs in school-life activities. I am studying speech, body position, gaze and use of artefacts in nineteen different activities from pre-school, pre-school class and first grade.

The results of the study include knowledge in two different, somewhat overlapping, areas. The areas, as I consider them, are those of *childhood* and *education*. I see multimodality as a way of going there. Firstly, the knowledge that this thesis generates gives more insight into what children are occupied with during school days and in addition to that what it is like to be a child at school today. Secondly, the focus on education can be summarized as ways of seeing education not only as learning school specific knowledge but also as seeing communication as a great issue of education more generally. The study also shows a somewhat different way of problematizing meaning making in educational settings where not much focus has been put on understanding meaning making through the lens of multimodality. The focus of studying modes as signs of individuality and group-participation is one way of understanding how meaning making is realized.

The analysis shows that the use of modes in the activities studied are at an analytical level about positioning yourself as an individual and at the same time as showing your belonging to a group of people in your class. These two positions are realized through different modal positions and different ways of relating to the artefacts and the physical room you are in. Through

directing modal attention to task-related artefacts the children switch focus from either individuality to group-participation or vice versa.

I have chosen to study two mainly occurring activities<sup>3</sup>, which takes place in the three different classroom settings where the study was carried out. I chose to call the activities 'play' and 'work' to differ between how the participants use the activities -a somewhat provocative differentiation.

Video recordings were done during the years 2000 – 2003 in one pre-school, one pre-school class and one first grade in a Swedish urban area. The video recordings were carried out in two periods each year. The video recording procedure is one way of collecting empirical material, but the video recordings are recordings of different kinds and therefore to be considered as different kinds of materials answering different kinds of possible research questions (see Pink, 2002; Heikkilä & Sahlström, 2003; Pink et al., 2004). The differences comprise of different zooming, different quality of sound and picture, different number of children in the picture and different activities recorded. These kinds of aspects have to be taken into consideration when using video recordings in a research study focusing on studying communication.

The technical aspects of video recordings mentioned above are not only technical, but also highly connected to the theoretical and ethical aspects of an entire research project (see Lindsay, 2000). Strict demands were therefore placed on the technical quality of the sequences, i.e. the different activities, which were to be chosen for an analysis. The demands included; being able to hear everything that was said; to see and interpret the direction of the gaze; and to be able to see the body positions of the children participating in the activity.

The ethnographic fieldwork yielded 100 hours of video recordings. These video recordings were made into a catalogue with the help of File Maker Pro software. Through repeated viewing of the sequences and by using the File Maker catalogue key elements, the later analysis became clear. 19 sequences became the empirical material used for analysis and the lengths of the sequences vary from 2 minutes to 6 minutes.

---

<sup>3</sup> Since the study was carried out 'ethnographically' I have empirical material covering the entire day the children were in pre-school, pre-school class and first grade and therefore I am able to support my analyses with this information.

All sequences are transcribed with a second to second basis. The transcripts are focusing on transcribing the occurrence of speech, body position and direction of gaze. The use of artefacts is included in the overall interpretation of the material, since they are seen in the pictures included in the transcriptions.

### **3. Social semiotics and multimodality**

The theoretical frame for the thesis is that of social semiotic multimodality (Kress & van Leeuwen, 2001; Kress, 2003; Jewitt & Kress, 2003; Hodge & Kress, 1988). Multimodality can be seen as one of many developments from a social semiotic discourse analysis. Multimodality can be described as more considered with aspects of representations of communication and as a theorization of how understanding communication is a way of understanding the realization of society and human life. In line with the social semiotic interest in studying social signs and the production and distribution of social signs, communication can be seen as a set of social signs produced by humans, i.e. modes.

Modes are, seen in a social semiotic multimodal way, realizations of ideologies, discourses and semiotic resources, which in communication carry different functions (Kress, 2003; Hodge & Kress, 1979/1993). The functions of modes vary from time to time and space to space. Modes are a way of understanding the realization of social semiotics, which emphasises what is done in time and in space. In this study focus is put on modes produced by humans, although artefacts and the physical spaces also are taken into the analysis. In the analysis, time is considered as a basic variable in the transcript and space as a basic variable when analysing artefacts and the physical spaces, in line with contemporary thinking in multimodality (Kress, 2003).

Functional specializations of the modes in the activities are of special interest in this study. The analysis shows that the children use the same modes differently depending on their position as an individual or as a group-participant in an activity. The function of one mode is changed from moment to moment, where as the same mode can be used to position individuality in one activity it is a clear way of emphasising group-participation in another.

Since one basic assumption in multimodal theory is that you can never choose not to communicate multimodally, all our time of being humans in

society is built on communication of different kinds. We have a limited set of modal variations to use and in the multiplicity of situations that we participate in we need to be able to use the same modes differently related to the task at hand. A considerable number of studies have emphasised speech and its variations as a central mode of communication, which most likely is true in many cases, but is that always the case? When working with multimodality the main hypothesis is that we don't know which mode is the most foregrounded and what functions it carry – most likely we have an idea. The question is most often how modes are related to each other (see for instance Jewitt, 2003; Mavers, 2003; Stein, 2003) in order to capture modes used for meaning making. The assumption is that by knowing how modes relate to each other and by knowing what modes carry what function we also know more about how society and human life is realized and categorized and how meaning making in this life is realized over and over again.

I here choose to consider education with the interest in meaning making, essentially as an empirical field within which different kinds of *changed realization* exist. In its widest sense changed realization can include different aspects of changing your insight into specific school knowledge, changing your way of looking at things around you, changing your friends or other aspects of your social life. In order to accomplish that change, you have to communicate. Activities like reading, talking, singing, playing, dancing and working include a perception based on, what you might call realizations of an understanding of both the situation you are in and the relations you have to the people around you. Your personal interest in that situation might for different reasons change, implying that your realization of the situation changes. This you do by changing your modes of communication – by directing your body towards different persons, by choosing what work to do and how properly to do it, by not singing along in songs, or by showing your painting, as the children do in the material, in a specific way.

#### **4. Concluding comments**

The title of this paper – ‘Modal differences in children’s communication’ – refers to the research focus of studying different modes children use when communicating. The results of the study show that the modes are used differently, i.e. the modes have different functions, in different activities. The activities and the modes are analyzed multimodally and afterwards regularities of the multimodal orchestration were searched for. Looking at

the analysis with the notions of individuality and group-participation gave new insights into the material. As briefly mentioned at the beginning of this paper these two aspects can be seen as central reference points for children within the school setting.

Together with a social semiotic multimodal analysis focusing on meaning making and realization of communicational modes the results of the study, have been presented in a very general way.

The framing of the activity, together with the possible use of artefacts, and the content of task changes the way modes are used. This is shown when focus is put on how the children modally refer to individuality and group-participation.

### **Acknowledgements**

The author would like to thank Sandra McAuley and Gunther Kress for commenting on both the content and the use of English in this paper.

### **References**

- Heikkilä, M., & Sahlström, F. (2003). Om användning av videoinspelning i fältarbete. *Pedagogisk Forskning i Sverige*. Vol 8(1-2): 24-41. Göteborg.
- Hodge, R., & Kress, G. (1979/1993). *Language as Ideology*. 2<sup>nd</sup> Edition. London/New York: Routledge.
- Hodge, R., & Kress, G. (1988). *Social Semiotics*. Cambridge/Oxford: Polity Press.
- Jewitt, C. (2003a). *A Multimodal Framework for Computer Mediated Learning: the Reshaping of Curriculum Knowledge and Learning*. Unpublished PhD thesis. Institute of Education, University of London.
- Jewitt, C. (2003b). Computer-Mediated Learning: The Multimodal Construction of Mathematical Entities on Screen. S. 34-55. In Jewitt, C. & Kress, G. (Eds.). *Multimodal Literacy*. New York: Peter Lang.
- Jewitt, C., & Kress, G. (2003) (Eds.). *Multimodal Literacy*. New York: Peter Lang.
- Kress, G. (2003). *Literacy in the New Media Age*. London/New York: Routledge.

- Kress, G., & van Leeuwen, T. (2001). *Multimodal Discourse. The Modes and Media of Contemporary Communication*. London: Arnold.
- Lindsay, G. (2000). Researching children's perspectives: ethical issues. pp.3-20. In Lewis, A. & Lindsay, G. (Eds.). *Researching children's perspectives*. Buckingham/Philadelphia: Open University Press.
- Mavers, D. (2003). Communication Meanings through Image Composition, Spatial Arrangement and Links in Primary School Student Mind Maps. Pp. 19-33. In Kress, G., & Jewitt, C. (Eds.). *Multimodal Literacy*. New York: Peter Lang.
- Pink, S. (2001). *Doing Visual Ethnography*. London/Thousand Oaks/New Delhi: SAGE Publications.
- Pink, S. , Kürti, L., & Afonso, A. I. (2004)(Eds.). *Working Images. Visual Research and Representation in Ethnography*. London/New York: Routledge.
- Stein, P. (2003). The Olifantsvlei Fesh Stories Project: Multimodality, Creativity and Fixing in the Semiotic Chain. pp. 123 – 138. In: Jewitt, C. & Kress, G.(Eds.).*Multimodal Literacy*. New York: Peter Lang.

## **Biography**

**Mia Heikkilä** is working at the Department of Education at Uppsala University as a researcher and teacher finishing her PhD by 2005. She has a background as EdM from Åbo Akademi University in Vasa, Finland, and a Master of Music Education from the Royal College of Music in Stockholm, Sweden. Her special interests are multi-modality, learning and meaning making, and she has been a visiting researcher at the Institute of Education, University of London several times during the last years. In her research she combines multimodality, theories of childhood and education.

### **Author's Address:**

*Mia Meikkilä  
Uppsala University,  
Department of Education,  
P O Box 2109,  
750 02 Uppsala,  
Sweden.  
phone: + 46-18-471 1659.  
e-mail: mia.heikkila@ped.uu.se*



# ON THE INTERACTION OF AUDIO AND VISUAL CUES TO FRIENDLINESS IN INTERROGATIVE PROSODY

*David House*

Department of Speech, Music and Hearing  
KTH, Stockholm, Sweden

## **Abstract**

*This paper investigates interactions between audio cues and visual cues to friendliness in questions in a series of perception experiments. Generally, the results indicate that high F0 peaks are perceived as friendlier than low peaks, late final peaks are perceived as friendlier than early final peaks, and final focus is in certain cases perceived as friendlier than medial focus provided the final focal peak is delayed. Friendliness in questions can also be effectively signaled by visual facial parameters such as a smile, head nod, and eyebrow raising. Data-driven visual synthesis can be used to convey friendliness in questions where the database is recorded by an actor instructed to portray happiness. Interactions between audio and visual cues show a strong influence of the visual cues and a consistent influence of audio cues over different visual stimuli.*

**Keywords:** Audiovisual prosody, question intonation, expressive prosody, talking heads

## **1. Introduction**

The signaling of interrogative mode in speech through intonation is a topic of standing interest in intonation research. Not only does question intonation vary in different languages but also different types of questions (e.g. wh, yes/no or echo questions) can result in different kinds of question

intonation (Ladd, 1996). In very general terms, the most commonly described tonal characteristic for questions is high final pitch and overall higher pitch (Hirst and Di Cristo, 1998). In many languages, yes/no questions are reported to have a final rise, while wh-questions typically are associated with a final low. Wh-questions can, however, often be associated with a large number of various contours (Cruttenden, 1986). This paper explores the role of audio and visual cues to interrogative mode and, more specifically, how the interaction of audio and visual cues can signal an attitude of friendliness in interrogative mode.

## 2. Question intonation in Swedish

In Swedish, interrogative mode is most often signaled by word order with the finite verb preceding the subject (yes/no questions) or by lexical means (e.g. wh-questions). Question intonation can also be used to convey interrogative mode when the question has declarative word order. This type of echo question is relatively common in Swedish especially in casual questions (Gårding, 1998). Question intonation of this type has been studied in scripted elicited questions and has been primarily described as marked by a raised topline and a widened F0 range on the focal accent (Gårding, 1979). In two recent perception studies, House (2002, 2003) demonstrated that a raised fundamental frequency (F0) combined with a rightwards focal peak displacement is an effective means of signaling question intonation in Swedish echo questions (declarative word order) when the focal accent is in final position.

In figure 1, stylised F0 contours of the manipulated portions of the stimuli used in one of the studies are presented schematically (House, 2002). The stimuli numbers 1-6 correspond to the timing location of the peaks in both the low-pitch and high-pitch set. The utterance was “*Hon vill bara flyga*” (She only wants to fly), in which the final word “*flyga*” was manipulated.

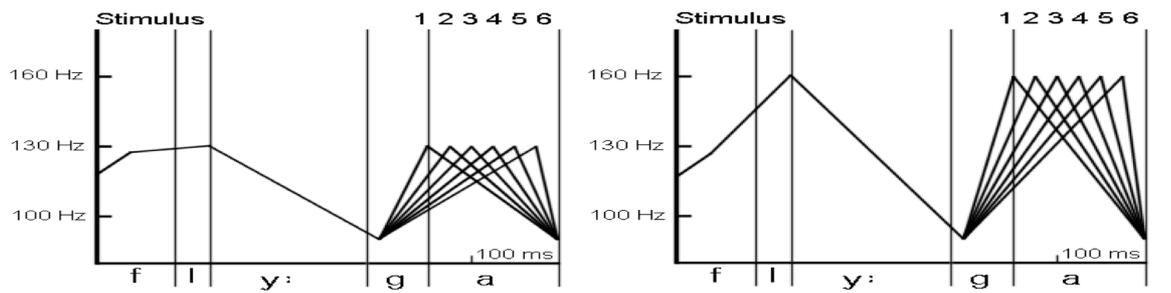


Figure 1. Schematic stimuli used in the perception test. The stimuli are numbered 1-6 in both the low-pitch set (left panel) and the high-pitch set (right panel).

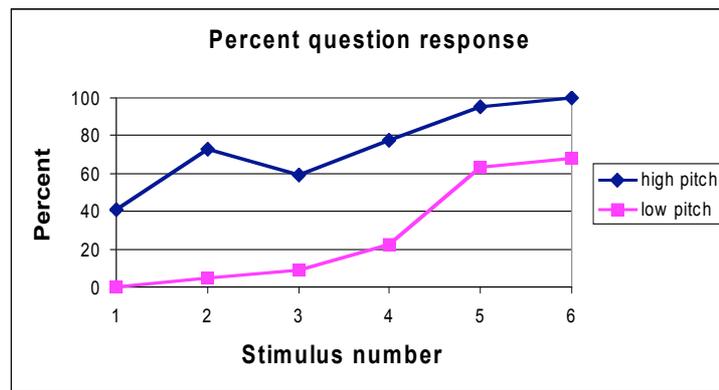


Figure 2. Results of the perception test showing percent question responses.

In figure 2, perception results are presented which confirmed the importance of timing where an early, low peak followed by a final falling contour was perceived as a statement while a late, high peak, resulting in a rise through the final syllable, was perceived as a question.

### 3. Visual cues to interrogative mode

In the second part of the study (House 2002), the same audio stimuli were presented with two different visual cue movement configurations involving the following facial gestures: smile, vertical head nod, eye narrowing and eyebrow lowering. The two configurations were hypothesized to convey interrogative or declarative mode. Two configurations combining different facial gestures were synthesized using an experimental version of the Infovox 330 diphone Swedish male MBROLA voice implemented as a plug-in to the WaveSurfer speech tool which also includes a visual parametric manipulation tool described in Sjölander and Beskow (2000). The parameter settings were inspired by those used in an earlier experiment

on positive and negative feedback cues (Granström et al. 2002). The parameters for the hypothesized interrogative mode consisted of a slow up-down head nod and eyebrow lowering. The parameters for the hypothesized declarative mode consisted of a smile throughout the whole utterance, a short up-down head nod and eye narrowing. Samples of the configurations are shown in Figure 3.



*Figure 3. The hypothesized interrogative configuration (left) and declarative configuration (right) sampled at the onset of the final vowel.*

In figure 4, the results of the audiovisual perception experiment are shown. They are generally very similar to the results of the audio stimuli alone (see figure 2) and thus indicate that the influence of the visual cues on the auditory cues was marginal. While the hypothesized cues for declarative mode (smile, short head nod and eye narrowing) reinforced declarative intonation, the hypothesized cues for interrogative mode (slow head nod and eyebrow lowering) led to more ambiguity in the responses. Similar results were obtained for English by Srinivasan and Massaro (2003). Although they were able to demonstrate that the visual cues of eyebrow raising and head tilting synthesized based on a natural model reliably conveyed question intonation, their experiments showed a weak visual effect relative to a strong auditory effect of intonation. This weak visual effect remained despite attempts to enhance the visual cues and make the auditory information more ambiguous.

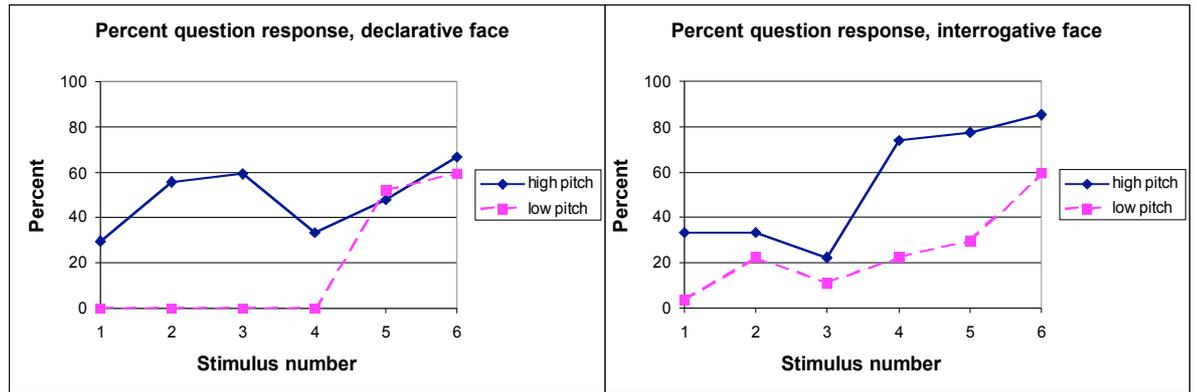


Figure 4. Results of the audiovisual test showing percent question responses for the declarative face (left panel) and for the interrogative face (right panel).

#### 4. Phrase-final rises, social intention and friendliness

As illustrated by the perception experiment presented above, phrase-final tonal characteristics, especially a final rise, can be of importance for signaling interrogative mode. The prosodic features of the ends of phrases can also reveal characteristics about questions which are important for the dialogue flow. For example, there has been recent interest in the automatic analysis of phrase final tones and short utterances with the objective of categorizing and extracting dialogue acts such as agreement, acknowledgement, backchannels, turntaking and speaker attitude (see e.g. Caspers, 2003; Cerrato, 2002; Ferrer et al., 2002).

In a recent study of phrase-final features in a set of 200 wh-questions extracted from a large corpus of computer-directed spontaneous speech in Swedish, it was found that final rises occurred in 22 percent of the utterances (House, 2005). Moreover, there was an indication that the questions ending in a final rise were more oriented to signaling a social interest while those with a final low were more oriented to a request for specific information. These results are consistent with a study on German spontaneous speech in which Kohler (2004) proposes that “rising pitch expresses friendliness, interest and openness towards the addressee, while falling pitch focuses on routine, lack of interest and categoricalness” (p. 207).

A series of audio-only perception tests was carried out in which subjects were asked to indicate which of a pair of questions was the friendlier one. As in the earlier experiments on question intonation, the manipulated

intonation parameters were F0 peak location and F0 peak height. In these experiments, a late, high peak was generally perceived as expressing more friendliness than an earlier, low peak. Figure 5 illustrates these trends where peak position 1 is early in the second syllable of the question “*Vad heter du?*” (What is your name?), position 4 is late in the final syllable and positions 2 and 3 are in between. A complete account of the experiments is presented in House (2005).

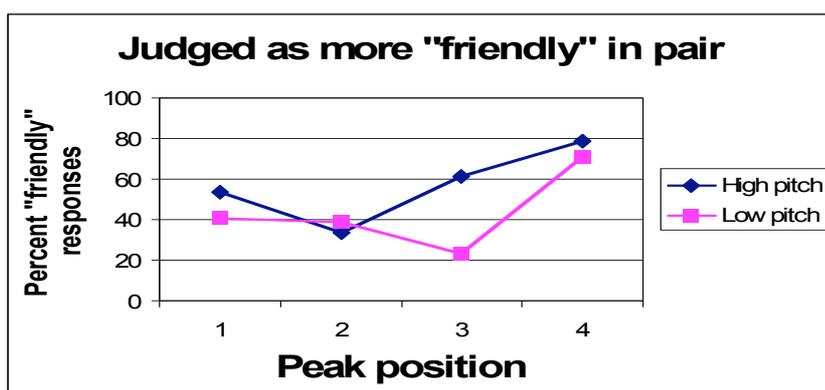


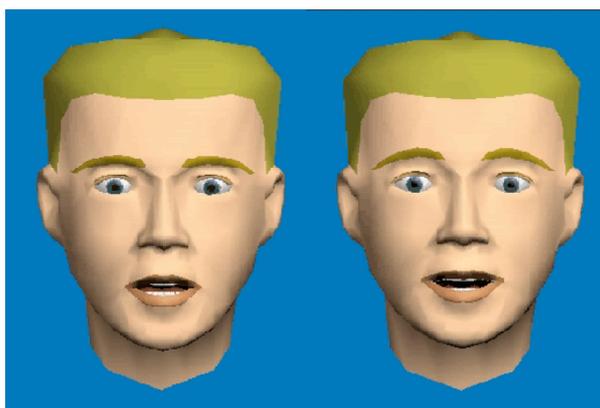
Figure 5. Percent “friendly” responses for each peak position and the two peak heights in all pair combinations.

## 5. Visual cues to friendliness (parametric synthesis)

To test the influence of visual cues on the perception of friendliness in an interrogative mode, two configurations combining different facial gestures were synthesized as in the question/statement experiment described above using an experimental version of the Infovox 330 diphone Swedish male MBROLA voice implemented as a plug-in to the WaveSurfer speech tool. The two configurations were designed to reinforce two of the audio examples described above, the low pitch in the early peak position (an information-oriented configuration) and the high pitch in the late peak position (a friendly/social-oriented configuration). For the information-oriented configuration, an early nod and a lowering gesture of the eyebrows were synchronized with the early focal accent (F0 peak) on the second syllable. For the friendly/social-oriented configuration, a late nod and a raising gesture of the eyebrows were synchronized with the late focal accent (F0 peak) on the final syllable. In addition, a smile was added throughout the utterance with increasing amplitude at the end of the

utterance after the nod. Samples of the configurations are shown in figure 6.

The two audio configurations were combined with the two video configurations making four stimuli. The stimuli were then converted to video files. A perception test was carried out in which the files were played using Windows Media Player and projected onto a screen in a classroom. The audio was played through high-quality loudspeakers. 27 native Swedish subjects were presented with three tokens of each of the four stimuli in random order. The subjects were asked to rate each stimulus on an unnumbered four-point scale where the endpoints were “friendly” and “less friendly.” Each stimulus was played twice in succession. The four-point scale was given point values from 0 to 3 where 3 represented the friendliest responses. Thus each stimulus could receive a maximum of 243 points as a “friendliness” score: 27 listeners x 3 tokens x max. 3 points = 243.



*Figure 6. The hypothesized information-oriented configuration (left) and social/friendly configuration (right) sampled in the middle of the second vowel of the utterance “Vad heter du?” (What is your name?).*

In figure 7, the results of the perception test are presented as a cumulative “friendliness score” for each of the four stimuli. It is clear that the consistent stimuli where both audio and visual cues are intended to convey the same attitude, are quite successful at conveying a low vs. a high degree of friendliness. For the inconsistent stimuli, the visual cues are stronger than the auditory cues. The friendly face combined with the early, low peak is perceived as friendlier than the info-face combined with the late, high peak.

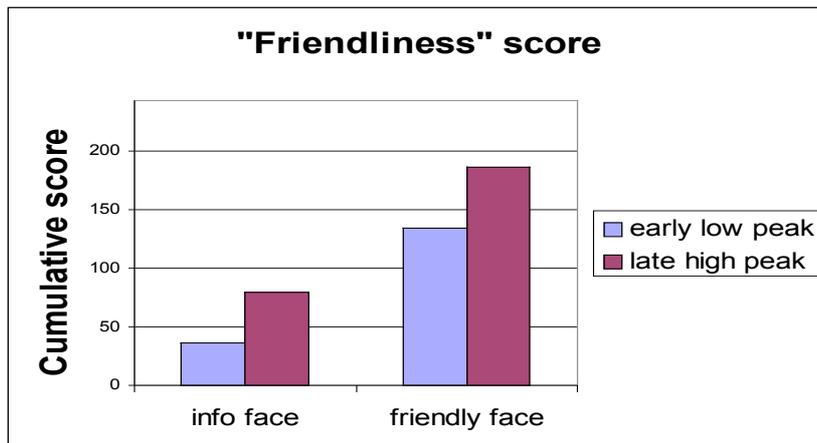


Figure 7. Results from the parametric synthesis test showing the cumulative “friendliness score” for each stimulus.

## 6. Data-driven visual synthesis

A different way of obtaining visual stimuli is by using data-driven visual synthesis. Facial movement data was collected by recording the positions of infrared markers on the face of an actor who was instructed to produce short sentences with different emotions (Beskow, et al. 2004). The 3D coordinates for each marker were registered and this information was then used to drive a talking head based on the MPEG 4 facial animation standard (Beskow and Nordenberg, 2005). Using the databases of different emotions results in talking head animations which differ in articulation and visual expression. For the current experiment, databases of angry, happy and neutral emotions were used to synthesize the same utterance as in the previous experiment, “*Vad heter du?*” (What is your name?). Samples of the three versions of the visual stimuli are presented in figure 8. As in the previous experiment, the three versions of the visual synthesis were combined with two audio configurations: low, early pitch peak and high, late pitch peak resulting in six stimuli. A perception test using these six stimuli was carried out in the same way and on the same occasion as the previous experiment using the same 27 subjects.

The results are presented in figure 9. It is quite clear that the face synthesized from the angry database elicited the lowest friendliness score. However, there is still evidence of interaction from the audio, as the angry face with the late, high peak received a higher friendliness score than did the angry face with the early, low peak. The faces from the other databases

(happy and neutral) elicited more friendliness responses, but neither combination of face and audio received as high a friendliness score as did the optimal stimulus from the parametric synthesis experiment. The happy face did not elicit more friendliness responses than did the neutral face, but the influence of the audio stimuli remained consistent for all the visual stimuli.



Figure 8. Visual stimuli generated by data-driven synthesis from the angry database (left) the happy database (middle) and the neutral database (right). All samples are taken from the middle of the second vowel of the utterance “Vad heter du?” (What is your name?).

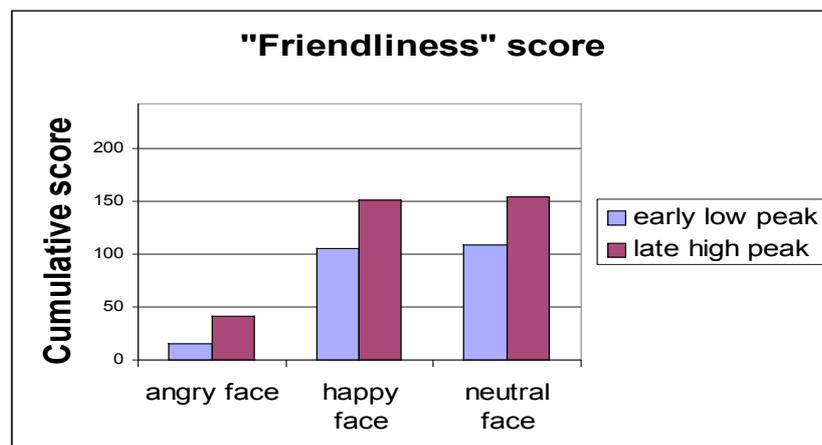


Figure 9. Results from the data-driven synthesis test showing the cumulative “friendliness score” for each stimulus.

## 7. Discussion and conclusions

The results of the experiments presented in this paper present evidence of the interaction of audio and visual cues to interrogative mode and to the signaling of friendliness in questions. Auditory cues dominated in signaling interrogative mode, while visual cues generally dominated in signaling attitude. The dominance of the auditory cues in signaling interrogative mode could indicate that auditory cues in question intonation may be less variable than the visual cues for questions, or we simply may not yet know enough about the combination of visual cues and their timing in signaling question mode to successfully override the auditory cues. Moreover, a final high rising intonation is generally a very robust cue to question intonation, especially in the context of perception experiments with binary response alternatives.

In these perception experiments, the visual modality was shown to be a powerful signal of attitude. However, the effects of the auditory cues for friendliness were clearly and consistently present in all the different versions of visual synthesis. This indicates that subjects make use of both modalities to make a judgement of speaker attitude and stresses the need to consider both the visual and audio aspects of expressive synthesis.

The results were clearer for the parametric synthesis than for the data-driven synthesis. This could partly be due to the fact that the parametric synthesis is more stereotypical and cartoon-like and therefore, when it successfully captures an attitude, it is perceived in clearer categorical terms. There may also be intrinsic differences in the way the two faces are perceived. In the data-driven synthesis, facial movement was limited mostly to the articulators (lips and mouth) and some eyebrow movement. The final smile and head movement present in the parametric synthesis was absent from the data-driven synthesis and may have contributed to the more successful “friendly” parametric face. It could be interesting and useful in future work to experiment with a combination of these techniques where for example head and eyebrow movement and the smile could be controlled within a data-driven visual synthesis framework.

## Acknowledgements

This research was carried out at the Centre for Speech Technology, a competence centre at KTH, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations. Special thanks to Jonas Beskow, Mikael Nordenberg and Magnus Nordstrand for creating the data-driven visual synthesis used in the perception test.

## References

- Beskow J., Cerrato L., Granström B., House D., Nordstrand, M., & Svanfeldt, G. (2004). The Swedish PF-Star Multimodal Corpora. In *Proc LREC Workshop, Multimodal Corpora: Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces*, 34-37. Lisbon.
- Beskow, J., & Nordenberg, M. (2005). Data-driven Synthesis of Expressive Visual Speech using an MPEG-4 Talking Head, In *Proceedings of INTERSPEECH 2005*, Lisbon, Portugal, 793-796.
- Caspers, J. (2003). On the function of low and high boundary tones in Dutch dialogue. *Proc 15th ICPHS*, Barcelona, 1771-1774.
- Cerrato, L., (2002). Some characteristics of feedback expressions in Swedish. In *Proceedings of Fonetik 2002*, TMH-QPSR 44, vol. 1, 101-104.
- Cruttenden, A. (1986). *Intonation*. Cambridge: Cambridge University Press.
- Ferrer, L., Shriberg, E; Stolcke, A. (2002). Is the speaker done yet? Faster and more accurate end-of utterance detection using prosody, In *Proceedings of ICSLP 2002*, Denver, Colorado, 2061-2064.
- Gårding, E. (1979). Sentence Intonation in Swedish, *Phonetica* 36, 207-215.
- Gårding, E. (1998). Intonation in Swedish, In D. Hirst & A. Di Cristo (Eds.) *Intonation Systems*. Cambridge: Cambridge University Press. 112-130.
- Granström, B., House, D., & Swerts, M. (2002). Multimodal feedback cues in human-machine interactions. In *Proc of the Speech Prosody 2002 Conference*, B. Bel & I. Marlien (Eds.). Aix-en-Provence: Laboratoire Parole et Langage, 347-350.

- Hirst, D., & Di Cristo, A. (1998). A survey of intonation systems, In D. Hirst and A. Di Cristo (Eds.) *Intonation Systems*. Cambridge: Cambridge University Press. 1-45.
- House, D. (2002). Intonational and visual cues in the perception of interrogative mode in Swedish, In *Proceedings of ICSLP 2002*, Denver, Colorado, 1957-1960.
- House, D. (2003). Perceiving question intonation: the role of pre-focal pause and delayed focal peak. *Proc 15th ICPHS*, Barcelona, 755-758
- House, D. (2005). Phrase-final rises as a prosodic feature in wh-questions in Swedish human-machine dialogue. *Speech Communication* 46: 268-283
- House, D. (2005). Phrase-final rises as a prosodic feature in wh-questions in Swedish human-machine dialogue. *Speech Communication*. (in press)
- Kohler, K. J. (2004). Pragmatic and attitudinal meanings of pitch patterns in German syntactically marked questions. In G. Fant, H. Fujisaki, J. Cao and Y. Xu (Eds.) *From traditional phonology to modern speech processing*, 205-214. Beijing: Foreign Language Teaching and Research Press.
- Ladd, D. R. (1996). *Intonation phonology*. Cambridge: Cambridge University Press.
- Srinivasan, R. J., Massaro, D. W. (2003) Perceiving prosody from the face and voice: Distinguishing statements from echoic questions in English. *Language and Speech* 46(1), 1-22.
- Sjölander, K., Beskow, J. (2000). WaveSurfer - a public domain speech tool, In *Proceedings of ICSLP 2000*, vol. 4, 464-467, Beijing, China.

## Biography

**David House** has a BA in English and Cinema Studies from the University of Kansas, 1975 and a Ph.D. in Phonetics, Lund University, 1991. He was Senior Lecturer, Department of Logopedics and Phoniatics, Lund University (1992-1994), Research Associate, Department of Linguistics and Phonetics, Lund University (1994-1997), and Senior Lecturer, Department of Languages, University of Skövde (1997-1998). He is now Associate Professor at the Department of Speech, Music and Hearing, KTH. Current research activity includes multimodal speech synthesis and prosody generation for spoken dialog systems, audio-visual perception of prosody, and educational applications of speech technology. Other research interests include phonetic variation in spontaneous speech and the perception of tone and intonation.

## Author's address

*David House*  
*Dept of Speech, Music and Hearing*  
*KTH (Royal Institute of Technology)*  
*Lindstedtsvägen 24*  
*SE-100 44 Stockholm*  
*Sweden*  
*phone: +468 790 7565 fax +468 790 7854*  
*email: davidh@speech.kth.se*  
<http://www.speech.kth.se>



# ACHIEVING TOPIC BY MULTIMODALITY IN EARLY DYADIC CONVERSATION

*Sari Karjalainen*

University of Helsinki, Finland

## **Abstract**

*The aim of this case-study is to analyse multimodality by looking at the structure and the content of early picture book conversations between an adult and a child at his preverbal stage. As, along with co-operation and intersubjectivity, the context is the central concept of the conversation analysis, my interest concerns contextuality. In this presentation, some aspects of context and their relation to topic and co-operative sequential structures of repair, are discussed.*

**Keywords:** Preverbal, conversation analysis, gesture, topic, contextuality

## **1. Introduction**

The process through which a child learns to speak constitutes a profound process of language socialization, through which the child becomes a competent socialized member of her society. The child's understanding of the world is calibrated by the local culture she is exposed to. The child's sensitivity to the 'context' develops through recognizing the existence of local understandings that are central to the young child's emerging grasp of the world of everyday life, in which her linguistic behaviour is situated, as Wootton (1997) has argued. The child becomes capable of grasping what is said to her when this talk is supported by nonverbal and contextual clues, and she develops a facility to initiate lines of conduct with an adult in ways

enabling her to have a degree of control over her immediate environment. The facility to attempt rectification and repair on occasions in which her communicative intentions have been misunderstood, is a critical skill (Wootton 1994; 1990; 1997). This facility will be analysed now by looking at how the two language users, differing in their linguistic competence, co-operate topically.

The 'preverbal' refers to a normal developmental stage when child's communication is founded rather on gesturing than on vocalization. From around the age of nine months the child displays an ability to refer to objects by deictic pointing gesture. During the child's second year of life, especially at the age of approximately 1 1/2 years, pointing is frequently used as a systematic way to respond to parental initiating acts. Moreover, the child's gestures can initiate sequences of action with the parent. Usually, an adult responds to a child's point by identifying or fetching an object or commenting on a quality or state of an object. The work on pointing suggests that pointings are not all of a piece: for example points can vary in their duration, or they can be accompanied by gaze and are usually accompanied by a vocalization (e.g. Lock et al. 1990; Wootton 1994; Franco and Butterworth 1996). As a context-bound gesture, pointing get its meaning in given referential situation.

Pantomimic gestures represent meanings more stable than those of pointing, so their semantic content does not change in different sequential contexts (Volterra 1981). Usually, pantomimic, or iconic, gestures appear later in development than deictics. Unlike pointing, children can differ a lot in using pantomimic gestures (Acredolo and Goodwyn 1990). Normally children are able to use symbolic gestures earlier than words, if they are encouraged to do so by their adult co-participants.

The pragmatic approach to language has introduced the term context. Bates (1976) defined pragmatics as 'rules governing the use of language in context'. The term seems to be defined by situated practice, and it does not seem to be possible to give a single, precise, technical definition of it. However, some dimensions of context have been noted (e.g. Goodwin and Duranti 1992): First, setting, i. e. the social and physical or spatial framework within which participants are situated. Second, behavioral environment, i.e. the way the participants use their bodies and behavior as a resource for framing and organizing their talk. Third, linguistic, also referred to as sequential context, i.e. talk itself provides context for other talk. At the preverbal stage the sequential context consists also of

nonverbal elements. Fourth, extrasituational context, i.e. background knowledge that extends far beyond the local talk.

Topic of a conversation has been included in the term context, as well. Topical sequence refers to contents of two (or more) utterances connected to each other by their common proposition. In early conversation the topic of the child's utterance, or the child's intention, is not always obvious. But, when an adult treats a child's utterance as being intentional, instances of joint attention occur, and different, more or less shared topics, are activated. It requires co-operation.

So, how are the shared meanings constructed? As Jones and Zimmerman (2003) have argued, intentionality is seen in participants' production and recognition of actions, in other words, the meanings become transparent in interaction between the child and the co-participant as it unfolds in a particular situation.

## **2. Data, method and transcription**

The method is qualitative and data-driven conversation analysis. The data base for the study is composed of mutual naturalistic picturebook conversations between child and adult videotaped in the homes of 6-8 normally developed Finnish children. Parents made between 6-10 hours of recordings of each child, on average having one-week recording period a month, from 1 to 2 years of age. Each period consists of episodes of picturebook reading varying between 5-15 minutes in length. From these episodes the data relevant to the research objectives are transcribed and analysed, case by case, according to the principles of conversation analysis. The current data, presented here, concerns one case, a boy, at an age of approximately 16 months. The co-participants are the child's caretaker (example 1) and the child's father (example 2).

The micro-analysis is focused on the sequential organization of participants' verbal and nonverbal action: especially adult's verbal and child's nonverbal turns including deictic and other gestures and gaze. Reflecting the focus of analysis, these aspects are always selectively represented in the transcript. Nonverbal elements in the transcript challenge the 'traditional' transcripts that include only speech turns, which can simply take the form of an orthographic gloss. The level of detail in the score-like transcripts, which I provide in subsequent presentation, is a result of two considerations. First, the critical evidence bearing on the

arguments being made is included in the transcript. As to this, to keep it simple, I provide only English translation of the talk. Second, it has to be intelligible to the standard reader. Various aspects of the behaviour of both parties at any moment are recorded on separate lines. When reading the score-like transcription, moving from left to right along these lines represent movement through time. The vertical relationships between the various aspects are represented as accurately as possible, especially in terms of their onset and cessation.

Further details about the conventions used within particular lines are given below:

<b>A</b>	adult
<b>E</b>	child
<b>Talk:</b>	in standard orthography (only English translations)
<b>Hand positions:</b>	rh (right hand), lh (left hand)
POINT P1*****	points to picture 1 (static hand position when pointing)
POINT P1*** ** *	points to picture 1 (movement of the hand when pointing)
GE ^^^^^^	other gesture-like movement
TB.....	touches book
TP.....	turns page
<b>Gaze:</b>	
A__book	gaze shift from A to book (from head <b>orientation</b> )
<b>Other:</b>	
1	beginning and end of act <b>or</b> movement within act
[	beginning of overlap between two actions
]	end of overlap
(1.0)	1 sec pause
(.)	micropause (< 0.5 sec)

### 3. Two examples of topic achievement

Reviewing the findings in the analysis, I take the view that gestures are this child's method to take part in the picture book conversation at preverbal stage (Karjalainen 2001). The child (here called 'E') is very skilful at using pointing and other gestures before his first words. In example 1, the child makes use of *pointing* gesture, in presenting his intention to his caretaker.

#### *Example 1*

- 1 A there's a bear sawing a tree. they make this kind of stock of it. then or in fact they are logs they are pulled with that kind of [tow-tractor [there's such a stump left  
[POINTP1\*\*\*\*\*]
- 1 E [eh eh eh [eh eh eh  
[POINT P1\*[\*\*\*\*\*[POINT SELF\*\* | \*\*\* | \*\*\* | \*\*\* | \*
- 2 A hmmm (1.0) what (.) now I don't understand (.) [ I don't understand (.) [there's a saw  
[POINT P2\*\*\*
- 2 E [eh eh eh [eh  
\*\*|\*\*\*|\*|\*\*|\*\*\*|\*\*\*|\*\*\*\*\*|\*\*\*\*\* [ POINT P2\*\*\*\*\*\* [\*\*\*\*\*
- 3 A (.) have you sawn too (.) [ have you sawn also you have as I remember also you
- 3 E [ NODS  
\*\*\*\*\*|
- 4 A should have that kind of saw somewhere there

When A is speaking about the picture, E identifies one of the pictures in the book (line 1). A comments on it. In her comment *there's such a stump left*, A focuses on the object *stump*, which is the product of the act of sawing. Then, simultaneously with A's comment, E points to himself and does not stop the repetitive act of pointing, until A has focused on the act, which is seen on line 2. A makes an explicit comment on not understanding E's pointing. Next, E points at a new picture on line 2. This pointing overlaps with A's repetition of not understanding. Later, on line 2, A comments on the picture E is now pointing at: *there's a saw* and, further asks *have you sawn, too*. As soon as E has finished his sustained pointing at picture two, he nods, on line 3, which is in line with A's suggestion. A still continues, by making an argument, on lines 3 and 4, which brings the negotiation to a conclusion.

How is the sequential context is organized here? When looking at these two pointing gestures, on line 1 and 2, and the interpretations A makes of them, we see two different sequential contexts.

So, A is willing to accept E's topical offer (on line 1) here. She treats E's turns as being intentional, and by saying *I don't understand* she accepts E having information, that she doesn't have herself. After E has pointed the second picture, A makes her suggestion: *there's a saw, have you sawn too*. During the turn the child points to the picture of the saw presented). This picture provides new information, which A indirectly seeks. Moreover, the new information given here, is combined with the information given by E in his first turn.

We can only speculate on the backgrounds of A's interpretative turns, but they make sense because they are the resources available to the participants, and previously I have noticed that the adult's interpretation reveals the immediate sequential context and the contents of the prior turn. So, the sequence beginning with A's turn *there's such a stump left*, is preceded by A's turn with multiple referents as a source of information to make exact propositional matching with E's pointing to himself. Additionally, the focus of A's turn "stump" may have a propositional relation to child's turn.

Whatever the roots of A's understandings are, the child uses his nonverbal modality precisely and successfully. As a consequence, the topic offered by E, is achieved.

In another example (2), mainly *pantomimic* gestures are used for the purpose of presenting a topic.

*Example 2*

1 E [TP.....]

1 A it's a race-car

2 E [rh GE 1^[[[GE 2 ^[[[ [A\_\_\_\_\_book  
[rh GE 1^[[[GE 2 ^[[[ llh TP.....llh TB.....  
[khoh [eh eh

2 A [a little mini-car

3 E [ A \_\_\_\_\_  
[rh up lGE 2^[[[2a^[[[llh TP.....llh TP.....  
llh TP..... l eh [eh eh eh

3 A oh you drive the race-track (1.5) what do you mean [rh up.....

4 E \_\_\_\_\_book  
.....GE 2 ^[[[GE 3 ^[[[ llh [TB.....]TP.....  
leh eh [kho khoh kho

4 A [oh]is it the loading-shovel (.) you're looking for

5 E .....[MOVES HAND TO BOOK...l TB.....

5 A [are you still looking for the loading-shovel [ TB.....  
let's look where it is

6 E [.....]TP.....both hands [TP..... l

6 A [turn page] after page so it will be found

7 E [TB.....TP.....

7 A hmmm (.) next pa [ge will have a shovel (1.0) not yet shall we look still

Before focusing on the E's act on line 4, we notice, that this sequence begins with A identifying a *race car* on line 1. Further, on line 2, A says *a little mini-car*, by this continuing to describe the race-car mentioned before. Then, at the same time with this, E makes two arm movements, one of which is a shovelling-like movement (gesture 2). After enacting this movement of right arm with twisted wrist, so called gesture 2 in the transcript, E turns the page with his left hand. Moreover, E makes a glance towards A, vocalizing at the same time (line 2).

Then, on line 3, E repeats the right hand movement. This movement seems to have two phases: the first, the arm enacts a brief up and down movement ('GE 2') and the second, referred as 'GE 2a' in the transcript, seems like a rotating movement of the arm, and while doing this, E is looking at A. After finishing the movement, the arm is sustained in an up position.

Now, on line 3, A makes an immediate candidate understanding: *oh you drive the race-track*, then continues immediately: *what do you mean*. On line 4, E continues with the right hand movement for a while, and then, he makes another act with both hands: clenching his fists he is enacting the repetitive movement of both arms moving back and forth, like holding the *steering sticks*, and at this point, simultaneously vocalizing with specific ('velar fricative') repetitive sound. Further, E moves his hand towards the page, which coincides with A's turn: A makes a new suggestion on line 4: *oh is it the loading-shovel you're looking for*. On line 5 E's hand has now reached the book and, then, again, like earlier on lines 2 and 3, he turns the page. A repeats his suggestion (also line 5). At the same turn, he accepts his own suggestion by saying *let's look where it is, turn page after page so it will be found*. With this, he concludes the negotiation, and both the participants start turning the pages, searching for the missing loading shovel.

By looking at the sequential contexts for child's gestures in this example, we see, that, having common propositional background, the phrase *oh you drive the track* has an uptake of a prior sequence (A's turn on lines 1 and 2). A interprets the 'rotating' movement of E's arm, as arising out of the topic *race-car*: really, the form of E's gesture may be related to driving the track. This was proved to be wrong interpretation. A's metalinguistic comment refers to having problems in identifying the topic for sure, and explicitly searching for an answer on line 3. On line 4 E's new way of referring (GE 3 or the 'steering stick'-gesture) gives new information on the topic, the information A asked for directly by asking *what do you mean*.

Again, there is a cumulative effect of these two gestures, increasing A's understanding. Additionally, E's simultaneous attempt to turn the page refers to searching for something. E is here very active with turning the pages, on lines 2 and 3 and also on lines 4 and 6, also referring to book with gaze. Moving the gaze from the book to A, and from A to book, on lines 2 and 4, it can be seen as an active moment of indication one's intention relating to ongoing topic. So, in this case of A's misunderstanding, E manages to modify her expression by using another iconic gesture, which directs adult's interpretation to a new, correct direction. Consequently, the intersubjectivity between the participants is maintained in the conversation.

#### **4. Conclusion**

I gave two examples of different negotiations, where the child uses different preverbal modalities and resources activating the context, both noticeable and non-noticeable aspects of it.

During the picturebook sessions, there are conversational sequences, where participants focus their attention to shared referents. Within these sequences, an elementary sequential structure is an adjacency pair, consisting of child's gesture and adult's verbal turn, by which adult interprets child's gesture. These structures can be linked up to lengthy sequences, in which the topic is modified. Further, there is, in some instances, topical shift from noticeable to non-noticeable referents. The shift may be fluent, when the content and form of adult's turn reveals the immediate sequential context, the topic of the prior turn or turns. The non-noticeable referents are retrieved from the child's real life experience, near or far beyond the local situation. This may challenge intersubjectivity: thus, the topical shift is not always fluent. Practically, during some of these episodes, explicit negotiations on meaning have to be followed through the episodes I have discussed here.

The pointing can be analysed as a situated activity system in which action is built by assembling diverse semiotic resources into locally relevant multimodal packages (Goodwin 2002). In example 1, the child first activates the non-noticeable context by pointing at himself, which causes understanding problems. Then he activates the physical, or noticeable, context, in a way that brings new information on his previous intention. The topic the child is offering comes far behind the current situation from the child's real life experience.

In example 2, the topic, or child's intention, has its roots also in the past, not as far as in example 1, but in a moment a few minutes earlier. Before the sequence presented here, there was a lengthy sequence, in which these two gestures were used, when referring to the loading shovel. The first was the shovel-gesture with one arm and a twisted wrist, and the second steering-stick gesture with both arms. The use of the first of these two is unsuccessful in a situation presented here, but the cue included in the second gesture, may bring the crucial information.

The gestures emerge, within a field, already endowed with meaning (see Goodwin 2002). It is this field, that may be the cause for understanding, and the misunderstanding, as well.

So, what I tried to show, having a grasp of different dimensions of context, is a basis of successful conversation. This is highlighted, particularly, in early conversation, where topic is achieved by multimodal negotiations on meaning. In both examples an adult is asking for the missing information with explicit, metalinguistic comment, which I see as a construction of shared knowledge. As we have seen here, these explicit negotiations are pursued, as due to the different levels of context and co-operative multimodal structures, and particularly here, sequential repair-like structures. As a consequence, even the non-noticeable topics can be achieved.

## References

- Acredolo, L. P., & Goodwyn, S. W. (1990). Sign Language Among Hearing Infants: The Spontaneous Development of Symbolic Gestures. In V. Volterra & C. J. Erting (Eds.) *From Gesture to Language in Hearing and Deaf Children*. Berlin, Springer-Verlag: 68 - 78.
- Bates, E. (1976). *Language and Context. The acquisition of pragmatics*. New York, Academic Press.
- Franco, F. & Butterworth, G. (1996). Pointing and social awareness: declaring and requesting in second year. *Journal of Child Language*, 23: 307 - 336.
- Goodwin, C. (2002). Pointing as situated practice. In S. Kita (Ed.) *Pointing: Where language, culture and cognition meet*. Hillsdale, NJ: Lawrence Erlbaum Associates: 217-241.

- Goodwin, C., & Duranti, A. (1992). Rethinking context: an introduction. In Goodwin, C., & Duranti, A. (Eds.) *Rethinking context. Language as an interactive phenomenon. Studies in the Social and Cultural Foundations of Language No. 11*. Cambridge, Cambridge University Press: 1 - 42.
- Jones, S. E. & Zimmerman, D. H. (2003). A child's point and an achievement of intentionality. *Gesture* 3:2, 155-185.
- Karjalainen, S. (2001). Sormi sanojen joukossa: Yhteistyö ja konteksti topiikin ylläpidossa aikuisen ja lapsen välisissä kuvakirjakeskusteluissa. [Finger among words: co-operation and context maintaining topic in picture book conversations between adult and child.] University of Helsinki, Department of Phonetics, Master's Thesis.
- Lock, A., Young, A., Service, V. & Chandler, P. (1990). Some Observations on the Origins of the Pointing Gesture. In V. Volterra & C. J. Erting (Eds.) *From Gesture to Language in Hearing and Deaf Children*. Berlin, Springer-Verlag: 42 - 55.
- Volterra, V. (1981). Gestures, signs, and words at two years: When does communication become language? *Sign Language Studies*, 33: 351 - 361.
- Wootton, A. J. (1994). Object transfer, intersubjectivity and third position repair: early developmental observations of one child. *Journal of Child Language*, 21, 543-564.
- Wootton, A. J. (1990). Pointing and interaction initiation: the behaviour of young children with Down's syndrome when looking at books. *Journal of Child Language*, 17: 565 - 589.
- Wootton, A. J. (1997). *Interaction and the development of mind*. Cambridge University Press.

## **Biography**

**Sari Karjalainen** received the M.A degree in logopedics from the University of Helsinki in 2001. From 2003 onwards she has worked for her Ph.D degree, early conversation as the subject, at the Graduate School for Language Studies (Langnet).

### **Author's address:**

*Sari Karjalainen  
University of Helsinki  
Department of Speech Sciences (Logopedics)  
Siltavuorenpenger 20 A/F  
P.O. Box 9  
00014 University of Helsinki  
Finland  
phone: + 358 9191 29385  
e-mail: sari.karjalainen@helsinki.fi  
fax: + 358 9191 29341*

# EVALUATION OF A MOBILE MULTIMODAL SERVICE FOR DISABLED USERS

*Knut Kvale, Narada Warakagoda and Marthin Kristiansen*  
Telenor R&D, Fornebu, Norway

## **Abstract**

*Multimodal interfaces make it possible to communicate with services in many different ways, and it is claimed that this freedom is particularly useful for disabled persons. To test this hypothesis we have developed a flexible multimodal interface to a public web-based bus-route information service for the Oslo area. The original service is text-based. Our new interface running on a PDA converts the service to a map-based multimodal service supporting speech, graphic/text and pointing modalities. Here users can choose to use speech or point on the map or even use a combination of tap and talk simultaneously (so-called composite multimodality) to specify the required information including arrival and departure stations. The response from the system is read aloud by speech synthesis while the same information is shown textually on the screen. We have carried out two rounds of user-evaluation for this multimodal interface. The first one was a scenario-based evaluation carried out for five disabled users. The main conclusion was that these users' attitude towards the multimodal system was positive and they recognized the advantages and the potential of such systems. Then we carried out an in-depth evaluation of how a dyslectic and an aphasic used our system. These two could neither use the corresponding public speech-based telephone service nor the text-based web-service, but they found the new multimodal interface very useful. Our paper presents the detailed results of these evaluations.*

**Keywords:** Multimodal interfaces, mobile terminals, disabled users, modality adaptation, aphasia, dyslexia

## 1. Introduction

Today's society faces the problem of an increasing exclusion of a growing number of elderly and disabled people for whom the employment of a wide range of new communication and information services becomes more and more difficult. One solution to this problem is to equip the electronic services and applications with intelligent modality adaptive interfaces that let people choose their preferred interaction style depending on the actual task to be accomplished, the context, and their own preferences and abilities.

In order to test the hypothesis that multimodal inputs and outputs really is useful for disabled people, we have developed a flexible multimodal interface to a public web-based bus-route information service for the Oslo area. The original public service on the web, which has both HTTP and WAP interfaces, is text based (i.e. only unimodal). The users have to write the names of the arrival and departure stations to get the route information, which in turn is presented as text. Our multimodal interface for small mobile terminals converts the web service to a map-based multimodal service supporting speech, graphic/text and pointing modalities as inputs. Thus the users can choose whether to use speech or point on the map, or even use a combination of tap and talk simultaneously to specify the arrival and departure stations. The response from the system is presented as both speech and text. We believe that this multimodal interface gives a freedom of choice in interaction pattern for all users. For normal able-bodied users this implies enhanced user-friendliness and flexibility in the use of the services, whereas for the disabled users this is a means by which they can compensate for their not-well-functioning communication mode.

In order to test whether multimodality of this kind actually may help disabled users, we have carried out two different user evaluations. First, a qualitative scenario-based evaluation followed by a questionnaire was carried out for five disabled users. The goal was to study the acceptance of the multimodal service by the disabled users. It also allowed us to gather the users' views about improvements needed for the system. Secondly, we performed an in-depth evaluation of how a dyslectic and an aphasic made use of our system.

The paper first describes the multimodal system architecture and bus information system. Then the user evaluations are discussed.

## 2. System Architecture

Our multimodal bus information system is a modified version of the MUST system (Almáida 2002), (MUST 2002). It is based on the Galaxy communicator (GALAXY 2002) and has a hub-spoke type architecture as shown in Figure 1. The server part of the system consists of six separate modules which can communicate with each other through the central facilitator module “hub”. All the server side modules run on a PC, while the client runs on a PDA, here a Compaq iPAQ. The client consists of two main components handling voice and graphical (GUI) modalities.

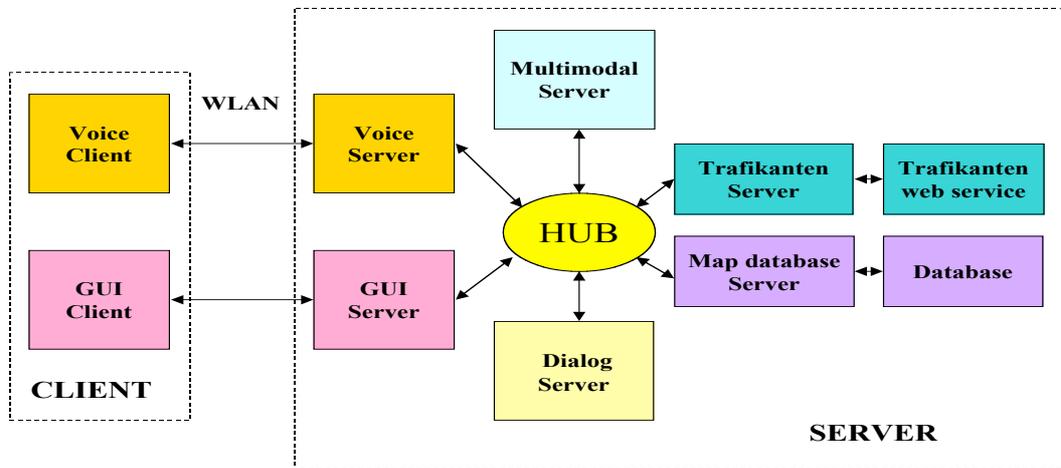


Figure 1. System architecture of the multimodal bus information system

When the user query contains both voice and pointing inputs the information flow through the system is as follows. First the client records the pointing and voice signals and transfers them to the server over the wireless connection, which in our case is a wireless local area network (WLAN) based on the IEEE 802.11b protocol. The GUI server and voice server collect these signals respectively. While the GUI server annotates the pointing signal in a suitable way, the voice server performs a speech recognition operation to extract the concepts carried by the speech signal. Then both the voice server and GUI server pass the concept values further to the multimodal server, which combines speech and pointing information to form a single message, if they lie within a predefined time window. The combined information is passed to the dialog manager which actually completes the multimodal integration process and interprets it to perform

the necessary action depending on the current dialog state. In a typical situation the appropriate action would be to contact the (map) database server and the "Trafikanten" web service to get the necessary information (Trafikanten, 2005).

The dialog server processes the results from the database and the "Trafikanten" server to produce a single message, before sending it to the multimodal server. The multimodal server splits this message (by a simple "fission" operation) into a voice part and a GUI part. These two parts are then sent further to the voice client and the GUI-client, where they are presented to the user.

We have applied the Scansoft SpeechPearl 2000 automatic speech recogniser (ASR) for Norwegian (Scansoft, 2005). The vocabulary consisted of 57 different bus stations. These names were defined both in the concept <from\_station> and under the concept <to\_station>. To improve the recognition accuracy, we defined some "composite words" containing word-pairs or word triples that were often spoken together: "and\_here", "here\_and", "and\_this", "and\_for\_this", "and\_this", "and\_for\_this". The total vocabulary consisted of 151 "words". For Norwegian Text-to-Speech (TTS) synthesis we used "Telenor Talsmann" (Talsmann).

More details about the system architecture are provided in (Kvale et.al. 2003a), (Warakoda et.al 2003), (Kvale et.al 2004).

### **3. The Multimodal Bus Information Service**

The interface of our multimodal service is provided by the client application running on a mobile terminal. When the client is started, and connected to the server, the main page of the server is presented to the user. This is an overview map of the Oslo area where different sub-areas can be zoomed into, as shown in figure 2. Once zoomed, it is possible to get the bus stations in the area displayed. The user has to select a departure station and an arrival station to get the bus route information.

The users are not strictly required to follow the steps sequentially. They can e.g. combine several of them, whenever it makes sense to do so.

Our service fulfils the W3C Multimodal Interaction Requirements (W3C 2003) for both *simultaneous* inputs (i.e. the speech and pointing inputs are interpreted one after the other in the order that they are received) and

*composite* inputs (i.e. the speech and pointing inputs at the “same time” are treated as a single, integrated compound input by downstream processes). Users may also communicate with our service unimodally, i.e. by merely pointing at the touch sensitive screen or by speech only. The multimodal inputs may be combined in several ways, for instance:

- The user says the name of the arrival bus station and points at another bus station at the map, e.g.: “I want to go from Jernbanetorget to here”
- The user points at two places at the screen while saying: “When goes the next bus from here to here”

In both scenarios above the users point at a bus station within the same time window as they utter the underlined word, “here”. In order to handle two pointing within the same utterance, we defined an asymmetric time window within which speech and pointing are treated as a composite input if:

- Speech is detected within 3 seconds after a pointing
- Pointing is detected 0.85 second before the speech signal ends

Both pointing and speech can be used in all operations including navigation and selecting bus stations. Thus the user scenarios can embrace all the possible combinations of pointing and speech input. The received bus route information is presented to the user as text in a textbox and this text is also read aloud by synthetic speech.

Thus we expect that the multimodal service may prove useful for many different types of disabled users, such as:

- Persons with hearing defects and speaking problems may prefer the pointing interaction.
- Blind persons may only use the pure speech-based interface
- Users with reduced speaking ability may use a reduced vocabulary while pointing at the screen.

Figure 2 and 3 show typical screen sequences for a normal user and a user with reduced speaking ability respectively. In both cases they want to go from “Fornebu” to “Jernbanetorget”.

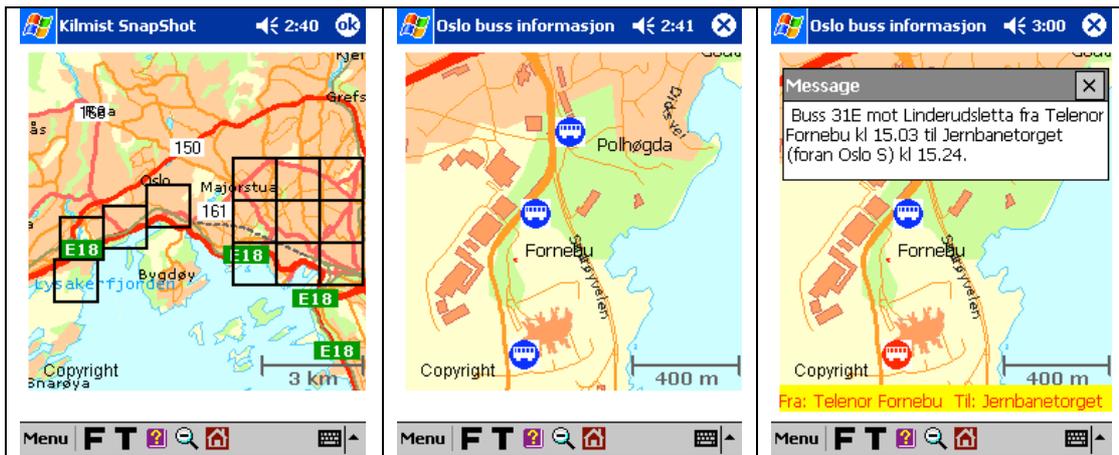


Figure 2. A typical screen sequence for a normal user.

- 1) Overview map: User taps on the submap (the square) for Fornebu
- 2) User taps on bus station Telenor and says "when is the next bus from here to Jernbanetorget
- 3) Bus information pops up on the screen and is simultaneously read aloud

#### 4. User evaluations

One key-question in these kinds of evaluations is how to introduce the service and the possibility of multimodal interaction to new users. Different introductions have impact on how people use the services (Kvale et.al., 2003b). We applied two different strategies for introduction:

- For the scenario-based evaluation we produced an introduction video showing the three different interaction patterns: Pointing only, speaking only, and a combination of pointing and speaking. We did not subtitle the video, so deaf people had to read the information on a text sheet.
- For the in-depth evaluation of the dyslectic and aphasic user we applied so-called model based learning, where a trusted supervisor first showed how he used the service and carefully explained the functionality.

#### 4.1 Scenario-based evaluation

A qualitative scenario-based evaluation followed by a questionnaire was carried out for five disabled users (Kristiansen, 2004). The goal was to study the acceptance of the multimodal service by the disabled users.

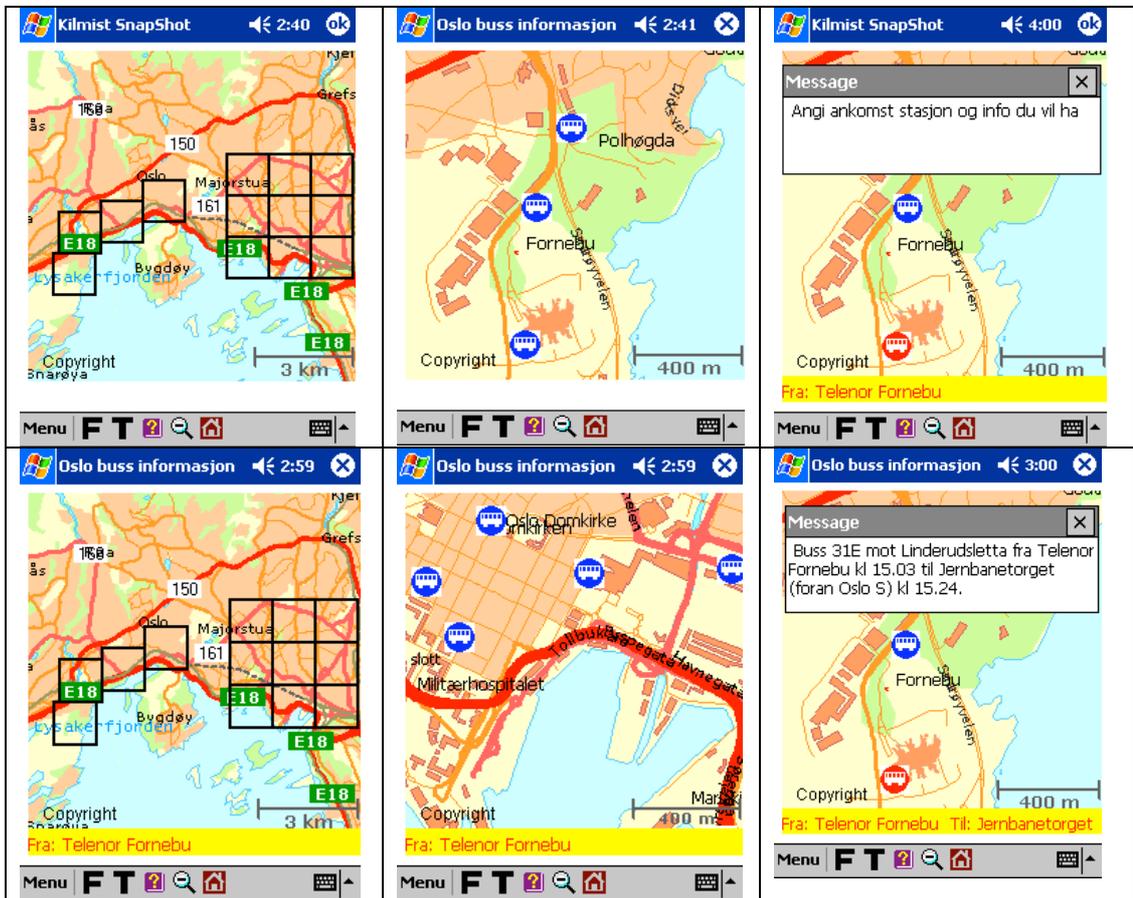


Figure 3 A typical screen sequence for a user with reduced speaking ability. 1) Overview map: User taps on the submap (the square) for Fornebu. 2) User says "next bus here Jernbanetorget" and taps on bus station Telenor. 3) System does not recognize the arrival station. Therefore the user selects it by using pen. But first user taps on the zoom-out button to open the overview map. 4) User taps on the submap, where bus station Jernbanetorget lies. 5) User taps on the bus station Jernbanetorget. 6) User can read the bus information

The users were recruited from Telenors handicap program (HCP) in the spring 2004. They were in their twenties with an education of 12 years or more. The disabilities of the five users are:

- Muscle weaknesses in hands
- Severe hearing defect and a mild speaking disfluency

- Wheelchair user with muscular atrophy affecting the right hand and the tongue
- Low vision
- Motor control disorder and speech disfluency.

The scenario selected for this evaluation involved finding bus route information for two given bus stations. The users should complete the task in three different manners: By using pen only, speech only and by using both pen and speech. The tests were carried out in a quiet room with one user at a time. All the test persons were able to complete the tasks in at least one manner:

- They were used to pen-based interaction with PDAs so the pen only interaction was easy to understand and the test users accomplished the task easily. Persons with muscle weaknesses in hands or with motor control disorder demanded the possibility of pointing at a bigger area around the bus stations. They also suggested that it might be more natural to select objects by drawing small circles than by making a tap (see also (Oviatt 1997)). The person with hearing defects and speaking disfluency preferred the pen only interaction.
- The speech only interaction did not work properly, partly because of technical problems with the microphone and speech recogniser and partly due to user behaviour such as low volume and unclear articulation.
- The multimodal interaction was the last scenario in the evaluation. Hence some persons had to be explained once more this functionality before trying to perform this task. The persons with muscular atrophy combined with some minor speaking problems had great benefit from speaking short commands or phrases while pointing at the maps.

In the subsequent interviews all users expressed a very positive attitude to the multimodal system and they recognized the advantages and the potential of such systems.

#### ***4.2 In-dept evaluation of two disabled persons***

We have performed an in-depth evaluation of how a dyslectic and an aphasic used our system.

### ***A dyslectic test user***

Dyslexia causes difficulties in learning to read, write and spell. Short-term memory, mathematics, concentration, personal organisation and sequencing may be affected. About 10% of the population may have some form of dyslexia, and about 4% are regarded as severely dyslexic (Dyslexia 2005).

Our dyslectic test person was fifteen years old and had severe dyslexia. Therefore he was very uncertain and had low self-confidence. He was not familiar with the Oslo area. Therefore we told him the two bus station names he could ask for: "From Telenor to Jernbanetorget". He had however huge problems with remembering and expressing these names, especially "Jernbanetorget" because it is a long name. Hence we changed the task to asking for the bus route information: "From Telenor to Tøyen". These words were easier for him, but he still had to practise a couple of times to manage to remember and pronounce the two bus stations.

However, after some training, he had no problem using the PDA. He quickly learned to navigate between the maps by pointing at the "zoom"-button. He also talked to the system. When the system did not understand his confirmation input, "yes", he immediately switched to pointing at the "yes" alternative on the screen (he had no problem with reading short words). If the bus station has a long name he could find it on the map and select it by pen instead of trying to use speech.

### ***An aphasic test user***

Aphasia refers to a disorder of language following some kind of acquired brain damage, for example, due to a stroke. Aphasia denotes a communication problem, which means that people with aphasia have difficulty in expressing thoughts and understanding spoken words, and they may also have trouble reading, writing, using numbers or making appropriate gestures.

Our test person suffered a stroke five years ago. Subsequently he could only speak a few words and had paresis in his right arm and leg. In the first two years he had the diagnosis *global aphasia*, which is the most severe form of aphasia. Usually this term applies to persons who can only say a few recognizable words and understand little or no spoken language (AFS 2005). Our test person is not a typical global aphasic any longer. He has made great progress, and now he speaks with a clear pronunciation and prosody. However, his vocabulary and sentence structure are still restricted, and he often misses the meaningful words - particularly numbers, important

verbs and nouns, such as names of places and persons. He compensates for this problem by a creative use of body language and by writing numbers. He sometimes writes the first letters of the missing word, and lets the listener guess what he wants to express. This strategy worked well in our communication. He understands speech well, but may have problems interpreting composite instructions. He is much better at reading and comprehending text than at expressing what he has read.

Because of his disfluent speech characterized by short phrases, simplified syntactic structure, and word finding problems, he might be classified as a Broca's aphasic, although his clear articulation does not fit completely into this classification.

He is interested in technology and has used a text-scanner with text-to-speech synthesis for a while. He knew Oslo well and was used to reading maps. He very easily learned to navigate with the pen pointing. He also managed to read the bus information appearing in the text box on the screen, but he thought that the text-to-speech reading of the text helped the comprehension of the meaning.

His task in evaluation was to get bus information for the next bus from Telenor to Tøyen by speaking to the service. These stations are on different maps and the route implies changing buses. Therefore, for a normal user, it is much more efficient to ask the question than pointing through many maps and zooming in and out. But he did not manage to remember and pronounce these words one after the other.

However, when demonstrated, he found the composite multimodal functionality of the service appealing. He started to point at the from-station while saying “this”. Then he continued to point while saying “and this” each time he pointed not only at the bus stations but also at function buttons such as “zoom in” and when shifting maps. It was obviously natural for him to talk and tap simultaneously. Notice that this interaction pattern may not be classified as a composite multimodal input as defined by (W3C 2003), because he provided exactly the same information with speech and pointing. We believe, however, that if we have had spent more time in explaining the composite multimodal functionality he would have taken advantage of it.

He also tried to use the public bus information service on the web. He was asked to go from “Telenor” to “Tøyen”. He tried, but did not manage to

write the names of the bus stations. He claimed that he might have managed to find the names in a list of alternatives, but he would probably not be able to use this service anyway due to all the problems with reading and writing. The telephone service was not an alternative at all for him because he was not able to pronounce the station names. But he liked the multimodal tap and talk interface very much and characterised it spontaneously as "Best!", i.e. the best alternative to get the information needed.

## **5. Discussion**

We have performed qualitative evaluations of how users with reduced ability applied the multimodal interface. Thus, the results are by no means statistically significant. We are aware that for instance aphasics are different and even the same person may vary his or her performance from one day to the next. Still, it seems reasonable to generalise our observations and claim that for aphasics a multimodal interface may be the only useful interface to public information services such as bus timetables. Since most aphasics have severe speaking problems they probably will prefer to use the pointing option, but our experiment indicates that they may also benefit from the composite multimodality since they can point at the screen while saying simple supplementary words.

In the evaluations we tried to create a relaxed atmosphere and we spent some time having an informal conversation before the persons tried out the multimodal service. In the scenario-based evaluations only the experiment leader and the test person were present. In the in-depth evaluations the test persons brought relatives with them. The dyslectic user had his parents with him, while the aphasic came with his wife. The evaluation situation may still be perceived as stressful for them since two evaluators and two teachers were watching the test person. The stress factor was especially noticeable for the young dyslectic.

## **6. Concluding Remarks**

Our composite multimodal interface to a map-based information service has proven useful for persons with muscular atrophy combined with some minor speaking problems, and also for a severe dyslectic and an aphasic. The severe dyslectic and aphasic could not use the public service by

speaking and taking notes in the telephone-based service or by writing names in the text-based web service. But they could easily point at a map while speaking simple commands.

## Acknowledgements

We would like to express our thanks to Tone Finne, Eli Qvenild and Bjørgulv Høigaard at Bredtvet Resource Centre for helping us with the evaluation and for valuable and fruitful discussions and cooperation. We are grateful to Arne Kjell Foldvik and Magne Hallstein Johnsen at the Norwegian University of Science and Technology (NTNU) for inspiration and help with this paper.

This work has been financed by the BRAGE-project of the research program “Knowledge development for Norwegian language technology” (KUNSTI) of the Norwegian Research Council.

- Bredtvet Resource Centre: <http://www.statped.no/bredtvet>
- BRAGE: <http://www.tele.ntnu.no/projects/brage/index.php>
- KUNSTI: <http://program.forskningsradet.no/kunsti/>
- NTNU: <http://www.ntnu.no/>

## References

- AFS 2005. Aphasia Fact Sheet, URL <http://www.aphasia.org/NAAfactsheet.html>
- Almeida, L. et al. 2002. “The MUST guide to Paris - Implementation and expert evaluation of a multimodal tourist guide to Paris”, Proc. ISCA Tutorial and Research Workshop (ITRW) on *Multi-Modal Dialogue in Mobile Environments*, (IDS 2002), pp. 49-51, Kloster Irsee, Germany.
- Dyslexia Institute URL - <http://www.dyslexia-inst.org.uk/>
- GALAXY Communicator: <http://fofoca.mitre.org/>
- HCP - Telenors Handicap-program. URL: <http://www.telenor.no/hcp/>
- Kristiansen, M. 2004. *Evaluering og tilpasning av et multimodalt system på en mobil enhet*”, Master Thesis NTNU (in Norwegian).
- Kvale, K., Warakagoda, N.D. and Knudsen, J.E. 2003a. *Speech centric multimodal interfaces for mobile communication systems*. *Elektronikk* nr. 2, 2003, pp. 104-117.

- Kvale, K., Rugelbak, J., & Amdal, I., (2003b). *How do non-expert users exploit simultaneous inputs in multimodal interaction?*, Proc. International Symposium on Human Factors in Telecommunication (HfT 2003), Berlin, pp. 169-176.
- Kvale, K., Knudsen, J.E., & Rugelbak, J., (2004). *A Multimodal Corpus Collection System for Mobile Applications*, Proc. Multimodal Corpora - Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces, Lisbon, pp. 9-12.
- MUST 2002. *MUST - Multimodal and Multilingual Services for Small Mobile Terminals*. Heidelberg, EURESCOM Brochure Series, May 2002. URL: <http://www.eurescom.de/public/projects/P1100-series/P1104/default.asp>
- Oviatt, S. et.al. (1997). *Integration and synchronization of input modes during multimodal human-computer interaction*. In Proceedings of Conference on Human Factors in Computing Systems: CHI '97 415-422. New York: ACM Press.
- Scansoft Speechpearl. URL <http://www.scansoft.com/speechpearl/>  
Telenor Talsmann.  
[http://www.telenor.no/fou/prosjekter/taletek/talsmann/tts\\_more.htm](http://www.telenor.no/fou/prosjekter/taletek/talsmann/tts_more.htm)
- Trafikanten. URL - <http://www.trafikanten.no>
- Warakagoda, N. D., Lium, A. S., & Knudsen, J. E. (2003). *Implementation of simultaneous co-ordinated multimodality for mobile terminals*. The 1st Nordic Symposium on Multimodal Communication, Copenhagen, Denmark
- W3C 2003. *Multimodal Interaction Requirements (W3C note)*  
URL- <http://www.w3.org/TR/2003/NOTE-mmi-reqs-20030108/>

## Biography

**Knut Kvale** (Ph.d.) and **Narada Warakagoda** (Ph.d) are research scientists at Telenor R&D in Norway. Knut Kvale also holds a Professor II position at the University of Stavanger, Norway. We have been working on various aspects of speech technology for more than a decade and since 2001 we have implemented and tested speech centric multimodal systems for small mobile terminals.

**Marthin Kristiansen** (M.Sc) carried out his Master Thesis for NTNU at Telenor R&D in the spring 2004, and is now employed at Kvaerner Engineering.

## Authors' addresses

*Knut Kvale, Narada Warakagoda & Marthin Kristiansen  
Telenor R&D  
NO-1331  
Fornebu  
Norway  
e-mail: knut.kvale@telenor.com  
narada-dilp.warakagoda@telenor.com  
marthin.kristiansen@akerkvaerner.com*

# WORD FINDING ABILITY AND BODY COMMUNICATION

*Ann-Christin Månsson*  
SSKKII Cognitive Science,  
University of Göteborg, Sweden

## **Abstract**

*In this paper, three groups of eight children each are studied; all the children are four to six years old. One group has language comprehension problems; in their production, semantics and the lexicon are primarily affected (SEM/LEX group). Another group has no language comprehension problems and their main production problems affect phonology and grammar (PH/GR group). These two specific language impairment groups are compared to a matched control group. One purpose of this study was to find out how word-finding problems interact with different categories of body communication (BC). BC that was significantly correlated with word-finding ability was of interest. The influence of specific activities—a naming task versus a spontaneous situation—is also discussed. A qualitative interpretation of the function of the BC used by the three groups of children when they had word-finding problems was done. Unlike the PH/GR group, children in the SEM/LEX group did not seek help with different body posture movements as much as they really needed to in the naming task. This is an interesting finding as it shows that the children in the SEM/LEX group may not be aware of their word-finding difficulties. This applies to all children in the SEM/LEX group: even at the age of 4, they did not seek help in learning new words. Clearly, this behavior could have a negative impact on their semantic development, as they have little opportunity to learn new words.*

**Keywords:** Specific language impairment, body communication, word-finding ability, semantic development.

## 1. Background

The term “nonverbal communication” is often used in the literature to denote something children have too much or too little of and may even be described as “normal” or “not normal.” In spite of the fact that little research has been done on what kind of “nonverbal communication” children with language impairments use, the term prevails as something that everyone thinks they know something about. Clinical workers may make such remarks as “He can’t sit still on a chair, he moves his body all the time, he must be hyperactive” or “He didn’t look at me during the whole session in the naming task; he avoids eye contact. Is that normal?” In our clinical experience, parents often ask whether they should pay no attention to the gestures or body communication their language-impaired children use when they want something and do not have the words to ask for it. These comments, among others, were the starting point for this thesis, as hardly any studies had been done on gestures and body communication. It is always helpful to fall back on serious research results when one is attempting to provide guidance for parents.

McNeill (1992) claims that speech and gesture must cooperate in order to express the person’s meaning. Speech and gesture are seen as completely different from body language, defined as a communication process made up of their body movements.

Iconic gestures and other body communication will both be included in the present study and the subgroup of iconic gestures will be analyzed separately. The purpose of studying iconic gestures is, first of all, to be able to compare the research in our study with other research that has been done on children’s gestures. Furthermore, one of the central purposes of the present study is to discover the semantic content of iconic gestures.

Body communication (BC) is the term used in this study for body posture movements, head movements, gaze direction, facial expressions and arm and hand movements (including iconic gestures). All of these movements have a communicative function, but the term “nonverbal communication” is not appropriate, mainly because they very often co-occur with speech. Even if body communication does not co-occur with speech, it still has a communicative function; often it actually has a word-like function.

One of the goals of this study is to find out whether younger children with normal language development or SLI children use more body movements and other kinds of BC.

Knapp and Hall (1992) state that body motion, or kinetic behavior, typically includes gestures, movements of the body (limbs, hands, head, feet and legs), facial expressions and eye behavior such as direction and length of gaze. Some gestures are speech-independent, like emblems, while others are speech-related; the latter are directly tied to or accompany speech, and serve to illustrate what is being said verbally. Posture is normally studied in conjunction with other body communication in order to determine the degree of attention or involvement, the degree of status relative to the other interaction partner, or the degree of liking for the other in the interaction. A forward-leaning posture, for example, has been associated with higher involvement and liking. Touching behavior are relics from earlier stages in life, when we were first learning how to manage our emotions, to develop social contacts, or to perform some task. The face is the primary locus of the expression of affect or emotion; the six primary affects are anger, sadness, surprise, happiness, fear and disgust. Facial gestures can also be used as regulatory gestures, providing feedback and managing the flow of interaction.

Allwood (1998) discusses features of flexibility and conflict prevention and how they are related to cooperation. The main focus in this study is on body communication from video-recorded and transcribed human-human dialog. Different gestures and their functions are described. The classification used for head movements includes, for example, nodding, tilting and shaking the head. Some functions of nodding the head that are mentioned include rejection of negative statements and reinforcing one's own turn. Shaking the head may function as a rejection or denial, both as a proper turn and overlapping with another person's turn. Facial expressions such as wrinkling the eyebrows may show lack of understanding of facts related, indicate that something is unpleasant in one's own utterance, express difficulty in finding the right word or indicate surprise at another's utterance. Raising the eyebrows may indicate surprise at another's utterance. Smiling functions, for example, to indicate insecurity, give confirmation, indicate friendliness, elicit confirmation, remove seriousness, remove the effects of one's own statement, apologize, remove danger, indicate humor, weaken opposition or even indicate that something is unpleasant. Gazing, such as gazing around at other interlocutors, may function to elicit confirmation and observe others' reactions. Gazing down

may break contact in silence, avoid confrontation and indicate insecurity. Gazing at the speaker shows attention to the speaker, and seeking eye contact elicits feedback. Laughing has, among other functions, that of releasing tension, both collective and individual. Among the various kinds of body posture, leaning forward shows interest, straightening the body indicates insecurity, and moving the shoulders indicates that something is to be taken as a rough estimate. Arm and hand movements include fidgeting with clothes, which expresses insecurity and releases tension, iconic illustrations whose function is to supplement content, baton gestures used for emphasis, and pointing which functions as symbolic or concrete deixis. These are just some of the functions mentioned by Allwood (1998).

Ekman and Friesen (1969, 1981) categorized body communication in general and found the following types: emblems, illustrators, affect displays, regulators and adaptors: 1) Emblems have a direct verbal translation, consisting of a word or two or perhaps a phrase, known by all members of the group, class or culture, and may repeat, substitute or contradict some part of the verbal behavior. 2) Illustrators are movements directly tied to speech, serving to illustrate what is being said verbally (repeat, substitute, contradict or augment the information provided verbally), which are socially learned. There are six types of illustrators: Batons are movements which time out, accent or emphasize a particular word or phrase; Ideographs gives the direction of an idea; Deictic movements points to an object in the environment; Spatial movements depict a spatial relationship.

“Spatial movement” is the term used when children show relations such as size and movements with their hands; it is not actually a good term since all hand gestures are spatial.

Kinetographs are movements which depict a bodily action; Pictographs are movements which draw a picture of their referent. 3) Affect displays have the face as the primary site (e.g. expressions of happiness, surprise, fear, sadness, anger, disgust and interest). 4) Regulators regulate the conversational flow and pacing of the exchange. The most common regulator is the head nod; another example is raising the eyebrows. 5) Adaptors are movements that were first learned as part of adaptive efforts to satisfy self-needs or to perform bodily actions to manage emotions. Self-adaptors involve touching one’s own face, hair, or lips to facilitate or inhibit sound production or speech. Hand-to-face adaptors are a rich source of information.

The terminology used for body communication in our study is mainly taken from Allwood (1998), with words to denote representational gestures from McNeill (1992) and Kita (1993, 2000), and emblems, spatial movements, affect display and self-adaptors from Ekman and Friesen (1969, 1981).

### *Research questions concerning the relation between the use of specific subcategories of BC and word-finding ability*

A purpose of the study is to provide further information about the specific subcategories of BC that are used in a situation where the children have word-finding problems and the interlocutor does not understand them. How do word-finding problems interact with different subcategories of BC? Do the kinds of BC used in different activities—in this study, the naming task and the spontaneous situation—differ? The possible meanings of the BC that is used due to the children's word-finding difficulties or communicative problems will be of interest. The main questions in this connection are:

1. What kinds of specific subcategories of BC are used more frequently by each group?
2. Do the children use BC to seek help from the interlocutor?
3. How do they seek help from the interlocutor?

## **2. Method**

### ***2.1 Subjects/Groups***

Twenty-four monolingual Swedish-speaking children were chosen for this investigation. The children were between 4;0 and 6;3 years of age; fifteen were boys and nine were girls. They were divided into three groups with eight subjects in each. One group had language comprehension problems; mainly semantics and lexicon were affected in their production (SEM/LEX group—mean age 5;3). The second group had no language comprehension problems and their main production problems affected phonology and grammar (PH/GR group—mean age 5;2). They were compared to a matched control group (NSP (no speech pathology) group) with a mean age of 5;2. (The children were included in the study after their parents and the ethical committee had given their consent.) Table 3.1 shows the age and sex of all children and whether there was any family history (FH) of dyslexia or Specific Language Impairment.

A number of criteria were used for the selection of the groups. The children were diagnosed by a speech/language pathologist as having SLI. It is taken into consideration that the children in the study are a heterogenous group. A psychologist tested the children with the WPPSI-R and/or Arthur Adaptation of the Leiter International Performance Scale (Arthur, 1952), and only children reported to have an IQ above 85 were included. None of the children had any mental handicap, autism, or overt neurological signs. All had normal hearing. The subjects had no dyspraxia. Language comprehension was tested in order to separate out the two SLI groups, using the Test for the Reception of Grammar (TROG; Bishop, 1998). Children who were 8 months behind the normal limits were diagnosed as having language comprehension problems. They were also tested with the SIT (Språkligt impressivt test; Hellqvist, 1989), which is used by speech pathologists in Sweden. SIT is a Swedish version of the Carrow-test. This test was designed to test the language comprehension. The SEM/LEX group scored 8 months behind the normal limits on this test as well. The grammar production was tested with The Lund Test of Phonology and Grammar (LUMAT) (Holmberg and Stenqvist, 1983). The section in the test with phonology was not used from that test. The Swedish Phoneme Test (Hellqvist, 1989) consists of confrontational naming of pictures and selection of the picture that a spoken word refers to. This test was designed to test children's production and comprehension of Swedish phonemic contrasts. The Phoneme test is not aimed for semantic correctness, but as it is used frequently in clinical practice, it may be used for semantic correctness as well as phonological ability. The children's responses to the naming task have been analyzed in terms of semantic and phonological correctness, in order to assess the nature of their word production problems. The children with mainly semantic-lexical problems (SEM/LEX group), were selected on the basis of 1) word-finding difficulties, i.e. a score of less than 82 (cf. Table 3.2) semantically correct word responses on the Phoneme Test (Hellqvist, 1991), and 2) difficulties explaining and understanding instructions, as reported by parents and preschool teachers. The children with mainly phonological and grammatical problems had more than 82 semantically correct word responses on the Phoneme Test (Hellqvist, 1991). Language comprehension was within the normal limits for all members of this group.

## 2.2 *Analysis of body communication*

The videotapes were examined several times. All the BC that occurred in the two different situations was observed. The BC was categorized according to a system influenced by different gesture researchers, e.g. Ahlsén (1985), Allwood (1998), Ekman and Friesen (1969, 1981), Gullberg (1998), Kendon (1972, 1980), Kita (1993) and McNeill (1992) (see Chapter 2). The BC was divided into five main categories: body posture movements, head movements, gazes, facial expressions and arm and hand movements. The neutral position in the naming task was that the child sat in an upright position, looking down at the pictures. The child's arms were on his or her lap. The neutral position was similar in the spontaneous situation, with the difference that the gaze was directed more generally forward and down. All changes from the neutral position were categorized and counted. The IL was face-to-face with the child. The children were seated in both situations. There were no other audience than the IL in the classroom in any of the situations. The five categories of BC were divided into subcategories for the movements and gestures that occurred. *Body posture movements*, e.g. leaning forward, moving to the side, straightening the body, retracting the body, moving the body forward and backward, getting down on the floor, moving back towards IL, lying down on the chair, jumping up from and sitting down on the chair, lying down on the table and shrugging the shoulders, were observed. Table 3.1 shows body posture movements compared to the classification in Allwood (1998), which in many respects was similar to the present study.

### 2.2.1 *Body posture movements*

Table 2.1. Classification of body posture movements in Allwood (1998) and the present study

<b>Allwood (1998)</b>	<b>This study</b>
Body posture	Body posture movements
Leans forward	Leans forward
	Moves body to side
Moving shoulders	Shrugs shoulders
	Retracts body
Raises body	Raises body
	Moves body forward and backward
	Gets down on floor
	Moves back towards IL
	Lies down on chair
	Gets up from and sits down on chair
	Lies down on table

*Head movements*, including nodding, shaking, tilting the head to one side, pushing the head forward, head down, moving the head down, jerking the head backward, turning the head towards IL and putting the head down, were observed as they occurred. Table 2.2 shows the classification of head movements in Allwood’s (1998) study and in this one.

### 2.2.2 *Head movements*

Table 2.2. Classification of head movements in Allwood (1998) and the present study

<b>Allwood (1998)</b>	<b>This study</b>
Head movements	Head movements
Tilts head to one side	Tilts head to one side
Pushing head forward	Head forward
Raises head	Raises head
Shakes head	Shakes head
	Head down
Nods head	Nods head
	Head down on arm
Jerks head backward	Jerks head backward
	Turns back of head towards IL
Rocks head	
Rocks head forward	

### 2.2.3 *Gaze direction*

The numbers given for gaze direction refers to the number of change in gaze direction, away from the neutral one.

Table 2.3. Classification of gaze direction in Allwood (1998) and the present study

<b>Allwood (1998)</b>	<b>This study</b>
Gaze	Gaze direction
Gazes at speaker	Gazes at interlocutor
	Gazes to side (not at IL)
	Gazes up (not at IL)
	Gazes straight forward (not at IL)
	Shuts eyes
Gazes at own hand gesturing	Gazes at own hand gesturing
Gazes down	Gazes down (further down than neutral gaze)
Gazes at manipulable artifacts	Gazes at manipulable artifacts
	Gazes away from IL
	Gazes at own hands
Gazes around at other interlocutors	
Seeking eye contact	

#### 2.2.4 Facial expressions

*Facial expressions* such as surprise, fear, disgust, anger, happiness, and sadness were observed (Ekman and Friesen, 1981), but the children were neutral apart from smiling and wrinkling their eyebrows. The only facial gesture (Allwood, 1998) observed in this study was wrinkling eyebrows, while raising eyebrows did not occur. Smiling and wrinkling eyebrows were classified as facial expressions. Laughter did not occur.

Table 2.4. Classification of facial expressions in Allwood (1998) and the present study

<b>Allwood (1998)</b>	<b>This study</b>
Facial gestures	Facial expressions
Smiles	Smiles
Wrinkles eyebrows	Wrinkles eyebrows
Raising eyebrows	

## 2.5 Arm and hand movements

*Arm and hand movements*, such as iconic illustrations (McNeill, 1992; Allwood, 1998), pictographs (Ekman and Friesen, 1981), beats (McNeill, 1992), spatial movements (Ekman and Friesen, 1981), handling artifacts, moving the hand forward and back, and crossing the arms (Allwood, 1998), pointing deictic movements (Ekman and Friesen, 1981), and the emblems of showing an amount with the fingers were also observed in this study. Self-adaptors (self-touching) (Ekman and Friesen, 1981) were observed and categorized under arm and hand movements

Table 2.5. Classification of arm and hand movements in Allwood (1998) and the present study

<b>Allwood (1998)</b>	<b>This study</b>
Arm and hand movements	Arm and hand movements
Fidgeting with clothes, hair	Self-adaptors
Pointing	Deictic, concrete
	Hand up
	Hand on table
	Stretching of the arms
Arms crossed	Arms crossed
Moving artifacts	Handles artifacts
Iconic illustrations	Iconic illustrations
Baton gesture	Beats
	Emblem, symbolic
	“Spatial movement”
	Hand to side
	Hand forward and back
Moving finger	

Hand up = the hand was raised with the palm towards the IL.

Hand on table = moves hand downwards with the palm down towards the table

Hand to side = the hand in a lateral position with an arc trajectory to the side

Hand forward and back = hand with an arc trajectory to both

## ***2.6. The relation between the subcategories of BC and word-finding ability***

The types of BC used in this study, i.e. body posture movements (BPM), head movements (HM), gaze direction, facial expressions and arm and hand movements, and the different subcategories were counted. The observations are from the three groups of children in the two situations. Word-finding ability depended on how the children performed on the naming task (SBP; Apt, 1994), i.e. the number of semantically correct word responses. Correlation analyses were done to assess the importance of each variable of the BC together with the variables word finding in the naming task. The same variables were used in both situations to see if there was a correlation effect. Potential correlation effects between different variables were tested.

## **3. Results**

### ***3.1. Body posture movements***

Word-finding problems caused changes in the use of BPM. In general, the **SEM/LEX** group made more BPM in the naming task, whereas the **PH/GR group** used more in the spontaneous situation. The naming task was difficult for the SEM/LEX group due to their apparent word-finding problems. The spontaneous situation was troublesome for the PH/GR group as they had primary problems with phonology but also word-finding problems, and might therefore tend to avoid communicating. The SEM/LEX group could fall back on their phonological competence to a greater extent than the PH/GR group in the spontaneous situation. The **PH/GR group** used leaning forward more than the NSP and SEM/LEX groups in the naming task. In the spontaneous situation, both **SLI groups** tended to move the body to the side more than the NSP group.

For the **PH/GR group**, word-finding problems were related to moving the body to the side in the naming task and retracting the body in the spontaneous situation.

### ***3.2. Head movements***

Word-finding problems were found to be correlated with tilting the head to one side for the **PH/GR group** in the naming task and for the **NSP group**

in the spontaneous situation. Word-finding problems were related to raising the head for the PH/GR group in the spontaneous situation. Word-finding problems were related to moving the head down for the **SEM/LEX group** in the naming task.

### **3.3. Gaze direction**

Word-finding problems were related to gazing to the side for the **PH/GR group** in both situations. Age was related to gazing to the side for the **SEM/LEX group** in the *naming task*. Word-finding problems were also related to gazing down and gazing at IL for the **PH/GR group** in the spontaneous situation. Word-finding problems were related to gazing at own hands gesturing for the SEM/LEX group in the naming task.

### **3.4. Facial expressions**

There were no significant differences between these groups' use of facial expressions in the *naming task* and in the *spontaneous situation*,

### **3.5. Arm and hand movements**

The **SEM/LEX group** used iconic gestures in the naming task when they had less semantically correct word responses in the naming task; “spatial movements” in the spontaneous situation were also used because of less semantically correct word responses in the naming task. The **PH/GR group** did not use arm and hand movements to cope with word-finding problems except for the younger children's showing numbers with their fingers (emblems) in the spontaneous situation.

## **4. Discussion**

Word-finding problems and phonological problems caused body posture movements to be used in different ways. In general, the SEM/LEX group used more body posture movements in the naming task, while the PH/GR group produced more body posture movements in the spontaneous situation. As we have seen, the naming task was more difficult for the former group, while the spontaneous situation was more problematic for the latter. In the spontaneous situation, the SEM/LEX group could fall back on their phonological competence to a greater extent than the PH/GR group.

Different body posture movements have different functions and thus may be more likely to be used when the children have difficulties with word finding or with phonology. For the PH/GR group, word-finding problems were related to moving the body to the side and retracting the body, possibly revealing the children's desire to escape from the situation; phonological problems were related to moving the body forward and backward in the naming task and "out of focus" movements in the spontaneous situation.

The SEM/LEX group had the most problems with word finding; nevertheless, they did not use as many different types and subcategories of BC as the PH/GR group. Word-finding problems were related to the total number of body posture movements in the naming task for the SEM/LEX group. The naming task was the most demanding for these children, since it exposed their naming difficulties. The increased use of body posture movements can be seen as compensating for their word-finding difficulties.

Professionals may also misunderstand the extensive use of body posture movements as a sign of hyperactivity. Diagnosing these children as hyperactive may be neither appropriate nor good for their subsequent language development, since their use of body posture movements may be a sign that they want to participate in communication. Unlike the PH/GR group, children in the SEM/LEX group did not seek help with different body posture movements as much as they really needed to in the naming task. This is an interesting finding as it shows that the children in the SEM/LEX group may not be aware of their word-finding difficulties. This applies to all children in the SEM/LEX group: even at the age of 4, they did not seek help in learning new words. Clearly, this behavior could have a negative impact on their semantic development, as they have little opportunity to learn new words.

Allwood (1998) mentions that leaning forward functions to show interest. In our study, forward-leaning movements were used more to appeal for help and elicit feedback by the SLI groups, and significantly more by the PH/GR group, compared to the NSP group. Knapp and Hall (1992) mention the possibility that leaning forward may be associated with greater involvement and liking. Leaning forward was, however, not found to be related to age, word-finding ability or phonological problems in this study.

Word-finding problems were related to tilting the head to one side for the PH/GR group in the *naming task*. The same phenomenon was true of the

NSP group in the *spontaneous situation*. The PH/GR group appeared to be more confident than the SEM/LEX group about seeking help from IL in the naming task. For the SEM/LEX group, word-finding problems were related to moving the head down in the naming task; this movement may indicate that the children were seeking clues from the picture when they had word-finding problems. Head movements have a communicative function; as in Allwood's (1998) study, they are used to convey mutual agreement and to give and elicit mutual support.

For the PH/GR group, word-finding problems were related to gazing to the side in the *naming task*, and to gazing at IL, gazing to the side and gazing down in the *spontaneous situation*. The PH/GR group sought help from IL and actively showed interest more often than the SEM/LEX group.

Feedback was elicited through raising the head and smiling, as in Allwood's (1998) study.

The SEM/LEX group was the only one that had a negative correlation between word-finding ability and iconic gestures; the worse the problems the children had with word finding, the more iconic gestures they used. There was also a negative correlation between word-finding ability and "spatial movements," suggesting that both iconic gestures and "spatial movements" are closely linked to conceptual development.

## References

- Ahlsén, E. (1985). *Discourse patterns in aphasia*. Gothenburg Monographs in Linguistics 5. Department of Linguistics, University of Göteborg, Göteborg, Sweden.
- Allwood, J. (1998). Cooperation and flexibility in multimodal communication. Unpublished manuscript, Department of Linguistics, University of Göteborg, Göteborg, Sweden.
- Apt, P. (1994). *SBP Svensk Benämningssprövning*. Standardized revision based on Naming. Escape. EU project. Malmö, Sweden
- Arthur, G. (1952). *Arthur adaptation of the Leiter International Performance Scale*. Chicago: The Psychological Services Center Press.
- Bishop, D. V. M. (1983). *Test of reception of grammar*. Published by the author and available from Age and Cognitive Performance Research Centre, University of Manchester.

- Bishop, D. V. M. (1989). *Test for the reception of grammar, manual* (2nd edition). [Swedish translation. Holmberg, E., & Lundälv, E. (1998). Göteborg, Sweden: SIH Läromedel] University of Manchester, Age and Cognitive Performance Research Centre.
- Knapp, M., & Hall, J. (1992). *Nonverbal communication in human interaction*. New York: Harcourt Brace Jovanovich College Publishers.
- Ekman, P., & Friesen, W. B. (1969). The repertoire of nonverbal behaviour: Categories, origins, usage and coding. *Semiotica, 1*, 49–98.
- Ekman, P., & Friesen, W. B. (1981). The repertoire of nonverbal behaviour: Categories, origins, usage, and coding. In A. Kendon (Ed.), *Nonverbal communication, interaction and gesture* (49–58). The Hague, Paris: Mouton Publishers.
- Gullberg, M. (1998). *Gesture as communication strategy in second language discourse: A study of learners of French and Swedish*. Lund: Lund University Press.
- Hellqvist, B. (1989). *Nya SIT- Språkligt Impressivt Test för Barn*: [Language Comprehension test for Children.] Löddeköpinge: Pedagogisk Design.
- Hellqvist, B. (1991). *Fonem testet. The Swedish Test of Phonemes*. Löddeköpinge: Pedagogisk Design.
- Holmberg, E. & Stenqvist, H. (1983): *Nya Lundamaterialet. Kartläggning och bedömning av barns språkliga förmåga*. [The Lund test of Phonology and Grammar. Description and assessment of children's linguistic abilities.] Malmö: utbildningsproduktion AB.
- Kendon, A. (1972). Some relationships between body motion and speech. An analysis of an example. In A. Siegman, and B. Pope (Eds.), *Studies in dyadic communication* (177–210). New York, Pergamon Press.
- Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In M. R. Key (Ed.), *The relation between verbal and nonverbal communication* (207–227). The Hague: Mouton.
- McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press.
- Kita, S. (1993). Language and thought interface: A study of spontaneous gestures and Japanese mimetics. Unpublished Ph.D. dissertation, University of Chicago, Chicago.
- Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.), *Language and gesture. Language, culture and cognition* (162–186). Cambridge: Cambridge University Press.

## Biography

**Ann-Christin Månsson**, PhD in Neurolinguistics at the Linguistic Department, Göteborg University, Sweden. The title of the thesis was "The relation between gestures and semantic processes: A study of normal language development and specific language impairment in children". She examined the body communication (BC) used by the three groups of children when they had phonological and word-finding problems. She also developed a method for observing semantic development revealed by children's speech and hand gestures.

Ann-Christin has a background as a speech and language pathologist. She is currently working on a project entitled "The internal content and function of children's gestures" supported by FAS (grant nr 2004-0143).

## Author's address

*Ann-Christin Månsson  
SSKKII Cognitive Science  
Dept. of Linguistics  
Göteborg University  
Box 200  
SE 405 30 Göteborg  
Sweden  
e-mail: anki@ling.gu.se*

# ARTIFICIAL GAZE PERCEPTION EXPERIMENT OF EYE GAZE IN SYNTHETIC FACES

*Gunilla Svanfeldt, Preben Wik and Mikael Nordenberg*

Department of Speech, Music and Hearing  
KTH, Stockholm, Sweden

## **Abstract**

*The aim of this study is to investigate people's sensitivity to directional eye gaze, with the long-term goal of improving the naturalness of animated agents. Previous research within psychology have proven the importance of the gaze in social interactions, and should therefore be vital to implement in virtual agents. In order to test whether we have the appropriate parameters needed to correctly control gaze in the talking head, and to evaluate users' sensitivity to these parameters, a perception experiment was performed. The results show that it is possible to achieve a state where the subjects perceive that the agent looks them in the eyes, although it did not always occur when we had expected.*

**Keywords:** Perception experiment, Talking heads, MPEG-4, Eye gaze, Embodied Conversational Agents

## **1. Introduction**

Today the default metaphor used in human-computer interaction (HCI) is the desktop, where the computer is compared with a desk containing a desktop, drawers and file holders. Once animated agents perform satisfactorily well, an important shift of the metaphor used in HCI may occur, using instead a person metaphor. Before this can take place a number of unresolved issues must first be addressed.

Well-synchronised speech-articulation in the animated agent is a necessity for a natural communication, and has been provided through the work of Jonas Beskow (Beskow, 2003). The lip-reading support that the articulating synthetic faces can provide in noisy environments or to hearing impaired is well established (Beskow et al., 1997, Agelfors et al., 1998). The next step is to equip the agent with the capacity of being more expressive by means of adjusted articulation and other facial movements, so that e.g. turn-taking signals and attitudes can be conveyed. The importance of the eye gaze is also necessary to consider. So far the eyes in the synthetic faces have been more or less neglected, partly because of lack of control facilities, but with the new MPEG-4 model that will be described below, the possibility of tailoring the eye gaze behaviour provides new opportunities to use this channel for information.

The aim of this paper is to investigate how to introduce more natural-like eye behaviour in animated agents. In order to test whether we have the appropriate parameters available when adjusting the eye gaze in the synthetic face, and to evaluate the users' sensibility to these parameters, an experiment has been performed.

## 2. Eye Gaze

Eye gaze has been studied quite extensively in human-human interaction within psychology. It has also gained attention in the area of animated characters. However, despite the discussion about the function and importance of gaze, it is rarely properly implemented. For example, in (Cassell et al., 1994) a rather detailed description of four categories of gaze is given, but they do not differentiate between head and eye movements in the implementation of the system.

One of the goals in the development of a virtual agent is to make it natural to interact with, besides other important issues such as good articulation and relevant acoustic output. In order to achieve effective interaction, it is necessary for the user to feel that the agent is interested in and focussed on the conversational exchange, and that it displays relevant signals for intentions, understanding and turn-taking. A prominent visual cue for this, except head or eyebrow movements, is the eye gaze.

According to Argyle & Cook (1976), speech related gazes have three main functions:

to send social signals

to open a channel to receive information

to control the synchronisation of speech

As to the first function, certain rules about the amount of gaze apply to different situations. If these rules are broken, people are likely to be offended, or at least confused. The amount of gaze transmits impressions of the temporary or permanent state of the user. Kleck and Nuesle (1968) found that persons (on film) with only 15% of their gaze directed to the conversation partner were perceived as cold, pessimistic, and defensive, while those with 80% gaze were considered friendly, self-confident and sincere. Argyle (1988) reports how people with higher levels of gaze are seen as more attentive, and that lack of eye contact indicates passiveness or inattentiveness. It is however not only a matter of total amount of gaze. There are norms concerning how long a glance should last, and the amount of mutual gaze also depends on the distance between the two persons, since the gaze can be seen as a regulator of the intimacy. Another determining factor is the sex of the two conversational parts – women tend to have less eye contact with men than with other women.

If the agent is not capable of acting according to the social rules, it will not be considered trustworthy in the users' eyes or might induce unwanted reactions of the user. In a study by Park Lee *et al.* (2002), a gaze tracking device is used to acquire data. After data processing and implementation in a synthetic face, a comparison of three different types of eye movements is performed. It showed that with no eye movements at all, the character was perceived as lifeless, but with statistically based eye movements the face character looked more natural and friendly. With random eye movements, the quality of the character was unstable.

In a study performed by Garau *et al.* (2001), similar results were found when comparing dyadic (i.e. two participants) conversations. There were four different conditions: video, audio-only, and two avatar conditions, where the avatar's head and eye movements were either randomly induced or based on research on face-to-face dyadic conversations. The video condition got the highest overall scores, and the inferred-gaze condition got better scores than the random case regarding the similarity to real face-to-face conversation, involvement, co-presence and partner evaluation.

The second function – to open a channel to receive information – is not implemented in the talking head used for the study in this report. The agent

cannot receive any visual information, although it hopefully will in the future. It may be discussed whether it still should simulate this ability or not in order to keep a fluent interaction with the user. However, if the agent signals that it can read the users' gestures, and then do not react to these, it might confuse the user. It is well known that a good interface should be clear with regards to what it can and what it cannot perform.

The possibility of controlling the synchronisation of speech, which is the third function of speech related gaze, is very appealing to explore with a talking head. Even though the signalling will only be in one direction, it may still lead to important improvements of the interaction with the agent. If the animated agent is able to signal turn-taking, there is likely to be less uncertainty and interrupts in the conversation. It will also mean less cognitive load for the user if the basic signals for conversation regulation are employed.

### **3. The talking head**

Animated talking heads capable of producing lip-synchronised speech have been developed at CTT (Beskow, 2003). The acoustic speech can be either synthetic or natural, and the model can also convey extra-linguistic signs such as frowning, nodding, and eyebrow movements.

To gain knowledge about how to animate the agents in terms of verbal and non-verbal behaviour, 3D facial data collected by means of an optical motion tracking system from Qualisys<sup>2</sup> was used. Reflective markers attached to the speaker's face were registered with infrared cameras and the system provided the 3D coordinates of those markers.

A new generation of talking heads are currently being developed at CTT using the MPEG-4 standard. MPEG-4 is known for being a high compression standard for coding audio and video, but the MPEG-4 (Version 2) standard also describe channels for face and body animation in a very low bitrate coding. The standard defines 66 low-level facial animation parameters (FAPs) that describe the animation of a face model (Ostermann , 2002).

Distances in the MPEG-4 models are, as in many 3D models, not described in metrics, but in units. This is because the size of the model does not correspond to anything absolute in the physical world. There are however relative distances between different parts of the model that need to be

described. A generic face has been modelled and measured and a set of standard sizes has been given. Deviations from this can then also be described with ease. For example, in the MPEG-4 face used in this experiment, the distance between the mouth and the tip of the nose is 5.3 units, the distance between the eyes is 12 units, and the diameter of the iris is 2.1 units.

In order to regulate the gaze of the synthetic face, the distance to the virtual viewpoint (or camera analogy) is used as basis for the eye positioning. The idea is to have the agent look into the camera, just as a person on a photograph seems to look at the observer, if the person was looking into the camera when the picture was taken. Having the virtual viewpoint as baseline, the point of focus can then be varied in a three dimensional space.

In the manual texture mapping of the synthetic face, textures risk being asymmetrically set. This means that even though the gaze is calculated depending on the camera, some skewness in the model may damage the effect of eye gaze. The parameters that can be manipulated for eye gaze control are thus rotation of the eyeballs in two directions, and the combined set-up of the eyes' direction will yield the focus point in space.

## **4. The experiment**

### ***4.1 Aim of the experiment***

The primary aim of the experiment was to evaluate if subjects – for any of the conditions – perceive that the agent looked them in the eyes. If so, to what parameters were the subjects sensitive: how much displacement of the eyes was needed for loosing the gaze, and was the sensitivity different depending on dimension or position of the head? This knowledge is crucial for future studies and implementation of eye movements.

### ***4.2 Method***

There were 15 participants in the experiment, 8 women and 7 men. The stimuli that were<sup>2</sup> presented to the subjects consisted of static pictures with different eye gaze and headposition. The eye gaze was varied in three dimensions – laterally (x), vertically (y) and in the depth (z) dimension. As described earlier, the face model is defined in units rather than ordinary metrics. In the x-and y-dimensions, the total variation for the focus point

---

<sup>2</sup> <http://www.qualisys.se>

was 40 units (which corresponds to an angle of approximately  $20^\circ$ ), and each step was 5 units. In the z-dimension, the steps were not linear. For focus points between the agent and viewpoint, the steps were 30 units. For distances beyond the observer increasingly larger steps were taken. The variation in the three dimensions is illustrated in figure 1. Due to the lack of previous studies of this kind, the steps and limits were chosen to cover a reasonable range of angles.

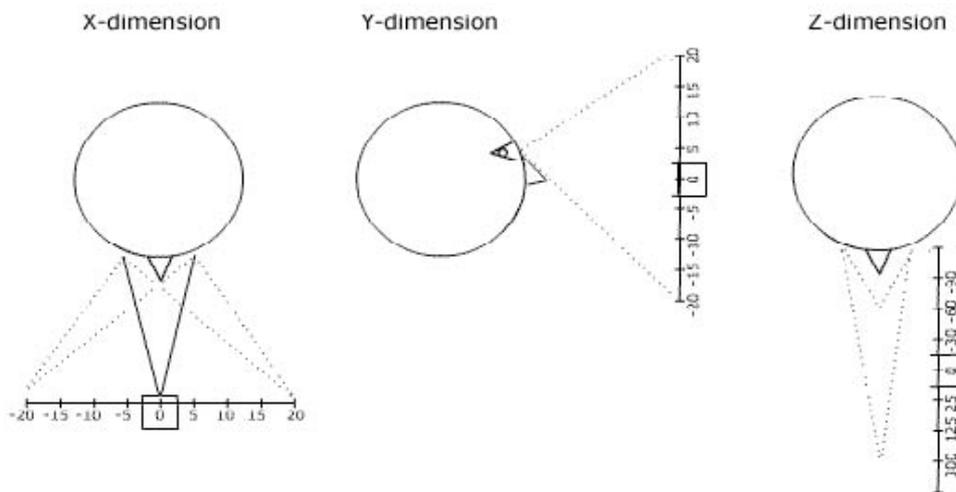


Figure 1. Illustration of the three dimensions in which the gaze point varies.

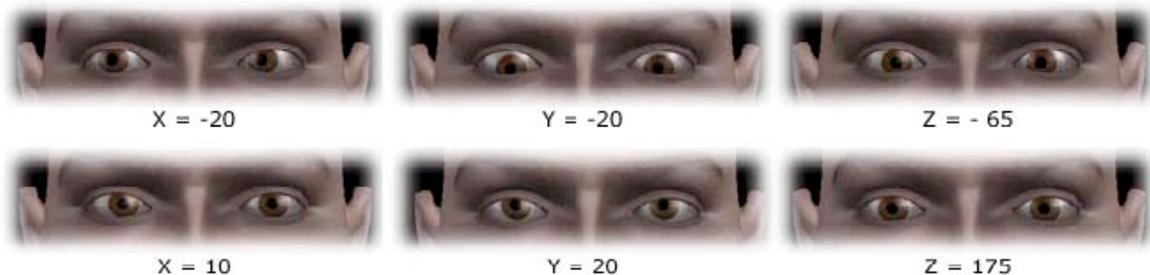
Two different head positions were evaluated, one front view, and one with the head turned to the side (see figure 2). The idea was to find a head position that typically could occur in a dialogue situation, yet large enough to make a perceivable difference. We found that an angle of approximately  $11^\circ$  to be suitable. The two different head positions, together with 23 eye gaze variations gave 46 different conditions. Each condition was presented to the subjects twice, in randomised order. Six additional pictures – the same for all the subjects – were inserted in the beginning of the test as dummies and were removed before the analysis. In total 98 stimuli were presented to each subject. Some examples of eye gaze are shown in Figure 3.

The introduction to the test was presented by another talking head with synthetic acoustic speech, where the aim of the experiment was explained, and instructions to the test were given. For each stimulus, the subject was asked to answer yes or no to the question “Is this man looking you in the eyes?”.



*Figure 2. The two head positions used in the experiment. Both have eye gaze that according to calculations should look straight into the viewpoint, and thus look the observer in the eyes.*

After the self-paced sequence of 98 stimuli, four additional pictures were shown where the subject was asked to more qualitatively describe where the agent was focusing its gaze. The subject was also asked about any difference in difficulty of determining the gaze direction of the agent when the face was in the frontal view as compared to the side view.



*Figure 3. Examples of variation in gaze direction in each dimension.*

Finally, the subject was invited to give his or her opinion to what the most prominent defect of the synthetic face was. The aim of this last question was to give us an idea of what possible distractions there might have been during the test, besides getting a hint of what is most urgent to improve.

### **4.3 Results**

The results show that it is possible to produce eye gaze in the synthetic face that observers think meets their gaze. However, the phenomenon did not always occur when we had expected. Some subjects were also more

permissive when judging the gaze than others, which can be seen in figure 4, where the total amount of positive answers is displayed.

The front view obtained more positive responses than the side view, as seen in figure 5. It received almost twice as many “yes”-answers as the side view, and this trend remained for all three dimensions in space, see figure 6, 7 and 8.

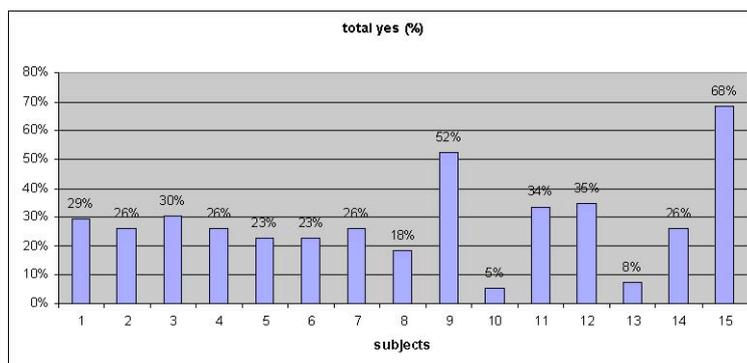


Figure 4. The graph shows the total percent of positive answers for each subject in the experiment. The question they responded to was “Does this man look you in the eyes?”.

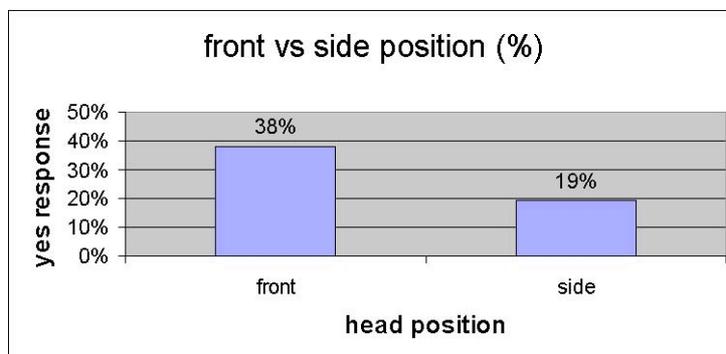


Figure 5. The diagram shows, for all subjects, how many pictures (in percent) that were perceived as looking the subjects in the eyes, according to viewpoint (front and side view).

Both the x-and y-dimensions show that there is an asymmetry in the responses. The positive responses are not centred on 0 (corresponding to the virtual viewpoint), which would have been expected. This trend is more striking for the side view than for the front view. In the x-dimension, the 10 unit displacement got the highest score, and in the y-direction it was the 15 unit case that obtained the most positive answers.

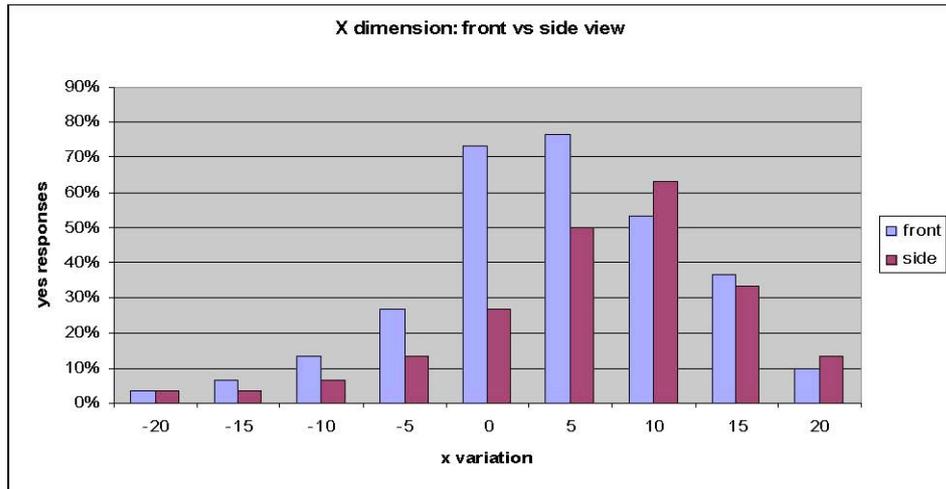


Figure 6 . Illustration of the distribution of responses judging that the agent looked the subjects in the eyes. The steps on the x-axis are in units, the total range from -20 to 20 corresponds to an angle of  $20j$ .

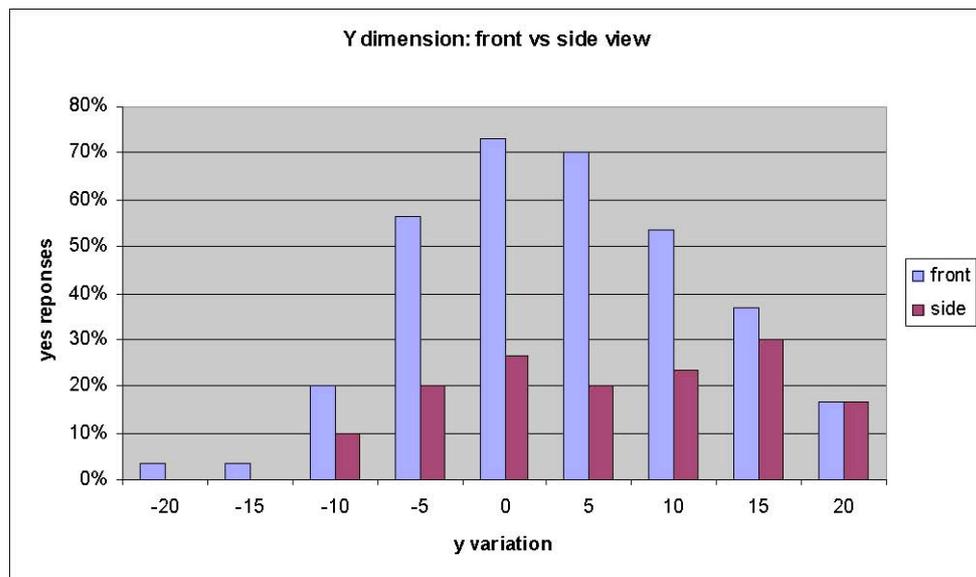


Figure 7. Illustration of the distribution of responses judging that the agent looked the subjects in the eyes. The steps on the x-axis are in units, the total range from -20 to 20 corresponds to an angle of  $20j$ .



Figure 8. Illustration of the distribution of responses confirming that the agent looked the subjects in the eyes. The steps on the x-axis are in units.

By the shape of the graphs in figure 6, 7 and 8, it can also be stated that the subjects were less sensitive to changes in the depth dimension than in the other two dimensions. The sensitivity also diminished when the head was turned to the side, especially in the y-dimension.

The result of the four pictures that were shown after the self-paced test was interesting in that a rather wide range of answers were given to where the agent focussed its gaze. The first picture had the focus point at 65 units in front of the virtual viewpoint, and thereby meant to be perceived as in between the screen and the subject. Out of the subjects, there were 7 who reported that this was the case, 3 subjects remarked that the agent looked to the left of the subject (which was not intended), another 3 thought the focus was on their chin or nose, and to one subject the agent seemed absent-minded or just unfocused. One subject considered the agent to look him in the eyes.

In the second picture, the agent was supposed to look beyond the subject, so the focus point was set at 175 units behind the viewpoint. As few as 3 subjects said the agent was fixating a point behind them, and another 3 thought the gaze was unfocused. Although there was no such intention, 6 subjects believed the agent looked to the left of the subject, in some cases in combination with behind. 2 subjects thought the agent was looking them in the eyes, and perhaps most surprisingly – since the aim was the opposite – 2 subjects thought the fixation point was in front of the subject.

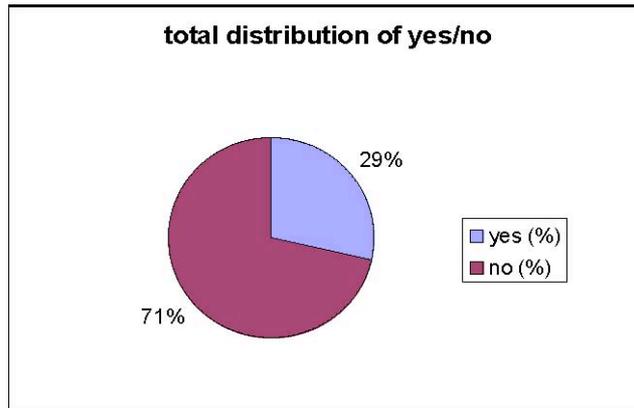
The opinion about the third picture was more unanimous. The fixation point was set 5 units below the virtual viewpoint, and 11 of the 15 subjects agreed. 6 of the subjects thought that the gaze focus was in between the screen and themselves. However, this may be a sequence effect, since the preceding picture had a fixation point beyond the subject.

Finally, the last of the four pictures had the focus point set 10 units (approximately 5 degrees) to the side of the virtual viewpoint, which would produce a gaze direction to the right of the subject. This was also reported by 8 of the subjects, and 6 of the subjects said that the agent had the gaze focus behind them. 8 subjects also reported that the gaze was above their own eyes, and curiously enough 2 subjects still thought the agent looked to the left of them (despite that the opposite was intended).

The difference in results between the front view and the side view could be supported by the outcome of the question the subjects were asked about whether there was any difference in difficulty in determining the eye gaze between those two conditions. The front view got more yes-answers, while the subjects reported that it was easier to discriminate different fixation points in the side view. Many of the subjects reported that the variations in gaze direction in the front view were vague. They felt more certain about whether the agent looked them in the eyes when the face was turned away. The reason for this may be that the position of the iris in the eye becomes more easily determined.

## **5. Discussion**

There is a problem with having yes/no-questions on a material that is not equally distributed in that sense. The subject may subconsciously strive to get a 50/50 distribution of their responses. (According to the implementation strategy of the agent only 4% of the stimuli should have a yes-response, or 30% if broadening the categories one step). Therefore, it is likely that the subjects have been too permissive in their judgement. The overall distribution of yes-and no-responses is shown in figure 9.



*Figure 9. Total distribution of yes-and no-responses for all subjects and all conditions.*

A common reaction to the cases where the focus point of eyes was set beyond the virtual viewpoint was that the agent looked absent-minded, not specifically focussing at anything else. It seems that the eye gaze behaviour is interpreted as in a dialogue context, rather than a specific estimate of a focus point in the three-dimensional space. It was considered especially confusing when the agent focussed somewhere in between itself and the virtual viewpoint (interpreted as between the screen and the subject), since there was nothing to focus on there – it is highly unlikely that someone will fixate a point in the air (unless a spider is hanging there).

The striking asymmetry in the x-and y-direction has several possible explanations. One is that the illumination of the agent was stronger from one side, so the shadowing might have influenced the perception of the gaze.

Another factor that may have contributed to the asymmetry is the manual mapping of texture. When carefully studying the face it can be noticed that one of the eyes (iris and pupil) is larger than the other, which is a texture mapping defect. A combined problem was that the larger iris and pupil was on the brighter side, which might have enhanced the problem. Normally the pupil gets smaller when the light increases, so the effect is likely to be confusing for the observer.

Notable is that some subjects reported a unconscious tendency to look only at one of the agent's eyes to begin with, and when noticing this and changed strategy, found that they were more unsure of the direction of the gaze. The two eyes were thus not consistent, it was like if the agent was squinting (strabismus). This may also be a result of the texture mapping problem.

Concerning the asymmetry in the y-dimension, the problem may be in the design of the eyes. Compared to photos of real eyes, it can be stated that the synthetic eyes show more of the iris than real eyes tend to do, and also more of the whites (see figure 10). Either the iris should be larger, or the eyes should be more closed. Probably the latter, or maybe a combination.



*Figure 10. Above the authors' eyes are shown as examples of real eyes. Below the default setting of the agents eyes. The proportion differences in how much iris that is shown, and how much of the whites that are visible illustrate possible clues to the asymmetry in results in the y-direction.*

## **6. Conclusions and future work**

The test results showed that we managed to produce eye gazes where the subjects perceived that the agent was looking them in the eyes, but in some cases this happened when we thought it would not, according to the calculations of eye gaze in relation to the virtual viewpoint. The subjects were less sensitive to changes in the depth dimension than in the other two dimensions. The sensitivity also diminished when the head was turned to the side.

It is worth to stress the interesting fact that the scores were not very high despite the fact that the set-ups were done manually. This highlights the need for more studies, to thoroughly investigate how to manipulate the parameters that we have in our use in order to control the eye gaze. It is possible that some adjustments of the face model are needed to ensure that not small mistakes during texture mapping or illumination risk to disturb the obviously very fine tuning that is required for eye gaze.

It is also possible that the perfect focus point is somewhere else, since we did not test combinations of the three dimensions. In a future study it would be interesting to narrow the range of each dimension and instead allow for

combinations as well as smaller steps. Another approach would be to perform a production experiment, where the subjects adjust the eye parameters into a position where they perceive that mutual gaze is achieved. That would yield more detailed information about the sensitivity of the subjects in this respect.

Another approach that would be interesting to combine with the method described above, is to – instead of just answer yes or no – mark on a scale where the subject perceives that the fixation point is.

In the experiment in this report, static pictures were used, but for obvious reasons, animated sequences are of high interest to us. As soon as eye gaze in static pictures can be achieved, producing realistic eye movements will be the next step. One challenge is the collection of data in order to accomplish natural and trustworthy eye gaze behaviour. But before that kind of implementation can be meaningful, we must learn how to control the eyes, and how different eye gazes are perceived by the users. When introducing eye movements, there are other aspects that become increasingly important, such as the use of the muscles surrounding the eyes, blinks, and other facial movements as well as head movements.

As with the rest of the facial movements in the talking head, it is desirable to have data driven methods for the eye gaze control. That means we have to collect data that is appropriate for a data driven animation method. The system that will be used for data collection, is the Tobii system<sup>3</sup>. The Tobii system uses video images of the person's face in combination with infrared light in order to track the 3D position of each eye, and to determine the target that each eye gaze is directed towards. This will permit studies of eye gaze behaviour for turn-taking signals and other communicative characteristics.

## **Acknowledgements**

We wish to thank all the participants of the perception experiment. This research was carried out at the Centre for Speech Technology, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations.

## References

- Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K-E., Öhman, T. (1998): Synthetic faces as a lipreading support. In *Proceedings of ICSLP'98*.
- Argyle, M. (1988). *Bodily Communication*. New York: Methuen & Company.
- Argyle, M. and Cook, M. (1976). *Gaze and Mutual Gaze*. Cambridge University Press, Cambridge.
- Beskow, J. (2003). Talking heads – models and applications for multimodal speech synthesis. *Ph.D. dissertation*, KTH, Stockholm, Sweden, 2003.
- Beskow, J., Cerrato, L., Granström, B., House, D., Nordenberg, M., Nordstrand, M., Svanfeldt, G. (2004). Expressive Animated Agents for Affective Dialogue Systems. In *Proceedings of<sup>3</sup> ADS'04*.
- Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K-E., Öhman, T. (1997). The Teleface project - Multimodal Speech Communication for the Hearing Impaired. In *Proceedings of Eurospeech '97*, Rhodos, Greece.
- Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., and Stone, M. (1994). Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In *Proceedings of ACM SIGGRAPH '94*
- Garau, M., Slater, M., Bee, S., Sasse, M.A. (2001). The Impact of Eye Gaze on Communication Using Humanoid Avatars. *CHI 2001*. Vol.3, Issue No.1.
- Kleck, R.E., and Nuessle, W. (1968). Congruence between the indicative and communicative functions of eye-contact in interpersonal relations. *Brit. J. Soc. Clin. Psychol.*, 7, 241-6.
- Ostermann, J. (2002). Face Animation in MPEG-4. In Pandzic, I. S. and Forchheimer, R. (Eds.) *MPEG-4 Facial Animation - the Standard, Implementation and Applications*, Chichester, England: John Wiley & Sons. pp. 17-56.
- Park Lee, S., Badler, J. B., Badler, N. I. (2002). Eyes Alive, *ACM Transactions Graphics* 21 (3), July 2002, ACM:NY, 637-644.

---

<sup>3</sup> <http://www.tobii.se/>

## Biography

**Preben Wik** has his academic education from the University of Oslo, Norway, and is currently a PhD student at the Department of Speech, Music and Hearing, KTH.

**Gunilla Svanfeldt** has a Master of Engineering from KTH, and is currently a PhD student at the Department of Speech, Music and Hearing, KTH.

**Mikael Nordenberg** has a Master of Engineering from KTH and is currently working for NetWise (+46 8 337400)

### Authors' address:

*Gunilla Svanfeldt and Preben Wik  
Department of Speech, Music and Hearing  
KTH  
Lindstedtsv.24  
100 44 Stockholm  
Sweden  
e-mail: [gunilla@speech.kth.se](mailto:gunilla@speech.kth.se)  
[preben@speech.kth.se](mailto:preben@speech.kth.se)  
[mikaeln@speech.kth.se](mailto:mikaeln@speech.kth.se)*

# MULTIMODALITY AND DESIGNS FOR LEARNING

*Anna-Lena Rostvall & Tore West*  
Stockholm Institute of Education, Sweden

## **Abstract**

*This paper brings forward an example of gestural, musical and verbal shaping of the process of breathing, as illustrated in video recorded music education. Neither speech, music nor actions seem coherent when viewed independently from one another. Each mode has its own semiotics and the meanings of modes are intertwined and contributes in cooperation to how instructions can be understood. Teaching is a complex social phenomenon with a long history that could be understood in a multitude of ways. This is something that poses several challenges to any study of teaching and learning. Video recordings of interaction in a classroom setting is one way to capture what actually goes on in the tuition, rather than what the participants can say about their practice. Video recordings create very large amounts of multimodal data, and a multimodal analytical approach makes even the most ordinary classroom interaction appear enormously complex. One reason for the relatively small number of empirical studies in this area could be the theoretical and methodological difficulties involved in handling and analysing such rich data. An ongoing Swedish study analyses the interface between language, music and gesture as semiotic resources, rhetorically orchestrated in the classroom. A study of modalities of communication is combined with study of interpretative practices of teachers and students. Several generally held views on education are challenged through such a multimodal analysis; that teaching and learning are mainly applied practical accomplishments, that education is not in need of rhetoric, and that learning is an acquisitive process in which students acquire information from the teacher and from method books. Learning is instead suggested to be seen as a dynamic process of transformative sign-making which actively involves both teacher and*

*students, and where an expansion of the access to diverse semiotic resources increase the repertoire of possible thoughts and actions within the field.*

**Keywords:** Education, interaction, video analysis, learning, multimodality, social semiotics.

## 1. Introduction

Teaching is a complex social phenomenon with a long history that could be understood in a multitude of ways. This is something that poses several challenges to any study of teaching and learning. Video recordings of interaction in a classroom setting is one way to capture what actually goes on in the tuition. Video recordings create very large amounts of multimodal data, and a multimodal analytical approach makes even the most ordinary classroom interaction appear enormously complex.

All systems for analysis are more or less impregnated with assumptions and theories since they bring different data to the front. Research on interaction in education has been criticized for building on a simple communication model of sender-message-addressee (Shannon and Weaver, 1949) that does not describe the actual functions inherent to complex communication (Eco, 1984; Fairclough, 1995). Research based on this model stresses a focus on individual descriptive goals that reveal little understanding for the institutional and societal context, which influences the interaction on the local level. The model of initiation-response-feedback proposed by Sinclair and Coulthard (1975) is typical for the critique of the previous model. Another example are the categories of Bellack et. al. (1966) that locks teacher and students into fixed roles and leave little room for a more dynamic view of their interaction. In addition, the speech-act theory proposed by Austin (1962) and Searle (1969) could be criticized for a rigid categorization that isolates utterances outside their contextual setting.

Interaction in education is a complex phenomenon. An analytical matrix from isolated disciplinary fields – such as education, musicology, sociology or psychology – could not provide all the concepts required for different aspects of the study if the aim is to interpret data and enhance an understanding of interactional dynamics in an educational setting. An interpretation of data applying a single theoretical perspective runs the risk of promising and explaining too much at a level where the concepts cannot

provide a logical explanation for the occasionally implicit questions asked. By applying a multidisciplinary theoretical framework on diverse levels of the study, where each level corresponds to a set of research questions, the risk is reduced. There has, however, to be a logical and theoretical coherence linking the various levels and concepts based on a general theoretical definition of the studied phenomenon identifying which qualities of the phenomenon to be studied in the chosen methodology.

Theories and results from other fields can be extrapolated and used as an explicit background for the analysis and interpretation. A study of multimodal interaction in a complex educational setting requires concepts that define a perspective on learning, teaching, music, communication, education, and interaction on a personal, interpersonal as well as an institutional level. These concepts are interdependent to the extent that it would not be possible to understand the individual without a society, or a society without individuals. From this standpoint it becomes necessary to apply a critical societal perspective that provides us with an understanding of how the institution is confined within the routine actions of the teachers and students. These actions have evolved and gained legitimization throughout the history of the institution.

## **2. Research questions**

The ongoing Swedish study *Interaction in music education* analyses the interface between language, music and gesture as semiotic resources, rhetorically orchestrated in the classroom. A study of modalities of communication is combined with study of interpretative practices of teachers and students.

Data consists of 12 hours of video recorded and transcribed instrumental lessons. The main object of the study is to describe and analyse the intricate processes of teaching and learning in detail in three modes: speech, music and gesture. The results are based on an analysis of the complex multimodal interaction and the study focuses on how various communication patterns affect the students' opportunities to learn. The study is divided into three levels, each with their own set of research questions and theoretical concepts. The differing levels reflect continuous movement from the close-up description of *how* teachers and students act and interact, through a systematic analysis of *the patterns of interaction* in relation to students' learning possibilities, concluding with an interpretation

on a macro level of *why* teachers and students are interacting in the way they do.

With the particular scope of the theoretical and methodological decisions made, such patterns could emerge with coherence and consistency between different modes; from explicit or implicit expectations; oblique or direct information; which controls the definition of the situation, and how; as well as many more. The results of this analysis are discussed and interpreted within a wider historical and sociological perspective.

### **3. Methods and methodology**

The project was challenged with finding a way to handle the large amounts of video data systematically and transparently utilizing a combination of commonly used office software. The analyzing and reporting transcription tool developed (ARTT), consists of two software packages – Apple QuickTime Player and Microsoft Excel – connected through a third software package – AppleScript – that makes it possible to program simple strings of code to control the system software to enable the different software programs to interact. The software tool developed rendered it possible to view the digital video synchronized to a spreadsheet containing connected fields for transcription and coding interaction in different modes, all on a single computer screen. The spreadsheet is programmed to aid the recognition of patterns in the interaction, as well as more quantitative modes of output.

A small digital camcorder with a wide-angle lens and a built-in microphone of good quality was placed on a tripod so that it captured both the teacher and the student(s). Researchers were not present during the recording to reduce the effect on the informants to a minimum. An ethical decision to keep the informants anonymous led to the decision to shoot three times the amount of video that was needed, and to make a randomized sample of the lessons to be analyzed.

The method of transcription focuses on the events during the lessons as a series of communicative messages (Green, 1999) in three often overlapping or simultaneously occurring communicative modes: music, speech and gesture. The transcription of the various modes in a single transcript chart renders it possible to analyze if the messages conveyed in the three modes were coherent or incoherent. This technique could reveal inconsistencies and conflicting messages within the communication

process. Therefore the possibility to simultaneously view the transcription of the different modes side by side is of great importance. This makes the graphical layout of the transcription as well as the coding of the modes vital in terms of what can be shown.

The representation of the teacher-student dialog is based on the message units, which are symbolized with a new cell in the transcript chart. The beginnings and endings of the message units are distinguished by 'contextualization cues' such as pauses, prosody, gestures, etc (Green and Dixon, 1994; Green, Franquiz and Dixon, 1997; Green, 1999). The transcript chart is divided into columns: 'time code', 'teachers' musical activity', 'students' musical activity' 'teachers' talk', 'student's talk', 'teachers' gestures', and 'students' gestures'.

Five educational functions of *language*, *music* and *gestures* are differentiated with inspiration from sociolinguistics and adapted in the previous study: *testing*, *instructing*, *accompanying*, *analytical* and *expressive* functions. Each transcribed message unit from the teacher and students is coded with the concepts and the frequencies of the different functions of speech, gesture and music usage are registered for each lesson, and for teacher and students respectively. The concepts of schemata and focus of attention are differentiated into four categories: *cognitive*, *motor*, *expressive* and *social*. Each utterance in the transcript is coded with one of the concepts. The interaction is discussed in relation to research on optimal learning conditions in music.

The cognitive concepts of *schema internalization* (Bartlett, 1932; Dowling and Harwood, 1986; Arbib, 1995) and *focus of attention* (Treisman and Davies, 1973; Shaffer, 1975; Treisman and Gelade, 1980; Allport, 1980; Navon, 1985; Bamberger, 1996; 1999) are used as explanatory models of musical learning at the individual level. The concepts of schemata and focus of attention are differentiated into four categories: *cognitive*, *motor*, *expressive* and *social*. Each utterance in the transcript is coded with one of the concepts. The theory of attention has shown that it is difficult to divide attention on different tasks. Some forms of communication render it difficult for the student to focus on, and internalize an adequate motor schema since most of their attention is geared towards decoding language and the different symbolic systems used by the teacher.

An analysis of how students and teachers shift their focus of attention during lessons renders it possible to trace *learning sequences* (Gordon,

1993) through the observation and coding of actions. By coding each message unit by its educational function, and the focus of attention respectively, we can relate the emerging patterns to previous findings on optimal learning situations.

On the third level the results from the description and analysis on individual and interpersonal levels are discussed in terms of their institutional and societal origins and how these patterns effects the distribution of musical knowledge in society in terms of power and hierarchy. The metalevel is the third stage in mapping the interaction and their consequences for students' possibilities to learn.

## **4. Main research findings**

### ***4.1 Messages in gestures and their decoding***

During lessons students' attention is divided since they have to shift focus between the printed score, complex motor control, grasping auditory feedback and decoding gestures of the teacher. Teachers' attention is primarily focused on giving ad hoc instructions about decoding symbols or on technical aspects. When the teachers' messages in gestures and language are contradictory to each other, the student has to deal with this by using a greater part of his/her attention in order to make sense of the situation. This will have a negative influence on the learning process since students will have less mental capacity available to focus on the issue at hand.

The gestures of the teacher are quite often used to communicate a disapproving message contradictory to the verbal affirmative instruction or evaluation. By doing so teachers can state values and feelings that would be controversial if communicated in words. Conflicting messages reduce the students' possibilities to address the problem directly. They still have to make some sense of the situation, maybe by putting the blame on themselves for a problem that has never been openly addressed by the teacher.

### ***4.2 Messages and modes***

Mode is a culturally and socially fashioned resource for representation and communication (Kress et al, 2001; Kress, 2003). Messages are communicated through message units in more than one single mode,

starting in one mode to transform into other modes; for instance from notational symbols, through verbal and gestural instructions, into musical sound. Speech, music and actions in the study seldom seem coherent when viewed independently from one another. Each mode has its own semiotics and the meanings of modes are intertwined and contributes in cooperation to how instructions can be understood. Different modes have their separate limitations and potentials. Time-based modes have logics and potentials for representation that differ from space-based modes (Kress, 2003).

Knowledge changes its shape when it is transformed and communicated through the different modal material. Kress (2003) proposes that learning therefore can be studied as a dynamic process of transformative sign making which actively involves both teacher and students. This reveals challenges to the designs for learning; for the space-based modes as well as the time-based modes and mixed modes, physical environments and artefacts as well as the tuition. Several generally held views on education are challenged through such a multimodal analysis; that teaching and learning are mainly applied practical accomplishments, that education is not in need of rhetoric, and that learning is an acquisitive process in which students acquire information from the teacher and from method books (Kress, 2003).

## **5. Conclusion**

From the perspective of multimodality we suggest that learning can be seen as an expansion of the access to diverse semiotic resources, increasing the repertoire of possible thoughts and actions within the field. This reflects a movement of focus from instruction to construction; from teaching and learning towards focusing the educational functions of both instruction and construction in designs for learning.

## **References**

- Allport, D. A. (1980). Attention and performance. From G. Claxton (Ed.) *Cognitive Psychology: New Directions*. London: Routledge and Kegan Paul.
- Arbib, M. A. (1995). Schema theory. In Arbib (Ed.) *Brain Theory and Neural Networks*. 830-834. Massachusetts: The MIT Press.

- Austin, J. L. (1962). *How to Do Things with Words. The William James lectures delivered at Harvard University in 1955*. Oxford: Clarendon Press.
- Bamberger, J. (1996). Turning music theory on its ear. *International Journal of Computers for Mathematical Learning*, 1 (1), 33-55.
- Bamberger, J. (1999). Learning from the children we teach. *Bulletin of the Council for Research in Music Education*. 142, 48-74.
- Bartlett, F. C. (1932). *Remembering*. Cambridge: Cambridge University Press.
- Bellack, A. A., Kliebard, H. M., Hyman, R. T., & Smith, F. L. (1966). *The Language of the Classroom*. New York: Teachers College Press.
- Dowling, W. J. & Harwood, D. L. (1986). Music Cognition Academic Press Series In *Cognition and Perception*. Orlando: Academic Press, Inc.
- Eco, U. (1984). *The Role of the Reader*. Bloomington: Indiana University Press.
- Fairclough, N. (1995). *Critical Discourse Analysis. The critical study of Language*. London: Longman Group Ltd.
- Garnica, O. & M. King (Eds.) (1978), *Language, Children, and Society*. 159-174. New York: Pergamon.
- Gordon, E. (1993). Learning Sequences in Music. Skill, Content, and Patterns. 1993 Edition. *A Music Learning Theory*. (First publ. 1980.) Chicago: GIA Publications.
- Green, J. & Dixon, C. (1994). The Social Construction of Classroom Life. *Encyclopaedia of English Studies & Language Arts*. A. C. Purvis (ed.), NCTE in collaboration with Scholastic Press.
- Green, J. L. (1999). Transcribing as a conceptual process: Exploring ways of representing classroom activity. Paper presented at a workshop 15 September 1999. Uppsala University, Institution of Pedagogy
- Green, J., Franquiz, M. & Dixon, C. (1997). The Myth of the Objective Transcript: Transcribing as a Situated Act, *TESOL Quarterly*, 31 (1), pag.172-176.).
- Kress, G. (2003). *Literacy in the New Media Age*. London: Routledge.
- Kress, G., Jewitt, C., Ogborn, J. and Tsatsarelis, C. (2001). *Multimodal Teaching and Learning. The rhetorics of the science classroom*. London: Continuum.
- Navon, D. (1985). Attention division or attention sharing. In M. I. Posener & O. S. M. Marin (Eds.) *Attention and performance vol. 11*. Hillsdale NJ: Lawrence Erlbaum.
- Searle J. R. (1969). *Speech acts: an essay in the philosophy of language*. Cambridge: Cambridge University Press.

- Shaffer, L. H. (1975). Multiple attention in continuous verbal tasks. In P. M. A. Rabbitt & S. Dornic (Eds.) *Attention and performance, vol 5*. London: Academic Press.
- Shannon, C. & Weaver, W. (1949). *The Mathematical Theory of Communication*. Urbana: University of Illinois press.
- Sinclair, J. McH. & Coulthard, R. M. (1975). *Towards an analysis of discourse: the English used by teachers and pupils*. London: Oxford University Press.
- Treisman, A. & Davies, A. (1973). Divided attention to ear and eye. In S. Kornblum (Ed.) *Attention and Performance IV*. London: Academic Press.
- Treisman, A. & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology* 12, 97-136.

## Biography

### **Anna-Lena Rostvall, PhD.**

PhD in 2001 on the thesis Interaction and learning. A study of music instrument teaching. Currently Assistant Professor at the Stockholm Institute of Education. More than 15 years of experience as a guitar teacher. At present leading a 3-year project funded by the Swedish Research Council: 'Interaction and learning in music education'.

**Dr. Tore West** is currently Assistant Professor at the Stockholm Institute of Education. His research focuses on analysis of video recorded classroom interaction, with main interest concerning what effect different patterns of multimodal communication can have on opportunities for learning.

### **Authors' addresses**

*Dr. Anna-Lena Rostvall*  
*Stockholm Institute of Education*  
*Box 34103*  
*10026 Stockholm*  
*Sweden*  
*phone: +46 8 737 95 69*  
*e-mail: anna-lena.rostvall@lhs.se*  
*URL: <http://www.didaktikdesign.nu/musik>*

*Dr. Tore West*  
*Stockholm Institute of Education*  
*Box 34103*  
*100 26 Stockholm*  
*Sweden*  
*phone: + 46 8 737 95 67*  
*e-mail: tore.west@lhs.se*  
*URL: <http://www.didaktikdesign.nu/musik>*

# MULTIMODALITY STIMULATION: UNDERSTANDING FLUENCY IN STUDY ABROAD PROGRAMS

*Karen Woodman*

University of New England, Australia

## **Abstract**

*Study abroad students often enter programs having significant knowledge of the target language. Traditional frameworks do not accommodate the impact of this previous knowledge on early increases in fluency. This paper synthesizes research on activation, acquisition and attrition to develop a model of language activation in study abroad programs.*

*The question of why study abroad (and immersion) programs are typically more successful in increasing language fluency than classroom environments has been approached from a number of perspectives, and across a number fields. However, little consensus appears to exist as to the key factors underlying and influencing this process. While it is generally agreed that it's 'something about the environment' that promotes such language change, there is considerable debate over the identification or role of specific factors (e.g., frequency, motivation, identity, cognitive factors, etc).*

*Drawing on recent research in the fields of psycholinguistics (e.g., Ellis, 2004; Jarvis, 2000), cognition and neurolinguistics (e.g., Paradis, 2000), this paper proposes a theoretical model in which the key to the operationalization of activation in the study abroad (or immersion) experience is the recognition of language learners as progressive bilinguals (e.g., with differing interlanguage levels of linguistic and sociocultural competence) with L2 knowledge and/or skills which are particularly activated by the multimodality stimulation of the environment typical of study abroad programs.*

*This paper synthesizes research incorporating the concepts of partial concept activation, multimodality activation of language, influence of working memory limitations on language processing, influence of environment on language dominance, frequency effects and activation thresholds.*

**Keywords:** Multimodality, multimodal stimulation, language activation, study abroad programs, fluency development, SLA

## 1. Introduction

The question of why study abroad (SA) programs appear more successful in increasing language fluency than foreign language (or ‘at home’ (AT)) classroom environments has been approached from a number of perspectives (Freed, 1995a, b, c; Collentine & Freed, 2004) and across a number fields (Gardner, 1985). However, little consensus appears to exist as to the key factors underlying and influencing this process. While it is generally agreed that it’s ‘something about the environment’ that promotes such language change, there is still considerable debate over the identification or role of specific factors (e.g., frequency, motivation, novelty, etc. – e.g., Freed, 1995, 2004; Ellis, 2002). Although many researchers appear to assume contributing factors include increased frequency and/or exposure to language (e.g., increased input and output) in the SA environment, the possible role of differences in modality of such inputs/outputs in the SA vs. classroom does not appear to have been explored to any degree. In this paper, I will review recent research in a number of fields to identify key features of the study abroad environment which may help to explain apparent increases in fluency in SA participants, specifically related to multimodality stimulation. I will suggest three categories of factors in the study abroad environment may be considered relevant to understanding increases in fluency, and attempt to relate them to the impact of multimodality stimulation.

These categories can be broadly defined as:

1. Type of input and output (e.g., modality, context-specificity)
2. Frequency of input and output (e.g., increased input and output; threshold effects in neurological activation)
3. Individual factors (e.g., motivation, learning style, and idiosyncratic reactions to the SA experience (e.g., ‘optimal stress’))

Although many researchers have recognized the roles of these types of factors in language acquisition, the underlying aspect of modality in these categories does not appear to have been examined. In this paper, I will explore the complex interaction between factors in these three categories

which may help explain much of the individual variation which has also been found in study abroad outcomes. Specifically, I will propose a role for multimodality stimulation in language activation in the explanation of observed increased linguistic fluency in study abroad (or SA) contexts identified by Freed (1995a, c) and others (e.g., Collentine & Freed, 2004; Ellis, 2002; Freed, Segalowitz & Dewey, 2004; Segalowitz & Freed, 2004; Woodman, 2001). It will be proposed that a key difference between most classroom-based experiences and the study abroad context is the modalities through which linguistic, paralinguistic, cultural, and non-linguistic input can be perceived, processed, and, ideally, remembered and produced (e.g., output) in terms of the type of modality available, as well as the amount or frequency of input/output. Further, it will be suggested that these differences have implications on the cortical and neurological level for language acquisition, language activation, and language production. In other words, study abroad contexts provide stimulation in modalities which are rarely invoked in the classroom, and that these differences may provide a theoretical framework for the observed increased fluency. These modalities, interacting with individual differences in learning style and strategies, may serve to (1) activate (or increase access to) previously stored language and language skills (e.g., increasing efficacy, perhaps to the level of automaticity), as well as to (2) support development of new memories or consolidation of new language and language skills.

## **2. Background**

For many years, folklinguistic tradition has assumed that time spent in a target language community (i.e., the study abroad context) is good for language acquisition and/or language development (Miller & Ginsberg, 1995). For these reasons, many language programs offer in-country study options, with thousands of students annually traveling overseas for language study (Collentine & Freed, 2004). What appears to underlie such assumptions is the belief that the SA context contributes to the development of more ‘natural’ – or more fluent - speech.

In fact, the research on study-abroad programs has provided support for a positive impact on the language proficiency of the majority of language learners who have spent time in the target language when compared with learners who have not had similar exposure to the target language environment – especially when ‘language proficiency’ is defined in terms of more fluent production, rather than more accurate production (e.g.,

Clément, Gardner & Smythe, 1977a & b; Collentine & Freed, 2004; Freed, 1995a, b, c; Freed, Segalowitz & Dewey, 2004; Gardner et al., 1974; Gardner, Smythe & Brunet, 1977; Hanna & Smith, 1979). Even the most recent research on language acquisition in study abroad contexts, by Freed and others [see Collentine & Freed, 2004], which has suggested that improvement is not necessarily found in all areas of language development and use for study abroad students (e.g., syntax, morphology, and literacy skills), still strongly supports Freed's (1995a,c) assertion that a key benefit of study abroad can be found in increased fluency.

Some of the other benefits identified for learning in study abroad contexts include:

1. Higher levels of proficiency (Brecht et al, 1995)
2. Significantly enhanced oral fluency and overall proficiency (as measured by speech rate, and hesitation phenomena) and cognitive fluency (lexical recognition, lexical access automaticity, as well as speed and efficiency of attention control) (Freed, 1995b; Segalowitz & Freed, 2004; Woodman, 1998)
3. Improved vocabulary acquisition (Milton & Meara, 1995)
4. Improved narrative abilities and semantic density (Collentine, 2004)
5. Higher levels of confidence about reading skills (than non-SA participants) (Dewey, 2004)

Some less positive outcomes of study abroad programs, which seem to be related to context-specific socio-cultural issues, include the following:

1. L1 discourse behaviours can have negative effects on social relationships in SA (Wilkinson, 1995)
2. Sexist attitudes in SA may negatively influence women's learning in SA (Polanyi, 1995)

Similarly, findings of individual variation appear to be related to how learners relate to their new environment, rather than the environment itself. The SA context seems to highlight individual differences in terms of:

1. Improvement in grammar (Guntermann, 1995)
2. Use of Japanese politeness strategies (Marriot, 1995)
3. Appropriate pragmatic use of Japanese politeness ('women's language') (Siegel, 1995)

Similarly, in other findings, Talburt & Stewart (1999) suggest that sociocultural attitudes related to gender and race in SA context influence efficacy of SA.

## 2.1 *Fluency and Study Abroad*

Freed (1995a) identifies a number of positive linguistic changes which are attributable to the study abroad learning experience. Specifically, she identifies positive fluency-based linguistic changes (perceptual and otherwise) as being a key difference between learners from SA programs vs. non-SA programs.

She states:

(Learners from study abroad programs tend)...

to speak with greater ease in competence, expressed in part by greater abundance of speech, spoken at a faster rate and characterized by fewer disfluency sounding pauses. Such students display a wider range of communicative strategies and repertoire of styles and their linguistic identities extended beyond the expected acquisition of oral skills to a new self-realization in the social world of literacy (p. 50).

In order to clarify this relationship, Freed (1995a) also indicates the need for more structured comparison of study abroad with other types of learning contexts. This recommendation resulted in a series of research collaborations with colleagues including Norman Segalowitz, Dan Dewey, Barbra Lafford and John Collentine [reported in a special 2004 edition of *Studies in Second-Language Acquisition*]. The SSLA (2004) studies systematically compare ‘at-home’ (AH) traditional foreign language classes, ‘at-home’ intensive summer language immersion programs (IM), and in target-language country study abroad (SA) programs. Languages studied by Freed and her colleagues include French, Spanish, Japanese, and Russian.

Of particular interest to this paper is one of the studies (Segalowitz & Freed, 2004) which used data sets from a larger study sponsored by the Center for International Education Exchange (CIEE). The semester long study involved 46 learners of Spanish in two learning contexts: AH (n=20) and SA (n=26) in Alicante, Spain. Participants completed the SATII Spanish test and the OPI. AT students were enrolled in intermediate or

junior-level classes at the University of Colorado. SA students were enrolled in three daily courses for foreign language students at the University of Alicante, and lived with host families. Participants also completed a Language Context Profile to assess the amount L2 was used outside the classroom. Segalowitz & Freed (2004) focus on the construct of fluency, which they operationalize as gains in oral fluency (e.g., measured by speech rate and hesitation phenomena), overall proficiency gains with measures of cognitive fluency measured by lexical recognition, lexical access automaticity, and speed and efficiency of attention control. They report significant effects for the SA context on learners' oral fluency and overall proficiency, with variables 'interacting in complex ways that certain cognitive threshold effects determine the degree of overall gains that learners will make' (Collentine & Freed, 2004: 162). Overall, they found greater gains in the acquisition of lexical breadth and narrative ability for SA students, ease and smoothness of speech, which is produced at more native-like speed.

The fluency-related findings of Freed (1995a, c; Freed et al, 2004; Segalowitz & Freed, 2004) are also mirrored in the work of Woodman (1998, 2001), who found significant positive changes in fluency-related phenomena for Japanese ESL students in a study abroad program, within as little as one week in-country. These findings raised a number of conceptual and theoretical issues, including how to explain an observed 'word spurt' which occurred within 3-4 days after arrival. While the literature has multiple terms for learning, acquiring, producing and comprehending language [see Ellis, 1994 for review], there did not seem to be a wholly satisfying explanation for the observed phenomena.

Woodman (1998, 2001) proposes a distinction between *language acquisition* (e.g., acquiring new knowledge, skills and competences, or "KSCs") and the *activation* of stored KSCs. It is proposed that it is primarily in the process of *activation* that the impact of multimodality stimulation can initially be seen (e.g., in terms of fluency in SA programs). It is the type and amount of stimulation in the environment, which is influenced by both the type (e.g., modality) and amount (e.g., frequency) of language-related stimulation.

Also, the concepts of 'language activation' and 'multimodality stimulation' find support from cognitive and neurolinguistic research in the field of bilingualism (e.g., Paradis, 2000) and from Ellis' (2002) work on frequency effects, Schumann and colleagues (2002; 2004) research on the role of the

neocortex, hippocampus, and amygdala in SLA, and Sapolsky's (2004) proposals concerning the positive impact of 'optimal stress' on memory.

### **3. Language activation. Key issues**

Key issues in the definition of language activation include:

1. a distinction between language learners and progressive bilinguals
2. a distinction between activation and acquisition
3. the identification of neuro-cognitive processes underlying activation, especially multimodality stimulation

The distinction between *language acquisition* and *language activation* is important conceptually to the understanding of increased fluency in SA programs and the impact of multimodality stimulation, since this terminological distinction explicitly acknowledges the role of prior learning/acquisition as underpinning this fluency development [i.e., the importance of increased access and automaticity – e.g., Bley-Vroman, 2002; Ellis, 2002a, b, c; Goldman-Rakic, 1992; Hulstijn, 2002]. This definition recognizes that many students in SA have studied the target language prior to taking part in programs (e.g., Huebner, 1995), so the extent to which any observed change in language production is taken as evidence of *language acquisition* (as opposed to *language activation*) should be limited to clear examples of linguistic skills or new lexicon known to have been unknown prior to the SA experience

#### **3.1 What is a progressive bilingual?**

Explicitly recognizing what it is that learners bring to the study abroad (second language) environment is important to understanding the possible impact of context (especially multimodal inputs). In fact, although many SLA researchers tend to use the term 'language learners' for participants in study abroad programs, usually these individuals have had some previous experience with the target language (Huebner, 1995). This reality is clearly implicit in the use of language placement testing in such programs which separates students into such groupings as beginner, intermediate, and advanced classes. However, the effect of prior knowledge on language production, ironically, is not clearly operationalized in the literature. Significantly, the term 'language learner' tends to obscure a vital factor in

the understanding of possible influences in increased fluency: increased access or activation of previously stored (learnt) knowledge, skills, etc.

Discussing the differences between the fields of SLA and bilingualism, Jansen (2000: 19) notes:

The subjects of investigations who are referred to in one field as *learners and in the other as bilinguals are often the very same people* [my emphasis], and the two fields (SLA and bilingualism) share common concerns about how these people acquire, process, use language, and perform culture. Whereas the two fields share common concerns, however, they often differ in the emphasis. Bilingualism has traditionally had its hands more deeply in issues related to the mind, brain and identity (e.g., aphasia, code-switching, language-concept associations), whereas second language researchers have been more concerned with issues related to the acquisition of target-language forms and functions (e.g., syntax, morphology, pragmatics).

One result of these distinctions has been an assumption in the field of bilingualism that language learners should be viewed as bilinguals (e.g., possessing two language or knowledge or conceptual systems) - albeit with the developing language reflecting the current level on the IL continuum (e.g., Jarvis, 2000; Paradis, 2000); whereas in much of the SLA literature, it is not explicit. Although considerable debate exists in the field concerning the definition of the conceptual framework of bilinguals [summarized in the 2000 issue of "Bilingualism: Language and Cognition 3(1)"], this recognition of the knowledge/skills/competences [KSC] that language learners possess is critical to the focus of this paper and the argument for the role of multimodality stimulation frequency and activation in study abroad situations.

Elsewhere, I have proposed that the term 'progressive bilingual' as providing a clearer acknowledgement of the L2 knowledge, skills and competences that individuals bring to the Study Abroad (and/or immersion) experience (Woodman, 2001). These L2 knowledges/skills/competences are part of the abilities that individuals bring to the SA environment, and are the reservoir from which 'activation' arises when multimodality stimulation occurs.

Progressive bilinguals, at all stages before 'true bilingual', possess L1 (mother tongue) and L2i (dynamic interlanguage).

1. The L1 system will include that individual's current KSC [knowledge/skills/competences] including level of native speaker linguistic and competence, as well as reflect their individual levels of communicative competences.
2. The L2i system may be conceptualized as a dynamic interlanguage system reflecting that individual's current KSC in the target language of different levels of linguistic and communicative competences.

The L1 linguistic and cultural KSC system will, of course, influence development of the L2i system. [Note: The degree and type of cross-over is beyond the scope of this paper.]

It is proposed that these L2 KSC which are part of the attributes of the progressive bilingual are key to understanding the early increases in language production ('word spurt') identified in Woodman (1996, 1998, 2001), comprehension and overall increased 'fluency' frequently observed in Study Abroad programs (e.g., Freed, 1995; Freed et al, 2004; Segalowitz & Freed, 2004). Specifically, it is proposed that much of the observed very early increase (or 'word spurt') in language production and/or fluency is the result of activation of stored L2 KSC, rather than part of newly acquired/learnt information.

Activation is meant here to differentiate language change attributable to:

1. Activation of the interlanguage system or KSC (L2i) at the time of entering an SA situation, from language change (2) L2i + based on up-take of new KSC after entering the SA environment
2. Thus, L2i (activation) provides the KSC for initial production and comprehension in the SA environment, and arguably provides the mechanism for the 'word spurt', and fluency changes

Some observed language change could be considered language activation, rather than language acquisition. If then the SA-based multimodality stimulation(s) resulted in an "activation" of passive knowledge (learning) or "reactivation" of rarely used "lapsed" active knowledge, we would predict a relatively rapid change in language production as re-activation occurs (perhaps in the microlinguistic level) rather than a slower development which would be necessary for language acquisition ( e.g., completely novel material to be acquired). Or in fact, both processes may occur. For example, an initial rapid change/increase in lexicon, etc may

occur, with a later, more gradual change in general language production also being found. This perspective is supported by research findings comparing SA and AH (e.g., Freed, Segalowitz & Dewey, 2004), wherein IM and AT students did better on reading and writing than SA participants.

### **3.2 *Frequency, memory and activation***

The impact of frequency on learning and memory has also been well-documented (e.g., Ellis, 2002; Paradis, 2000; Sapolsky, 2004). Increased frequency of exposure helps consolidate learning by increased efficiency and potentiation of neurons. Also, certain types of neurons, which are related to memory and learning, experience threshold effects, so that neurons will not fire until a certain level of activation is achieved. Arguably, these differences are relevant to language activation in SA contexts as different “types” of memory could be expected to ‘behave’ in different (and differently optimal) ways under different types of stimulation (environmental or other), with individual variation in impact and results<sup>1</sup>. In addition, as will be discussed below, stimulation from – and in – different modalities, also appear to differentially influence learning and memory.

#### **3.2.1 *Memory and Learning***

Memory and learning involve activation at a number of levels:

1. Neuronal level: Memory and learning are instantiated at the level of the neuron
2. Learning results from the strengthening of synapses between neurons, meaning a higher probability of firing under stimulation and;
3. Patterns of activation: Memories, or learned material, are realized in the brain as patterns of activation

In addition, as noted by Robert Sapolsky (2004: 30), “memory is not monolithic, but comes in different flavours”. He identifies these ‘flavors’ as including:

1. Short-term memory and long-term memory

---

<sup>1</sup> Ellis (2002a, b,c) provides a comprehensive review of the importance of these cognitive differences in SLA, especially in relation to the role of frequency effects on learning and language production.

## 2. Explicit (declarative) memory and implicit (procedural) memory

These different types of memory also involve different parts of the brain, including the roles of the cortex, the hippocampus, and the cerebellum [see Sapolsky, 2004]. And, at the neuronal level, “knowledge is stored in the patterns of excitation of vast arrays of neurons (e.g., neuronal networks)” (p. 30). Thus, (as noted previously) learning and storage of memories involves the strengthening of some network branches more than others.

According to Paradis (2000), from a neurolinguistic perspective, concepts are realized as patterns of activation with zones of relations from constituent to peripherally related (e.g., some are core parts, some are less core). And to the extent that ‘concept’ includes all the knowledge that an individual has about a thing or event, *‘a concept is never activated in its entirety at any given time; only those aspects that are relevant to the particular situation are activated’* (Damasio, 1989 – cited in Paradis, 2000) as the various components are scanned by working memory (i.e., are in awareness). Further, mental patterns can be modality-specific or multisensory. The implication then is that in any given situation only some aspects relevant to the specific situation in which concepts are evoked reach activation threshold, and that this activation can be multisensory or multimodality.

Paradis comments:

Further, the exact same portion of the network is not always activated every time a given word is heard or uttered. Not just different connotations, but different denotative aspects of the referent are activated in each context. In other words, concepts are not only dynamic (i.e., changing over time, as Pavlenko rightly claims, but they are also fractionable in that only those portions of a concept relevant in the particular situation of its use are activated (i.e., present in consciousness) (p. 22)

Related to Paradis’ comments is the role of glutamatergic synapses. According to Sapolsky (2004: 30), glutamatergic synapses have two properties critical to memory:

1. They are nonlinear in their function: when a certain threshold is passed, a massive wave of excitation follows in the second neuron (e.g., learning)

2. Under the right conditions, when a synapse has had a sufficient number of superexcitatory glutamate-driven experiences, it becomes persistently more excitable (i.e., has just learned something or is potentiated, or strengthened), and therefore takes less of signal to recall a memory

Thus, graduated levels of concept activation (e.g., from a neurolinguistic perspective) may help to explain the phenomenon of the perception of gradual remembering (e.g., partially remembered) of languages when entering the L2 environment (e.g., in SA or travel), the ‘word spurt’ reported in Woodman (1998, 2001), or even what others have described as ‘din in the head’.

Paradis also emphasizes the role of multi-sensory, or multi-modality, stimulation in the activation of concepts. In terms of this discussion, he therefore provides support for the perspective that the more stimulation in more modalities, the more complete conceptual activation will occur, facilitating increased lexical access and increased fluency.

Paradis notes:

Irrespective of the means by which a concept has been acquired, its relevant components will be equally activated by auditory, visual somesthetic, olfactory, or verbal stimulation. The sight of a cat, the sound of a cat, the smell of a cat, or the spoken or written word ‘cat’ will all activate the relevant portions of the concept {cat}. In English, the word ‘cat’, given the appropriate context (e.g., the big cats of the African wildlife reserves) may refer to and evoke the concept of a lion or cheetah. Chat, the French translation equivalent of cat, would not and could not - unless in the context of a statement such as ‘this tiger looks like a cat (Paradis, 2000:22)

In this discussion, Paradis is clearly acknowledging the impact of multimodal inputs on memory – and specifically, memory and language.

### *3.2.2 Memory, stress and activation*

Sapolsky (2004) suggests that mild to moderate stressors actually enhance memory. He refers to this sort of optimal stress as ‘stimulation’, and he indicates that ‘optimal stress’ makes us feel alert and focused. He explains that the sympathetic nervous system, indirectly arouses the hippocampus

into a more alert, activated state, which in turn, facilitates memory consolidation. This involves an area of the brain that is also central to understanding anxiety, the amygdala [i.e., this area is also a focus of John Schumann's (2004) recent research, but Schumann is concerned with the role of motivation in SLA]. The sympathetic nervous system also helps the energy needs of potentiating neurons to be met by mobilizing glucose into the bloodstream increasing the force with which blood is pumped into the brain. He notes that 'a mild elevation in glucocorticoids levels smooth the progress by which synapses in the neocortex and hippocampus become more sensitive to glutamate signals', and 'the long-term potentiation is the building block of learning'.

Of particular interest to the issue of the differential impact of stimulation (multimodal or otherwise) on fluency in study abroad programs is Sapolsky's statement that:

A mild elevation in glucocorticoids levels smooth the progression of long-term potentiation in the hippocampus as well...(results in) moderate, short-term stress mak(ing) sensory receptors more sensitive. *Taste buds, olfactory receptors, the cochlear cells in the ears all require less stimulation, under moderate stress to get excited and pass on the information to the brain* [my emphasis] (p. 32)

In other words, under moderate stress, which could arguably result from the novelty of being in an SA environment, learners could in fact be more receptive or more sensitive to inputs in multiple modalities than they would be (or would have been) in a normal (e.g., non-novel) learning environment, such as AT.

In addition, in terms of possible explanations of individual variation in language change study abroad programs, is the other side of the coin – that is negative reactions to the stress of the SA environment could potentially (and literally) limit the amount of activation and/or acquisition, or in fact suppress language functions (what is usually referred to as 'culture shock').

Therefore, I would suggest that the impact on fluency of multimodality stimulation in study abroad programs, may be related to optimal stress, although clearly further research is necessary. In other words, it may be that Sapolsky's 'optimal stress' is what is triggered by the novelty of the SA environment, which may in turn play a role in reinforcing or supporting (or even activating) memory and learning through heightening levels of awareness in multiple modalities (e.g., sight, smell, tastes, etc).

Whether it is possible to duplicate in the classroom levels of optimal stress which could provide/reinforce learning and memory in multiple modalities - especially to cultural and linguistic dimension – remains to be seen.

In addition, the characteristics and behaviours of any given individual will also have a strong impact on how they react to and in a specific context. So, while some people will find the study abroad environment that optimal learning situation, others may react in a manner which minimizes their contact and receptivity to frequency and type of input and output. Some issues here include stress, learning style, motivation, and language level (e.g., Collentine & Freed, 2004; Dewey, 2004; Juffs, 2002; Lafford, 1995, 2004; Polanyi, 1995; Siegel, 1995).

### ***3.3 Multiple modalities, fluency and the Study Abroad Environment***

In summary, I will propose the following differences related to multimodality stimulation in different learning contexts (specifically, in-country study (SA) and at-home (AH) contexts) may be found to exist.

1. Visual modality: The target language community may provide significantly more visual input and output for learning and activation. For example, visual stimulation can be found in signs on buildings, in stores, on the bus, etc. this provides language in context. There also may be more reading and writing required (e.g., in terms form-filling, newspapers, reading subtitles in movies, TV, etc.), exposing learners to more language, and a much wider variety of vocabulary, than would normally be encountered in the classroom. Non-linguistic visual stimulation may also be prevalent in this environment. Some examples may include differences in clothing, architecture, food, and ‘ways of doing things’.
2. Verbal modality: The target language community may provide significantly more opportunities for linguistic output, in realistic (authentic) social situations. For example, opportunities for use of routines and formulas may be increased, with increased possibilities for direct or indirect positive or negative feedback (e.g., output understood or not). Exposure to, and the need for, appropriate use of register, context-specific dialect variations, etc., may also be heightened in the SA environment.
3. Auditory modality: The target language community also may provide

significantly more auditory input for learning and language activation. First, the learner is surrounded by native speakers of the language talking to each other. The language may also be prevalent on the radio, on television, in the movies, etc. If the learner takes advantage both the target language community in terms of developing friendships and relationships etc., the impact of this modality may be higher.

4. Olfactory modality: It may seem strange to talk about the ‘smell of a community’. But memory has been strongly linked to smell (Paradis, 2000). The ‘smell of a language community’ would be linked to its geography, its cuisine, and its way of doing things. Immersion in the target culture may provide opportunities for memories linked to language to be activated and acquired. This modality would be difficult to duplicate in the classroom (especially for extensive periods of time).
5. Taste: Taste is also strongly linked to memory (Paradis, 2000). Each culture, and community has its own tastes, typically linked to traditional foods and food-related rituals and routines. Eating tends to take place in social situations, which would be rich in sociocultural and sociolinguistic learning opportunities. In fact, if we think in terms of the experience of students in SA programs - going shopping, having dinner with their host family, or even going to restaurant could provide an environment which stimulates all of the above modalities (e.g., reading the menu or signs, ordering, smelling and tasting the food).
6. Kinesthetic (touch, movement): The traditional language classroom tends not to exploit the body-memory modality (Paradis, 2000). However, SA environments provide enriched learning opportunities for observing native speakers using body language (gestures, non-verbal communicative strategies). Understanding and using appropriate body language is often critical to communication in such situations. In addition, body language appears to play a significant role in the perception of fluency (especially by native speakers) (e.g., Freed, 1995a). The need for verbal communication strategies (Lafford, 1995; 2004) may also be minimized by strategic use of culturally/linguistically appropriate gesture. Furthermore, physicality is also linked to reinforcement of memory (learning) through multiple connections (e.g., modalities) (Paradis, 2000).

**Conclusions: *Multimodality stimulation, language acquisition, language activation, and fluency***

This paper has proposed that multimodality stimulation is implicated in both language acquisition (e.g., formation of memories, or learning) as well as language activation (e.g., accessing memory, learning). It is argued that the SA environment provides multi-modality activation: living in the culture, hearing the language, seeing the language, tasting the language, inside and outside the class. This all-encompassing L2 environment can trigger more general activation -and more quickly - than in the limited exposure environment of the classroom. In fact, it might be suggested, the phenomenon often referred to as 'din in the head' (e.g., within the first days in the L2 environment, words and phrases in the new language seem to ring in the head) may be explained by the massively increased frequency and multisensory aspects of the SA environment on first arrival triggering unusually widespread increases in activation throughout the language centers of the brain as L2i areas which were previously sub-threshold due to lack of activation (e.g., being limited to the L2 classroom, for example). The phenomenon of increased fluency is attributable to the same forces - increased activation, easier retrieval.

The role of frequency effects and/or multimodality/multi-sensory stimulation on levels and degree of activation is also a crucial question from an applied perspective: if these factors are in fact critical to language activation - can they be incorporated into non-SA classroom practice? In other words, if 'activation' is a key to fluency in SA programs, then underlying this is the idea that multimodality stimulation results in diverse areas of activation, and eas(ier) access for production and comprehension.

Finally, if the process is cumulative (arguably accounting for general threshold effects) - like a spreading activation connectionist models, frequent or repeated activation of a particular area would predispose (or sensitize or load) that area such that subsequent activations would occur at a much faster rate (stronger synapses), and arguably, consume less resources (e.g., automaticity). As images, sounds and words sort themselves out, some connections within the language are strengthened via frequency effects, become stronger, and therefore more strongly related, and more automatic, and language fluency gains should result.

## References

- Bley-Vroman, R. (2002). Frequency in production, comprehension, and acquisition. *Studies in Second Language Acquisition*, 24 (2), 209-213.
- Brecht, R., Davidson, D., & Ginsberg, R. (1995). Predictors of foreign language gain during study abroad. In Barbara Freed. (Ed.). *Second language acquisition in a study abroad context* (pp. 37-66). Amsterdam: John Benjamins.
- Clément, R., Gardner, R.C., & Smythe, P. (1977a). Motivational variables in second language acquisition: A study of francophones learning English. *Canadian Journal of Behavioural Science*, 9, 123-133.
- Clément, R., Gardner, R.C., & Smythe, P. (1977b). Inter-ethnic contact: Attitudinal consequences. *Canadian Journal of Behavioural Science*, 9, 205-215.
- Collentine, J. (2004). The effects of learning contexts on morphosyntactic and lexical development. *Studies in Second Language Acquisition*, 26 (2), 227-248.
- Collentine, J., & Freed, B. (2004). Learning context and its effects on second-language acquisition. *Studies in Second Language Acquisition*, 26 (2), 153-171.
- Dewey, D. (2004). A comparison of reading development by learners of Japanese in intensive domestic immersion and study abroad contexts. *Studies in Second Language Acquisition*, 26 (2), 303-327.
- Diaz-Compos, M. (2004). Context of learning in the acquisition of Spanish second language phonology. *Studies in Second Language Acquisition*, 26 (2), 249-273.
- Ellis, N. (2002a). Frequency effects in language processing. *Studies in Second Language Acquisition*, 24 (2), 143-188
- Ellis, N. (2002b). Reflections on frequency effects in language processing. *Studies in Second Language Acquisition*, 24 (2), 297-339.
- Ellis, N. (2002c). *The cognitive neuroscience of noticing and fluency*. Paper presented at American Association of Applied Linguistics Conference (Salt Lake City, UT).
- Freed, B. (Ed.). (1995a). *Second language acquisition in a study abroad context*. Amsterdam: John Benjamins.
- Freed, B. (Ed.). (1995b). Language learning and study abroad. In Barbara Freed. (Ed.). *Second language acquisition in a study abroad context* (pp. 3-34). Amsterdam: John Benjamins.

- Freed, B. (Ed.). (1995c). What makes us think that students who study abroad become fluent?. In Barbara Freed. (Ed.). *Second language acquisition in a study abroad context* (pp. 123-148). Amsterdam: John Benjamins.
- Freed, B., Segalowitz, N., & Dewey, D. (2004). Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. *Studies in Second Language Acquisition*, 26 (2), 275-301
- Garcher, R., Smythe, P., & Clement, R. (1979). Intensive second language study in a bicultural milieu. *Language Learning*, 29, 305-320.
- Gardner, R.C. (1985). *Social psychology and second language learning*. London: Edward Arnold.
- Gardner, R.C., Kirby, D., Smythe, P., Dumas, G., & Zelman, M. (1974). Bicultural excursion programs: Their effects on students' stereotypes, attitudes, and motivation. *Alberta Journal of Educational Research*, 20, 270-277.
- Gardner, R.C., Smythe, P., & Brunet, G. (1977). Intensive second language study: Effects on attitudes, motivation and French achievement. *Language Learning*, 27, 243-261.
- Goldman-Rakic, P. (1992). Working memory and the mind. *Scientific American*, 267(3), 110-117.
- Guntermann, G. (1995). The Peace Corps experience: Language learning and training and in the field. In Barbara Freed. (Ed.). *Second language acquisition in a study abroad context* (pp. 149-170). Amsterdam: John Benjamins.
- Hanna, G., & Smith, A. (1979). Evaluating summer bilingual exchanges: A progress report. *Working papers on Bilingualism*, 19, 29-58.
- Huebner, T. (1995). The effects of overseas language programs: Report on a case study of an intensive Japanese course. In Barbara Freed. (Ed.). *Second language acquisition in a study abroad context* (pp. 170-194). Amsterdam: John Benjamins.
- Hulstijn, J. (2002). What does the impact of frequency tell us about the language acquisition device? *Studies in Second Language Acquisition*, 24 (2), 269-273.
- Juffs, A. (2002). *Working memory as a variable in accounting for individual differences in second language performance*. Paper presented at American Association of Applied Linguistics Conference (Salt Lake City, UT).
- Lafford, B. (2004). The effect of the context of learning on the use of communication strategies by learners of Spanish as a second language. *Studies in Second Language Acquisition*, 26 (2), 201-225.

- Lafford, B. (1995). Getting into, through and out of a survival situation: A comparison of communicative strategies used by students studying Spanish. In Barbara Freed. (Ed.). *Second language acquisition in a study abroad context* (pp. 97-122). Amsterdam: John Benjamins.
- Miller, L., & Ginsberg, R. (1995). Folklinguistic theories of language learning. In Barbara Freed. (Ed.). *Second language acquisition in a study abroad context* (pp. 293-316). Amsterdam: John Benjamins.
- Polanyi, L. (1995). Language learning and living abroad: Stories from the field. In Barbara Freed. (Ed.). *Second language acquisition in a study abroad context* (pp. 271-292). Amsterdam: John Benjamins.
- Sapolsky, Robert (2004). Stress-out memories. *Scientific American Mind*, 14(5), 22-33.
- Segalowitz, N., & Freed, B. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in At Home and Study Abroad contexts. *Studies in Second Language Acquisition*, 26 (2), 173-199.
- Schumann, J. (2002). *Variation in neural structure across individuals*. Paper presented at American Association of Applied Linguistics Conference (Salt Lake City, UT).
- Schumann, J., Crowell, S., Jones, N., Lee, Namhee, Schuchert, S., & Wood, L. (2004). *The neurobiology of learning: Perspectives from second language acquisition*. Mahwah, NJ: Lawrence Erlbaum.
- Siegal, M. (1995). Individual differences and study abroad: Woman learning Japanese in Japan. In Barbara Freed. (Ed.). *Second language acquisition in a study abroad context* (pp. 225-244). Amsterdam: John Benjamins.
- Woodman, K. (1996). Time to Change? Exploring the temporal dimension of SLA. In *Proceedings of the 16th Congrès International des Linguistes* (CIL 16). Paris, France.
- Woodman, K. (1998). *Linguistics, perceptual and pedagogical change in a short-term intensive language program*. Unpublished PhD dissertation. University of Victoria (Canada).
- Woodman, K. (2001). *Linguistics, perceptual and pedagogical change in a short-term intensive language program*. ERIC.

## **Biography**

**Dr. Karen Woodman** is a Senior Lecturer in the online Master of Arts (Applied Linguistics) in the School of Languages, Cultures and Linguistics at the University of New England, Armidale, NSW Australia. Her current research interests include language activation and fluency in study abroad programs, issues in online teaching and learning, learning disabilities in ESL, and teacher development in intensive TESL programs.

### **Author's address:**

*Dr. Karen Woodman  
School of Languages, Cultures and Linguistics  
University of New England  
Armidale, NSW Australia 2351  
phone: +61-2-6773-3381  
e-mail: Karen.woodman@une.edu.au  
fax: +61-2-6773-3735*

# MULTIMODALITY IN A THREE-DIMENSIONAL VOICE CHAT

*Therese Örnberg Berglund*  
Umeå University, Sweden

## **Abstract**

*In recent years we have seen an increase in computer applications that support multimodal online communication. This paper centers on communication in one such multimodal application, Traveler, a graphical three-dimensional environment that allows for voice communication. Based on preliminary results from a study investigating communication patterns and negotiation strategies in this environment this paper gives some examples of how modal density is created here, building on the methodological and theoretical framework proposed by Sigrid Norris (2004). Through qualitative analysis of the material, both actual and self-perceived behavior of beginner participants is compared to that of more accustomed users. The paper also includes a discussion on the relationship between modal density and notions such as Common Ground and Presence.*

**Keywords:** Computer-mediated Communication, Modal density, Conversational negotiation, Common Ground, Presence

## **1. Introduction**

### ***1.1 Multimodal interactions***

Face-to-face interaction is always multimodal. As human interlocutors involved in communication we do not only have verbal language at our disposal, but by combining gestures, facial expressions, intonation,

positioning and, admittedly, in most cases also language, we can transmit complex messages on several levels simultaneously. In addition, also appearance and spatial configuration influences our interactions. The interrelation between different modes is something which an increasing number of linguistic scholars are paying attention to when analyzing interactional meaning. If only language is taken into consideration, the full complexity of meaning construction cannot be accounted for, and the analyst misses out on important information which the interaction partner has access to when decoding the messages.

In recent years, many of our daily interactions have moved online, a transition which has consequences for communicative practices. Relevant in this context is the way in which technological mediation affects the modalities which can be drawn on in interaction and communication. In this paper, it is argued that these effects are not necessarily negative, but rather different media have different affordances (Gibson 1977, Norman 1988, Hutchby 2001). Nevertheless, if we want to learn how to identify the platforms that are best suited for our purposes, it is important to investigate how the different modes available are being employed, and how language and communication are adapted to fit these modalities by both beginners and more accustomed users in different online environments.

In this paper, I refer to preliminary results from a qualitative study of interaction and communication in a graphical three-dimensional voice communication environment, *Traveler*. Based on an analysis of the communicative interaction and on answers to questionnaires I give examples of how modal density is created in this environment, how multimodal behavior differs between beginner and non-beginner users, as well as how modality interrelates with concepts such as Common Ground and Presence. Before we turn to the preliminary findings, I will give a brief overview of previous research on multimodality and Computer-Mediated Communication (CMC), as well as an introduction to the theoretical and methodological framework on which this analysis builds.

## ***1.2 Multimodality and Computer-Mediated Communication***

The fact that the verbal is not the only means of communication is something which has been acknowledged by researchers of computer-mediated communication over the years. However, theories such as that of *media richness* (Daft and Lengel 1984) and the *cues filtered out* approach (see Walther 2002 for a summary) have seen face-to-face communication

as the ideal speech situation, and argued that the modes should imitate those of face-to-face communication as closely as possible. Some CMC researchers, for instance Joseph B. Walther (2002), have heavily criticized these approaches and shown that just because some cues are missing this does not necessarily mean that communication will break down or that relationships will not thrive. On the contrary, Walther (1996) argues that interpersonal relationships that develop in, for instance, written modes may instead become *hyperpersonal*, a term which indicates that the specificities of the media used for communication may facilitate even more intimate relationships than face-to-face interactions. A similar pattern in change of emphasis can be identified when surveying the developments within presence research. Whereas Short et al.'s (1976) notion of social presence as communicating under face-to-face like conditions is still prevalent today, for instance in virtual reality research, there are also other movements within presence research where focus is turned to the larger social context, including attitudes and social equality (c.f. McIsaac & Gunawardena 1996), and where also the role of imagination is taken into consideration (c.f. McLellan 1996).

In this paper I would like to suggest an alternative approach to multimodality and CMC, which in line with the reasoning of J.B. Walther and recent research on presence does not claim that less modes equals less efficient communication, but instead centers on the different affordances of the different media and on how modal density is created.

## **2. Theoretical and methodological framework**

### **2.1 *Modal density***

According to multimodality researcher Sigrid Norris (2004), it is never possible to count the modes available in a communicative situation, since they are merely heuristic units of analysis. Instead she advocates an approach to multimodality where focus is on the creation of modal density. Modal density relates to levels of attention, and there is no inherent hierarchy among modes, but it all depends on the situation. In Norris' view, modal density can be achieved either by *intensity*, which means that one mode is best suited to deliver a message under present circumstances. As an example of this, Norris points out how the verbal language is given prominence when speaking on the phone. *Complexity* may also result in modal density, in cases when several different modes are used

simultaneously to deliver the same message and none of the channels is given higher prominence than the others.

By applying the theory of modal density to computer-mediated communication this would indicate that in fact only one mode is needed in order for modal density to occur – only more attention will be devoted to this one mode, as in the case of written CMC. Thus, here it is argued that an approach which focuses on modal density rather than on perceptual realism offers a worthwhile possibility when analyzing multimodality in CMC.

## 2.2 *Conversational negotiation*

Building on the work of interactional sociolinguists (Goffman, Gumperz, Goodwin, Duranti) this paper argues that interaction is the key to an understanding of how meaning is construed. Here, the notion of conversational negotiation is used to refer to the recurring subconscious strategies through which reciprocal co-construction of content, structure and context, and thus discursive coherence, is enabled. Negotiation in communication can take on many different forms, depending on both the level of negotiation and the strategies employed. Participants in conversation subconsciously negotiate in order to make sense of what the other expresses through an utterance as well as how to continue the conversation structurally, in line with Common Ground theory (Clark and Brennan 1991) Furthermore context is co-constructed in conversation (c.f. Duranti & Goodwin 1992), which for example can be seen in the negotiation for face (c.f. Goffman 1967), identity (c.f. Weedon 1997), and solidarity and support (c.f. Aston 1986, 1993). Table 1 shows some examples of negotiation strategies on the different levels of negotiation.

Table 1. The three levels of negotiation

<b>Negotiation level</b>	<i>Content</i>	<i>Structure</i>	<i>Context</i>
<b>Negotiation type</b>	Negotiation of meaning; shared conceptualizations	Negotiation of process	Negotiation of solidarity, support, face, identity, roles etc.

### 3. Material

#### 3.1 *The Traveler environment*

Traveler is a three-dimensional voice-enabled virtual environment, which is mainly used for socializing, and where there is a strong sense of community among the regular users. The program, which is free to download to your computer, was created in the mid 90's and despite some improvements it still does not demand very much of your computer or your internet connection. Upon entering Traveler you choose and customize your avatar. The avatars are mainly big heads, which in itself has some interesting implications for multimodality. The graphics are on a quite basic level, as are the different non-verbal expressions that you can make your avatar express by clicking on certain buttons. In Traveler you communicate via voice, and there is lip sync between the sound and the avatar – however, not with any phonological detail. The sound is distance-attenuated, which means that the further away from someone you move the less well you hear what that person is saying and vice versa.

#### 3.2 *Modalities in Traveler*

The modes available in Traveler can be grouped under three headings: Audible, Visual and Spatial modes. Table 2 lists the modes that have been identified as central in Traveler, to be further discussed in the section on preliminary findings.

Table 2. Modalities in Traveler

<b>Audible modes</b>	<b>Visual modes</b>	<b>Spatial modes</b>
Language, prosody, pause, extralinguistic audile markers	Facial expressions, push-to-talk, appearance	Layout, proxemics, movements

#### 3.3 *The filmed gatherings*

My material consists of five filmed gatherings in Traveler, where different groups of people have met to discuss different issues. Four of the gatherings have been filmed from two different perspectives. The participants in the discussions are academic language teachers, a group of researchers, a student group who participated in a correspondence course of

English within an education program and parts of the Traveler community. Apart from community members, most participants are beginner users of Traveler. The number of participants has varied between 5-15. The majority of participants have not had any contact prior to the meetings, but some have known each other from before, either through virtual encounters or through face-to-face meetings. In three of the gatherings I have participated myself and in the other two I was present as a passive camera doing the filming.

## **4. Preliminary findings**

### ***4.1 Analysis of the communicative interaction***

Generally speaking, the modes connected to the medium of sound have the highest modal density in Traveler. On the content level, verbal language is used to generate topics and meaning, to clarify, to repair, to repeat, to elaborate, as well as to emphasize and to give feedback on the content of interaction. Intonation and pauses are also used to indicate emphasis, and prosody and extralinguistic markers are sometimes used as back-channeling devices. On the structure level, the different modes connected with the sound all are used as cohesive devices. On the level of context, these modes indicate conversational style, which of course is an important part of a person's identity. Pauses are sometimes also employed in face-saving strategies.

An important cohesive device in Traveler is the push-to-speak function. In order to be heard you have to push the ctrl button. This is both audible and visible to the other participants, and if someone has indicated that he/she wants to take the floor by pushing down ctrl, this person will often also gain access to the floor. Apart from the push-to-speak function there are not many visual indicators that someone wants to speak, something which at times makes turn-taking difficult. It should also be noted here that an additional reason why turn-taking sometimes is difficult in Traveler is the short time lag on the server. People who know about this might choose to have short pauses between utterances to avoid overlaps, without this making them uncomfortable.

As far as the visual modes are concerned, my material indicates two quite interesting phenomena. For one, the ability to use facial expressions is employed only to a limited extent, and in most instances by accustomed

users, either as face-saving strategies or to display level of attention. This indicates that in this environment emotional expressions to a great extent depend on the sound capabilities. In addition, the role of the avatar in identity construction is debatable. In my material, I have found indications that the visual representations, that is, the avatars chosen, start to matter less once you get to know the person behind the avatar. However, if someone changes avatar, this is often commented on as problematic, which indicates that even though these visual representations are not the focal point of attention they are still important. The anonymity of the avatars is another factor which influences roles and relationships, in that, at least in an initial stage, people are more equal – then these contextual factors are negotiated via other modes.

In *Traveler*, spatiality seems to be of greater importance than visual qualities. People tend to form circles and those who feel comfortable manoeuvring their avatars usually turn toward the person who is speaking. Spatial cues indicating that one is paying attention are important strategies on all levels of negotiation. This form of gaze is also apparent when accustomed users want to show whether they are addressing someone in particular or the whole group. Apart from movements where the avatars turn to face one another, also head-movements, most often nods, are used for back-channelling – mainly by accustomed users or those who have been specifically instructed on how to use this feature. These cues mainly occur on their own rather than in combination with verbal contributions. Another spatial cohesive device that some accustomed users employ to indicate level of participation is by moving their avatars back and forth to show that they want to take the floor and then open it up again.

One further example of how the spatial mode has effects on interaction is the fact that the spatial layout itself influences the contributions of the participants, in that it sets the expectations on level of formality, etc. This is illustrated in the following excerpt from a transcript of a student session in *Traveler*:

Table 3. Example from a session in Traveler

((M and O move around in the space and then return to S; O turns to face M))  
M Do you think that we should stay here at the eh at the gate or if we should gather around a table to be more structured I don know  
O I don't think it matters since eh it's only us here  
M Okay I was just eh thinking about ehm being able to like to to stay to stay in line of eh of everybody's view so that so that eh everybody sees everybody cause it's it's I don't know it probably doesn't matter  
O But perhaps it's more professional if we sit at the table  
M Shall we try it?  
((M moves toward the table))  
O Yeah, okay  
((O turns to face the table and moves in that direction))  
S Okay  
((S follows))

## ***4.2 Participants' attitudes in relation to two central analytical concepts***

In order to get some insights into participants' own views of interacting in this environment, those taking part in the online discussions have been asked to fill out questionnaires. In the following, some of the answers received will be exemplified by relating them to two analytical concepts that are central to my thesis, namely Common Ground and Presence. I have chosen to include answers from three different people, to ensure that both attitudes representative of beginners and more accustomed users will be presented. Two of these participants are newbies, but with quite different experiences, and one is an extremely accustomed user of Traveler.

### **4.2.1 Common Ground**

Common Ground theory deals with the ways in which people negotiate shared understanding, as regards both process and content of communication. It is established through *grounding*, a process by which participants in communication validate that they share a common understanding. This verification will take on varying forms depending on both the purpose of conversation and the media used. (Clark and Brennan 1991)

As we have seen in the previous section, negotiation for content and structure can be accomplished through several different modes. However,

in my material, language is the most common mode used for this purpose. One possible reason for this is that most participants in these gatherings in fact are newbies, and whereas initially the employment of these strategies take an effort from the participants, after a while they appear to become internalized. To illustrate this I have chosen some answers which indicate how three of the participants in a gathering for language teachers think that visual cues influence their conversation management. A and B are newbies, whereas C is an accustomed user.

Table 4. Examples of replies (Common Ground)

<b>Question:</b>	<b>Did you use many non-verbal cues, and did you notice if others did? If so, do you recall on what occasions these were used? Did the non-verbal cues add anything, and when they were not used, did you miss them?</b>
<b>A (beginner user)</b>	“D and I played around with the emotions –angry, happy etc but we couldn’t see much of a change. Also tried nodding. I think it would be essential when using this in a class, to go through the ways that you express non-verbal clues. I noticed that E seemed to zoom in and out and nodded which made it clearer how she was feeling and how she wanted to participate.”
<b>B (beginner user)</b>	“I tried to use as many of the non-verbal affordances as possible – this was one of the main interests I had in joining the meeting and using Traveler. I used movement of the avotar, nodding agreement, smiling, change of position to face speaker, location to be inclusive, and reversing to indicate ‘resting’. The non-verbal behaviour adds a new dimension to online communication and made a significant difference for me – though not all positive.”
<b>C (accustomed user)</b>	“I often nod my head in agreement or bow in response to a “thank you.” I think it adds the yes or no responses without interrupting. I can’t say that anything is missing if others don’t. It’s just handy to respond without stopping somebody’s train of thought.”

Here we can see how these three participants have all given this aspect of the environment some thought, and how the two newbies have tried to use the non-verbal cues in an experimental way, whereas the accustomed user has started using the ones that he finds most functional in his interactions. By combining sound and those visual and spatial cues that can be used for feedback without interrupting, modal density is created. It should be added that this specific accustomed user also has been observed using emotes on a number of occasions where they have had a social/contextual function rather than a structural.

#### 4.2.2. Presence

Closely related to Common Ground is the notion of Presence. As illustrated in the section on previous research, presence is a complex notion with many different definitions. In this paper, presence refers to a sense of sharing a space, which can either be accomplished by perceptual stimuli or by less technologically advanced techniques that nonetheless can cause a feeling of immersion, such as personal and intimate language use. In this environment, the sense of shared space has implications for the quality of interaction. The feeling of being present can for instance be seen in the use of deictic expressions and reference. Address is another indicator of presence – sharing a space like this makes it possible to address the whole group or parts of it with personal pronouns or even with spatial cues only, rather than with proper names. The type of multimodal representation which Traveler allows for might both make communication smoother and create presence through relating to a basic spatial communicative situation. The answers to the following question illustrate participants’ experiences of presence when communicating in Traveler.

Table 5. Examples of replies (Presence)

<b>Question:</b>	<b>Did you feel as if you were transported to a place which you shared with the other participants, or did you think about how this all took place on your computer screen?</b>
<b>A (beginner user)</b>	“Yes and no-I was involved in the place but started to realize that I could be quite rude as well and check my mail or surf the net while people were talking.”
<b>B (beginner user)</b>	“It took me to another world and was a real adrenaline buzz. It was on my screen and I was conscious of it always, but I was definitely virtually gone from my usual habitat. It took me a little while to come down again....”
<b>C (accustomed user)</b>	“I am always immersed. I throw my mind into the environment easily. It doesn’t matter that the environment is artificial. My house is man-made too but I prefer it to a cave. I think of the place as real, even though I understand better than most the mechanics of how Traveler worlds are constructed.”

The answer of B is representative of most answers to this question. Interesting to note is that both one of the beginners (B) and the accustomed user (C) express similar attitudes, since this indicates that the habituation effect has not made the experience less immersive for the accustomed user. Of course, there are probable side explanations to why the accustomed user expresses these attitudes, for instance the social aspect – C has made friends in Traveler and regularly meets with them here. The fact that A has

been paying attention to other things while participating in the meeting might explain in part her low level of immersion. A has also expressed a dislike for the surreal avatars in answers to previous questions, which most likely will have had an effect on her experiences.

Most users in my material seem to be comfortable with the surrealistic avatars and environment, and instead the realism of the situation – people meeting in a shared space and discussing via voice – appears to have a great influence on their experiences. This indicates that personal relationships and involvement in combination with a realistic sense of shared space are the greatest generators of a sense of presence in this environment.

## **5. Conclusion**

In my material, modal density is most often created through intensity. Especially beginner users, but also more accustomed ones mainly depend on the sound resources, and thus, sound carries the greatest modal density in this environment.

Interesting to note is what appears to be a lack of complexity between the different modes. Only on a few occasions is modal density achieved through complexity. I believe that this can be explained by the newbies' relative unfamiliarity in using the visual and spatial cues, and the more accustomed users' ability to select the most functional cues from the different modes. However, the visual mode is nevertheless important, since especially with larger groups involved, turn-taking, feedback on the content level as well as emotional support and face saving strategies will be smoother when not having to rely on sound only. Those who have learnt how to make use of the visual and spatial modes have a great advantage here.

One possible explanation to why participants sometimes express frustration when visual cues appear to be missing is that the situation in itself so closely resembles face-to-face interaction that beginner participants expect the same conventions to apply here as in face-to-face. Accustomed users, on the other hand, learn new conventions specific for this type of environment in the social context of communicating here, and thus learn to make the most of the modes available. One example of this is the way in which accustomed users are starting to employ the spatial modes from a

functional perspective, for instance by manoeuvring the avatars to show either general or specific address, or by moving them back and forth as part of communication management.

The fact that visual and spatial expressions in Traveler depend on deliberate actions on behalf of the user has consequences for grounding. Participants in interaction need to focus their attention and intentionally send out signals that will help in the grounding process. After having learnt the social conventions, also these expressions will become internalized, and only then can complex modal density help participants reach common ground.

Modal density is related to levels of attention, and it is through modal density that one shows situational and interactional focus. The ability to express level and focus of attention has implications for both negotiation of context, structure and content. Further, my findings indicate that by making it possible to detect the level of attention of the others, modal density influences the sense of presence that participants experience. Another strong generator of presence here is the way in which the three-dimensional environment creates an illusion of shared space.

In sum, these preliminary findings support the argument that when studying multimodality in CMC a focus on modal density may be more fruitful than one on perceptual realism, since this approach draws the attention to environment specific conventions and affordances rather than to face-to-face interaction. In order to compare the suitability of the modal density framework over other multimodal approaches more investigations need to be undertaken, especially so in different types of environments with different modal affordances.

## References

- Aston, G. (1986). 'Trouble-shooting in interaction with learners: the more the merrier'. *Applied Linguistics*, 7(2): 128-143.
- Aston, G. (1993). 'Notes on the interlanguage of comity'. In Kasper G. & S. Blum-Kulka (Eds.), *Interlanguage Pragmatics*. Oxford: Oxford University Press: 224-50.
- Clark, H. H., & Brennan, S.E. (1991). 'Grounding in Communication'. In Resnick et al. (Eds.), *Perspectives on Socially Shared Cognition*. The American Psychological Association: 127-149.

- Daft, R. L. & Lengel, R. H. (1984). 'Information Richness: a new approach to managerial behavior and organization design. *Research in Organizational Behavior*, 6: 191-233.
- Duranti, A. & Goodwin, C. (Eds). (1992). *Rethinking context: Language as an interactive phenomenon*. Cambridge: Cambridge University Press.
- Gibson, J. J. (1977). "The theory of affordances". In R.E. Shaw & J. Bransford (Eds.), *Perceiving, acting and knowing*, Hillsdale, NJ: Lawrence Erlbaum Associates.
- Goffman, E. 1967. *Interaction Ritual: Essays in Face-to-Face Behavior*. Chicago: Aldine Publishing Company.
- Gumperz, J. J. (1982). *Discourse Strategies*. Cambridge: Cambridge University Press.
- Hutchby, I. (2001). *Conversation and Technology. From the telephone to the internet*. Cambridge. Polity Press.
- McIsaac, M. S. & Gunawardena, C. N. (1996). 'Distance Education'. In Jonassen D.H. (ed.), *Handbook of Research for Educational Communications and Technology*. New York: Macmillan Library Reference: 403-437.
- McLellan, H. (1996). 'Virtual Realities'. In Jonassen, D.H. (ed.), *Handbook of Research for Educational Communications and Technology*. New York: Macmillan Library Reference: 457-487.
- Norman, D. A, (1988). *The Design of Everyday Things*. New York: Doubleday.
- Norris S. (2004). *Analyzing Multimodal Interaction. A methodological framework*. New York & London: Routledge.
- Short J., Williams E. & B. Christie. (1976). *The Social Psychology of Telecommunications*. New York: John Wiley & Sons.
- Walther, J. B. (1996). 'Computer-mediated communication: impersonal, interpersonal and hyperpersonal interaction'. *Communication Research*, 23: 3-43.
- Walther, J. B. & M. R. Parks. (2002). 'Cues filtered out, cues filtered in: computer-mediated communication and relationships'. In Knapp M.L. & J. A. Daly (Eds.), *Handbook of interpersonal communication*. Thousand Oaks, CA: Sage: 529-563.
- Weedon, C. (1997). *Feminist Practice and Poststructuralist Theory*. 2<sup>nd</sup> edn. Oxford: Blackwell.

## Biography

**Therese Örnberg Berglund** is a doctoral student at Umeå University, Sweden, affiliated with both the Department of Modern Languages/English (<http://www.eng.umu.se>) and the humanities computer lab HUMlab (<http://www.humlab.umu.se>). Her research project deals with emerging communication patterns, and she is also interested in questions to do with ICT and (language) education. Therese coordinates the online activities of a Swedish national network for ICT in academic language education, ITAS (<http://www.humlab.umu.se/itas>), and is the national representative for EUROCALL in Sweden (<http://www.eng.umu.se/eurocall>). She also has a teacher's degree in German and English for upper secondary school.

Visit Therese's blog, Emerging Communications, for more information:  
<http://emergingcommunications.net>.

### Author's address:

*Therese Örnberg Berglund  
Department of Modern Languages/English  
Umeå University  
901 87 Umeå  
Sweden  
phone: +46 90 786 61 58  
e-mail: [therese.ornberg@engelska.umu.se](mailto:therese.ornberg@engelska.umu.se)*

# SECOND NORDIC CONFERENCE ON MULTIMODAL COMMUNICATION

## Contributors:

Dominic W. Massaro  
Isabella Poggi

---

Jonna Ahti  
Jens Allwood, Elisabeth Ahlsén, Johan Lund and Johanna Sundqvist  
Jens Allwood and Nataliya Berbyuk  
Jens Allwood, Loredana Cerrato, Kriistina Jokinen, Patrizia Paggio  
& Costanza Navarretta  
R. Atladottir, J. Gay, K.L. Jensen, R.B. Jensen, I. Kun, L.B. Larsen, S. Larsen  
Tom Brøndsted  
Loredana Cerrato  
Loredana Cerrato and Gunilla Svanfeldt  
Fang Chen  
Pierre Gander  
Mia Heikkilä  
David House  
Sari Karjalainen  
Knut Kvale, Narada Warakagoda, and Marthin Kristiansen  
Ann-Christin Månsson  
Gunilla Svanfeldt, Preben Wik & Mikael Nordenberg  
Anna-Lena Rostvall and Tore West  
Karen Woodman  
Therese Örnberg Berlund

Series: Gothenburg Papers in Theoretical Linguistics  
Dept of Linguistics, SSKKII, Göteborg University

ISSN 0349-1021



Printed in Sweden  
Göteborg University  
Göteborg, 2006