NEALT PROCEEDINGS SERIES
VOL. 15

Proceedings of the 3rd Nordic Symposium on Multimodal Communication

May 27-28, 2011
University of Helsinki
Finland

Editors
Patrizia Paggio
Elisabeth Ahlsén
Jens Allwood
Kristiina Jokinen
Costanza Navarretta

NORTHERN EUROPEAN ASSOCIATION FOR LANGUAGE TECHNOLOGY

Proceedings of the 3rd Nordic Symposium on
Multimodal Communication

NEALT Proceedings Series, Vol. 15

ISSN 1736-6305

Published by
Northern European Association for Language
Technology (NEALT)
http://omilia.uio.no/nealt

Electronically published at
Tartu University Library (Estonia)
http://dspace.utlib.ee/dspace/handle/10062/22532

Volume Editors
Patrizia Paggio
Elisabeth Ahlsén
Jens Allwood
Kristiina Jokinen
Costanza Navarretta

Contents

Preface

The articles collected in this volume are a selection of the papers presented at the 3th Nordic Symposium on Multimodal Communication that was held at the University of Helsinki on 27-28 May 2011. The symposium, which was organised by the Nordic project on multimodal corpora NOMCO (http://www.sskkii.gu.se/nomco/), and funded by the NOS-HS NORDCORP programme, is the latest event in a series of Scandinavian symposia and workshops dedicated to multimodal communication that was initiated more than a decade ago. The list includes the Swedish symposia on multimodal communication held in 1997, 1998, 1999 and 2000, the two Nordic symposia on multimodal communication held in Copenhagen in 2003 and Gothenburg in 2005, and the workshop at the 2009 NODALIDA conference in Odense. Following this tradition, the Helsinki symposium aimed to provide a forum for researchers from different disciplines who study multimodality in human communication as well as human-computer interaction.

A number of the studies presented at the symposium and published in this volume have been carried out under the auspices of the NOMCO project, and deal with the corpora of first acquaintance conversations in various languages developed and annotated as part of the project. The remainder of the papers, however, provide additional perspectives through a wide choice of topics including the analysis of listener responses, speaker clustering, or multimodal behaviour in aphasics. They address a range of communication situations and languages, and make us of quantitative as well as qualitative analysis methods.

The paper on co-activation by Allwood and Lu investigates the issue of multimodal behaviour adaptation in face-to-face communication. The authors look especially at repetition and reformulation in two Chinese-Chinese and two Chinese-Swedish first acquaintance conversations, and find that the more similar conversational participants are in terms of ethnic, gender and linguistic terms, the more co-activation takes place.

The study by Berbyuk-Lindström also addresses the cross-cultural dimension by analysing recordings of medical consultations between Swedish patients and Swedish or foreign doctors. In particular, the author looks at linguistic repetitions and reformulations. She finds that the foreign physicians use more repetitions and reformulations than their Swedish colleagues when interacting with Swedish patients. Thus, her results partly disconfirm the conclusions in the Allwood and Lu paper on co-activation. The question is, of course, whether the difference is due to the two very different communication situations.

Jokinen and Pärkson deal again with the way in which conversation participants attune their behaviour to one another. The topic of the paper is alignment of gestural behaviour and repetition of words or syntactic patterns across participants in three party conversations in Estonian. The authors note that the presence or absence of synchrony and repetition reflects the level of agreement and cooperativeness among participants.

Boholm and Lindblad analyse Swedish speakers in first acquaintance conversations, in particular the relation between words, prosody and head movements in Swedish interactions, and find systematic relations between certain word tokens or prosodic features and accompanying movements. The study also finds interesting regularities in the temporal alignment and mutual duration of words and nods.

Also the paper by Paggio and Navarretta explores multimodal characteristics of first acquaintance conversations, this time in a Danish linguistic context, and focuses in particular on the way feedback is expressed in words and gestures. It is shown that all modalities, i.e. head, face and eyebrows, contribute to the expressions of feedback, with repeated nods and smiles as the most frequent feedback gesture types.

Lu and Allwood look at feedback in Swedish, Chinese, and Swedish-Chinese first acquaintance conversations. On the basis of their mono-cultural and cross-cultural data, they describe similarities and differences between Chinese and Swedish participants in using unimodal and multimodal feedback.

De Kok and Heylen study multimodal listener behaviour from a number of different perspectives by comparing data from a corpus of listener responses with judgments on response appropriateness on the one hand, and experimentally induced responses on the other. By contrasting the three perspectives, they find that there are moments in which a user response is highly appropriate, inappropriate, controversial or neutral, and that different contextual cues can be used to discriminate these moments. The study is relevant for predictive models of listener behaviour.

The paper by Nishida, Ishikawa and Yamamoto is an example of how certain aspects of conversational behaviour can be modeled. In particular, it addresses the issue of speaker clustering in multi-party conversations, and proposes a method based on the two notions of *speaker subspace* and *phonetic subspace*. The method is quite successful at clustering speakers in a large corpus of conversational Japanese.

Vincze and Poggi provide a very different, largely qualitative analysis of different ways in which blinks and eye-closure are used in a corpus of political debates. Their aim is to describe a number of signal-meaning pairs to be used in the definition of a lexicon of gaze behaviours.

The last two papers look at multimodal behaviour in the context of impaired conditions.

The paper by Ahlsén looks at the relation between speech and gestures in aphasic patients. The communication situation is informal face-to-face interaction, and the data analysed are gesture samples from subjects with and without aphasia. The study points to the fact that gestures in aphasic patients to some extent are affected by the impairment, but also that they can be used to compensate for word finding difficulties.

The study by Fyrberg and Ahlsén, finally, looks at the multimodal communicative ability of a young subject suffering from moderate traumatic brain injury in communicative situations involving one or two interlocutors. The authors show that the adoption of a triangulation of methods, including the analysis of multimodal behaviour together with more conventional neuropsychological and speech assessments, provides a fruitful approach to the diagnosis and treatment of communication impairment after traumatic brain injury.

On behalf of the organising committee,

Patrizia Paggio

Organising committee

Elisabeth Ahlsén
Jens Allwood
Kristiina Jokinen
Costanza Navarretta
Patrizia Paggio

Acknowledgements

We would like to thank the members of the reviewing committee:

Elisabetta Bevacqua, CNRS - Telecom ParisTech
Nick Campbell, University of Dublin
Loredana Cerrato, Acapela Group Sweden
Jens Edlund, KTH Royal Institute of Technology
Marianne Gullberg, Lund University
Dirk Heylen, University of Twente
David House, KTH Royal Institute of Technology
Michael Kipp, DFKI Germany
Brian MacWhinney, Carnegie Mellon University
Isabella Poggi, Roma Tre University
Andrei Popescu-Belis, Idiap Research Institute
Matthias Rehm, Aalborg University
Kari-Jouko Räihä, University of Tampere
Rainer Stiefelhagen, Karlsruhe University
Nadia Mana, Bruno Kessler Foundation

Unimodal and Multimodal Co-activation in First Encounters ---- A Case Study

Jens Allwood
SCCIIL, interdisciplinary Center
University of Gothenburg
Göteborg, Sweden
jens@ling.gu.se

Jia Lu
Div. of Communication and Cognition
University of Gothenburg
Göteborg, Sweden
jia.lu@gu.se

Abstract

In human communication, people adapt to each other and jointly activate behavior in different ways. In this pilot study, focusing on one individual (Cf2) in four interactions two types of co-activation, i.e. repetition and reformulation in two modalities, vocal-verbal and gestural are investigated in two Chinese-Chinese and two Chinese-Swedish video-recordings of university students' first encounters. The aim, on the one hand, is to explore features of co-activation that might be specific to Chinese interactions or common to Chinese-Swedish interactions and, on the other hand, to try to see how one person Cf2 adapts to different strangers. In our analysis, we have considered both culture and gender dependent differences. We find that co-activation is more often unimodal than multimodal, and more often involves gesture than speech. We also find that the more similar interlocutors are regarding cultural/ethnic, linguistic, and gender/biological background, the more co-activation takes place, especially in the form of repetition.

Key Words:

Unimodal, multimodal, co-activation, monocultural, intercultural, Chinese, Swedish, vocal-verbal, gestural, culture, gender, interaction

1 Introduction

There are several different approaches to the area of co-activation in communication. One such approach is based on the hypothesis that so called 'mirror neurons' underlie both the production and the perception of movement (Rizzolatti & Arbib, 1998; Arbib, Bonaiuto & Rosta, 2006). Based on neurological studies of 'mirror movement' (Farmer, 2005; Bhattacharya & Lahiri, 2002) and 'mirror neuron' (Gallese & Lakoff, 2005; Arbib, 2005), mechanisms for acting, perceiving, imitation, and pantomime have been identified (Rizzolatti & Arbib, 1998; Ahlsén, 2008). Other theories concerning what we are calling "co-activation" have been labeled 'behavioral adaptation' (Galegher & Kraut, 1992), 'adaptive response' (Buck, 1984; Burgoon, Stern & Dillman, 1995; Cappella, 1991), 'imitation' (Ahlsén, 2008; Arbib, 2005), bodily coordination (Ivry & Richardson, 2002; Semjen & Ivry, 2001), 'alignment and automatized coordination' (Pickering & Garrod, 2004), and the phenomena considered are usually regarded by the cited authors as a basic and crucial part of human communication and language development. The terms chosen in the mentioned approaches all point to different but probably related aspects of 'bodily coordination'. In this study, we use the term 'co-activation' to refer to the occurrence of similar vocal-verbal and gestural behaviors that occur in different communicators either sequentially or simultaneously, in order to serve the purpose of coordinating human communication. We use the term "gestural" for all visible communicative body movements and the term "vocal-verbal" to distinguish verbal expressions that are vocal from verbal expressions that are gestural, e.g. the gestural words of deaf sign language or the head nods and head shakes used in feedback which we also regard as gestural words.

2 Types of Co-activation

We will take both vocal-verbal and gestural co-activation into account. An interesting part of the relevant behavior consists of communicative feedback (cf. Allwood, Ahlsén & Nivre, 1992; Allwood & Cerrato, 2003; Grammer, Allwood,

Ahlsén & Kopp, 2008). Co-activation can occur vocally through words or phrases, some of which consist of repetitions or reformulations, e.g. B says 'that's all right' after A says 'that's all right' (repetition), or B says 'that's fine' after A says 'that's all right' (reformulation). Co-activation can also occur through gestures; we have coded head movements (down-nod, up-nod, and shake), facial expressions (eyebrow frown, eyebrow rise, gaze up, gaze down, gaze at the other interlocutor, gaze sideways i.e. gaze left or right, smile, scowl (mouth open in a circle, and mouth corners down), posture shifts, shoulder movements (mainly shoulder shrugs), and hand movements as well as through combinations of vocal and gestural behavior, i.e. laughter, chuckle (basically a smile plus a laughing sound with a low pitch and intensity) or giggle (a smile plus a laughing sound with a high pitch and intensity, which are repeated or reformulated, e.g. B smiles after A smiles (repetition), or B chuckles to express friendliness after A has smiled in a friendly way (reformulation). The idea is that a gestural repetition involves use of "the same gesture" in terms of both function and expression, while a gestural reformulation also often involves use of a "similar gesture" and a "similar function". However, the requirement on similarity in function is stronger than the requirement on similarity in expression since, for instance, a negative headshake can be reformulated as a negative hand movement. We admit that as far as reformulations go, the boundaries concerning what is to be regarded as "similar" are somewhat vague both with regard to vocal and gestural expressions and their functions. Operationally, we have tried to restrict what is regarded as similar fairly narrowly to units that serve the same function in a fairly clear sense.

Below, we will use the term "unimodal" for co-activation that is vocal-verbal (only) or gestural (only) and "multimodal" for co-activation that is vocal-verbal plus gestural. In this paper, we restrict our study of co-activation to repetitions and reformulations, while not denying that the concept of co-activation has a wider application.

3. Purpose

This paper primarily investigates three questions. First, what vocal-verbal and gestural behaviors occur in unimodal and multimodal co-activation? Second, are different types of co-activation used in mono-cultural and intercultural interactions? Third, are there any gender differences?

4. Data and Method

The study is based on four video-recordings of face-to-face dyadic dialogs between Chinese and Swedish university students. In order to make a pilot case study of co-activation with respect to differences in culture and gender, one Chinese female subject (Cf2) was studied both in two Chinese-Chinese and two Chinese-Swedish dialogs that varied in the gender of her interlocutors (see Table 1). This allows us to see how the gender of a communicative partner might influence one and the same person (Cf2). Thus, in the mono-cultural interactions, Cf2 was studied with a Chinese female (Cf1) and a Chinese male (Cm1) and in the intercultural interactions, Cf2 was studied with a Swedish female (Sf2) and a Swedish male (Sm2). Since the number of examined recordings is small, a more representative study will require more data.

Recording	*Participants*	*Time Length*	*Language*
Dial.1	Cf2--Cf1	7:00 min.	Chinese
Dial.2	Cf2--Cm1	7:00 min.	Chinese
Dial.3	Cf2--Sf2	7:00 min.	English
Dial.4	Cf2--Sm2	7:00 min.	English

Table 1: The studied video-recordings ***(Note:*** *C=Chinese, S=Swedish, f=female, and m=male.)*

Our study is focused on how strangers who have no earlier acquaintance go about the task of getting to know each other. Each interaction was video-filmed by three video cameras (left-, center-, and right-position) with each interlocutor in a standing position (see Figure 1). The main subject Cf2 was video-recorded four times, and her counterparts Cf1, Cm1, Sf2 and Sm2 were video-recorded once each to provide different adaptation contexts for Cf2. Each video recording lasted approximately seven to ten minutes, but only the first seven minutes were analyzed in detail in the present study.

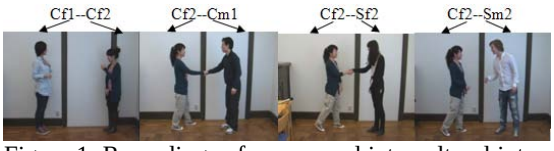

Figure 1: Recordings of mono- and intercultural interactions

The video-recorded data was transcribed and checked according to the GTS (Göteborg Transcription Standard) version 6.2 (Nivre, 1999). To increase reliability, each video recording has one transcriber and two independent checkers. All the video-recordings were manually annotated following the MUMIN multimodal coding scheme (Allwood, Cerrato, Jokinen, Navarretta & Paggio, 2007).

5. Analysis and Results

Below we will now analyze the four recorded dialogs from the perspective of whether the co-activation occurring is multimodal or unimodal.

5.1 Overview

Results concerning co-activation through repetition and reformulation, for all five participants, are presented in Tables 2 and 3. Table 2 shows that there is more unimodal gestural than unimodal vocal-verbal co-activation (171-69), while in contrast, there are only 19 cases of multimodal co-activation, for all participants in the four recordings.

Modality	***Type***	***Total***
Vocal-verbal	Repetition	57
Unimodal	Reformulation	12
	Total	**69**
Gestural	Repetition	111
Unimodal	Reformulation	60
	Total	**171**
Vocal-verbal	Repetition	6
+ Gestural	Reformulation	13
Multimodal	***Total***	**19**

Table 2: Total number of unimodal and multi-modal co-activations (including both Chinese and Swedish participants)

Modality	***Type***	***Mon.***	***Int.***	***Total***
Vocal-verbal	Rep.	12	11	23
(only)	Ref.	2	0	2
	Total	**14**	**11**	**25**
Gestural	Rep.	31	34	65
(only)	Ref.	14	15	29
	Total	**45**	**49**	**94**
Vocal-verbal	Rep.	3	0	3
+ Gestural	Ref.	3	1	4
	Total	**6**	**1**	**7**

Table 3: Cf2's unimodal and multi-modal co-activation *(Mon.=mono-cultural, Int.=intercultural, Ref.=reformulation, Rep.=repetition)*

In addition, we can see (Table 3) that the main subject Cf2 exhibits the same proportions between vocal-verbal and gestural and multimodal co-activation as those observed for the group as a whole (Table 2), but that the differences between Cf2's behavior in the mono-cultural and intercultural situation, are too small to be significant.

5.2 Unimodal Co-activation

In this section, unimodal co-activation i.e. vocal-verbal (vocal-verbal only) and gestural (gestural only) co-activation is studied more in detail.

5.2.1 Unimodal Vocal-verbal Co-activation

Below we will exemplify unimodal vocal-verbal co-activation as it can be observed through repetitions and reformulations. Excerpt 1 shows how the vocal-verbal expression 'wang you' ('turn to the right' in English) is repeated by speaker Cf2, while Excerpt 2 shows how 'hello' is reformulated to 'hi' by speaker Cf2.

Excerpt[1] 1 vocal-verbal unimodal repetition:	
Original Transcription	*Literal English Trans.*
$Cf1: <1 en >1 /// <2 wo men shi wang zuo >2 /// ni men shi **wang you** …	$Cf1: <1 yeah >1 /// <2 we turn to the left >2 /// you **turn to the right** …
@ <1 VFB; CPU confirmation >1 @ <2 VFB; CPU confirmation >2…	
$Cf2: <1 a /// dui dui dui >1 <2 **wang you** >2 …	$Cf2: <1 ah /// right right right >1 <2 **turn to the right** >2 …
@ <1 VFB; CPU confirmation >1 @ <2 VFB; CPU confirmation >2…	
Excerpt 2 vocal-verbal unimodal reformulation:	
$Sf2: **hello** $Cf2: **hi** < | > e1 @ < general face: giggle >, < hand start: Sf2, Cf2 shake hands >	

The vocal-verbal unimodal co-activations can be classified in terms of phrase categories and parts of speech. In Excerpt 1, 'wang you' ('turn to the right' in English) is a verb phrase that is repeated as feedback; in Excerpt 2, 'hello' and 'hi' are both interjections.

[1] The excerpts in this paper are extracted from transcriptions of the studied recordings. In GTS, $ identifies a speaker. Angular brackets < > indicate the scope of a comment, and the number identifies a corresponding comment. The symbol @ initiates the corresponding comment. The number of slashes (/, //, ///) indicate the length of a pause. Curled brackets { } contains letters of a written word form that were not pronounced in the spoken form. < | > indicates that a gesture without vocal-verbal information is inserted in a pause. In our coding, VFB= vocal-verbal feedback, GFB= gestural feedback, CPUE/A= contact, perception, understanding, emotion/attitude.

Feature	Frequency	Examples of repeated expressions
N/NP	37 (65%)	Hobbies; The American idol
V/VP	9 (16%)	Yao qiu *'require';* Hai pa jin qin *'(be) afraid of intermarriage'*
Adj	3 (5%)	Similar
Sentence	2 (4%)	Vad sa du *'what did you say'*
Int	2 (4%)	Hej *'hi'*
Adv	2 (4%)	Just
Pron	1 (1%)	Ta-men *'they'*
Prep	1 (1%)	(Shi) zai *'(be) at'*
Total	57 (100%)	
Relation to FB: 34 repetitions, 60%, are feedback		

Table 4: Grammatical categories of all vocal-verbal unimodal repetitions (The intercultural dialogs, although mainly in English, include a few Swedish expressions)

Table 4 shows the grammatical categories of the unimodal vocal-verbal repetitions; N (noun) and NP (noun phrase) (65%), V (verb) and VP (verb phrase) (16%). We may note that 60% of all the unimodal vocal-verbal repetitions have a feedback function, which indicates that co-activation and feedback are closely connected.

Feature	Frequency	Example
N/NP	5 (42%)	Bei jing *'backgound'* → Gong zuo bei jing *'working background'*
Adj	3 (25%)	Ting hao de *'(it is) very good'* → Bu cuo *'not wrong'*
V/VP	2 (17%)	Guo guo *'pass pass'* → Pass (English)
Pronoun	1 (8%)	I saw it → You saw it.
V/Prep	1 (1%)	Wang you *'(turn) to the right'* → (zai) you bian *'on the right'*
Total	12 (100%)	
Relation to FB: 3 reformulations, 25%, are FB		

Table 5: Grammatical categories of all unimodal vocal-verbal reformulations

Concerning unimodal vocal-verbal reformulations, the most common types are N/NP (42%), Adj (adjective) (25%), and V/VP (17%) (cf. Table 5). 25% of the vocal-verbal reformulations have a feedback function, which again, although weaker than for repetition, shows a link between co-activation and feedback.

We have seen in Table 2 (see also Table 6 below), that there are 57 repetitions and 12 unimodal vocal-verbal reformulations, altogether 69 unimodal vocal-verbal instances of co-activation (produced by both Chinese and Swedish paticipants). Thus, the number of vocal-verbal unimodal repetitions is approximately five times as large as that of vocal-verbal unimodal reformulations.

Vocal-verbal unimodal	*Dial.1*		*Dial.2*		*Dial.3*		*Dial.4*		*Total*
	Cf1	*Cf2*	*Cm1*	*Cf2*	*Sf2*	*Cf2*	*Sm2*	*Cf2*	
Repetition	9	7	10	5	5	3	10	8	57
Reformulation	2	0	3	2	2	0	3	0	12
Total	11	7	13	7	7	3	13	8	69

Table 6: Vocal-verbal unimodal co-activation in the recordings

We have chosen to study the Chinese subject Cf2, varying the gender and/or culture of her interlocutor. Cf2 shows the same tendency as the group as a whole using more unimodal (23) vocal-verbal repetitions than reformulations (2), as can be seen from Table 6. She used roughly the same number of unimodal vocal-verbal repetitions and reformulations in the Chinese mono-cultural interactions (12 (i.e. 7+5) and 2 (i.e. 0+2)) as in the intercultural interactions with the Swedes (11 (i.e. 3+8) and 0 (i.e. 0+0)).

With respect to the gender differences in using unimodal vocal-verbal co-activation, Cf2's interactions are illustrative. As shown in Table 6, Cf2 had slightly more vocal-verbal unimodal co-activation with males (Cm1(13) + Sm2(13)) than with females (Cf1(11) + Sf2 (7)). The number of cases is too small to allow any claim about gender difference in Cf2's interactions with Chinese interlocutors.

Vocal-verbal unimodal	***Dial.1 with Cf1***	***Dial.2 with Cm1***	***Dial.3 with Sf2***	***Dial.4 with Sm2***	***Total***
Repetition	7	5	3	8	23
Reformulation	0	2	0	0	2
Total	7	7	3	8	25

Table 7: Cf2's unimodal vocal-verbal co-activation

However, turning to repetitions and reformulations, in Dialogs 3 and 4 (see Table 7), Cf2 used more unimodal vocal-verbal repetitions with the Swedish male (8) than with the Swedish female (3) and Cf2 did not use any unimodal vocal-verbal reformulations with Swedish interlocutors.

5.2.2 Unimodal Gestural Co-activation

We have found totally 171 instances of unimodal gestural co-activation in all four analyzed dialogs. Of these 111 were repetitions and 60 reformulations (see Table 8).

Gestural unimodal	*Dial.1*		*Dial.2*		*Dial.3*		*Dial.4*		*Total*
	Cf1	*Cf2*	*Cm1*	*Cf2*	*Sf2*	*Cf2*	*Sm2*	*Cf2*	
Repetition	7	20	13	11	13	23	13	11	111
Reformulation	10	5	7	9	8	7	6	8	60
Total	17	25	20	20	21	30	19	19	171

Table 8: Unimodal gestural co-activation in the recordings

Thus, the number of unimodal gestural repetitions is approximately twice as many as that of unimodal gestural reformulations.

Excerpt 3 gestural unimodal smile repetition:	
Original Transcription	*Literal English Trans.*
$Cf2: <1 en /// >1 <2 | >2	$Cf2: <1yeah///>1 <2 |>2
@ <1 VFB; CPU confirmation >1, <1 GFB head: nods; CPU confirmation >1 @ <2GFB general face:**smile**;CPUE/A friendliness>2	
$Cm1: <1 | >1 <2 ou >2 <3 wo shi >3 <4 wo shi <5 hui zu >5 >4	$Cm1: <1|>1<2 oh >2 <3 i am >3 <4 i am (from) <5 hui nationality >5 >4
@ <1 GFB general face: **smile**; CPUE/A surprise/happiness >1 @ <2 VFB; CPU >2…	
Excerpt 4 gestural unimodal reformulation:	
$Cf2: [2 <1 oh >1 <2 yeah similar >2]2 // [3 in the]3 pronunciation [4 <3 // >3]4 // and … @ <3 general face: **giggle** >3 $Sf2: [3 <1 yeah >1 <2 | >2]3 @ <1 VFB; CPUE/A agreement >1, <1 GFB head: nods; CPUE/A agreement R >1 @ <2 GFB general face: **chuckle**; CPUE/A friendliness >2	

Excerpt 3, above shows how a smile is repeated unimodally by Cm1, and Excerpt 4 how Cf2's giggle is reformulated unimodally into a chuckle by Sf2. The unimodal gestural co-activations in Excerpts 3 and 4 are both related to the behavioral group smile/ giggle/ laughter/chuckle which often express friendliness, surprise or happiness, all of which are expectable and fairly common in first acquaintance dialogs.

In general, we have found (see Table 9, below) that unimodal gestural repetitions most frequently involve the following body parts; head (50%), general face (especially smile/ giggle/chuckle/ laughter) (37%), and gaze (6%), and that 69% of the unimodal gestural repetitions have a feedback (FB) function.

Co-activated gestures	**Freq.**	**Example**
Head (nod/ up-nod/ shake/ tilt/ others)	55 (50%)	$Cf2: <1 i'm li yun / <2 nice to meet >2 you >1 … @ <1 hand: **Cf2, Sm2 shake hands** >1 @ <2 GFB head: **Sm2 nod**; CPU >2, <2 head: **nod** >2 $Sm2:…<2i'm jesper>2 @ <2head: **Cf2 nods**>2 $Cf2: < oh > @ < VFB; CPU >, < GFB head: **nod**; CPU >
General face (smile/ giggle/chuckle/laughter)	41 (37%)	
Gaze (up/ down/ side-ways/ around)	7 (6%)	
Posture movement	4 (4%)	
Hand movement	3 (3%)	
Arm movement	0 (0%)	
Total	110(100%)	
Relation to FB: 76 (69%), have a feedback function		

Table 9: Body parts involved in gestural repetition

In Table 10 below, we can see the corresponding figures for gestural reformulation.

Co-activated gestures	**Frequency**	**Example**
General face (smile/ giggle/ chuckle/ laughter)	77 (62%)	$J: <1 yeah >1 it's kin+ i wou{ld} think it's kind of hard for you to <2 understand swedish [49 // >2 <3 elle{r}]49 sevenska >3 @ <1 VFB; CPUE/A agreement >1, <1 GFB **gaze: down**; CPUE/A hesitation O >1 … $L: [49 < (...) >]49 @ < **gaze around** >
Head (nod/up-nod/ shake/ tilt/ others)	17 (14%)	
Gaze (up/ down/ side-ways/ around)	13 (10%)	
Hand movement	8 (6%)	
Posture movement	8 (6%)	
Arm movement	2 (2%)	
Total	125(100%)	
Relation to FB: 71 raw frequencies, 57%, are FB		

Table 10: Body parts involved in unimodal gestural reformulation

Unimodal gestural reformulation is most frequently facial (especially smile/ giggle/ chuckle/ laughter) (62%), head (14%), and gaze movement (10%) (see Table 10), and 57% of the unimodal gestural reformulations have a feedback (FB) function.

Gestural unimodal	***Dial.1 with Cf1***	***Dial.2 with Cm1***	***Dial.3 with Sf2***	***Dial.4 with Sm2***	***Total***
Repetition	20	11	23	11	65
Reformulation	5	9	7	8	29
Total	25	20	30	19	94

Table 11: Cf2`s unimodal gestural co-activation

Turning back to Cf2, Table 11, above, shows that she used more than twice as many unimodal gestural repetitions (65) as reformulations (29). She further used almost the same number of unimodal gestural repetitions and reformulations with Chinese as with Swedish interlocutors: Repetitions; Chinese 31 (i.e. 20+11)) and Swedes

34 (i.e. 23+11); Reformulations; Chinese 14 (i.e. 5+9) and Swedes 15 reformulations (i.e. 7+8).

Concerning gender differences, Cf2 used roughly twice as many repetitive gestures when she interacts with females (43) as with males (22), irrespective of culture (cf. Table 11) and she used slightly more unimodal gestural reformulations with males than with females (as 9 to 5 in monocultural dialogs, and 8 to 7 in intercultural dialogs). That is, in both mono-cultural and intercultural interactions, Cf2 had more unimodal gestural repetitions with females and slightly more unimodal gestural reformulations with males.

5.3 Multimodal Co-activation

We now turn to multimodal co-activation. As can be seen from Table 12, there are totally 19 instances of multimodal co-activation, including both Chinese and Swedish subjects.

Multimodal V+G	*Dial.1*		*Dial.2*		*Dial.3*		*Dial.4*		*Total*
	Cf1	*Cf2*	*Cm1*	*Cf2*	*Sf2*	*Cf2*	*Sm2*	*Cf2*	
Repetition	0	2	1	1	1	0	1	0	6
Reformulation	1	0	1	3	1	1	6	0	13
Total	1	2	2	4	2	1	7	0	19

Table 12: Multimodal co-activation (*V+G=vocal-verbal+gestural)*

Of these, 6 are multimodal repetitions (see Excerpt 5) and 13 reformulations (see Excerpt 6, below). Thus, the number of multimodal reformulations is approximately twice as many as that of the multimodal repetitions.

Excerpt 5 multimodal repetition:	
$Sm2: we <1 call it >1 <2 **peking** >2 @ <1 general face: Cf2 chuckle >1 @ <2 name: city >2, <2 **smile** >2 $Cf2: <1 | >1 <2 yeah >2 <3 **peking** >3 [5 //]5 <4 en >4 // and u1… @ <3 VFB; CPU confirmation >3, <3 GFB general face: **smile**; CPUE/A friendliness O >3, <3 name: city >3	
Excerpt 6 multimodal reformulation:	
Original Transcription	*Literal English Translation*
$Cm1: < **hai** >	$Cm1: < **hi** >
@ < **right hand shake** >, < **smile** >	
$Cf2: < **hai ni hao** >	$Cf2: < **hi hello** >
@ < **right hand shake** >, < **smile** >	

In Excerpt 5, the multimodal unit, 'peking' + a smile, is repeated by speaker Cf2. In Excerpt 6, the multimodal unit 'hai' ('hi' in English) plus handshake and smile, is reformulated by speaker Cf2 into 'hai ni hao' ('hi/ hello' in English) plus a handshake and smile.

Returning to Cf2, she did not repeat or reformulate multi-modally very often in either monocultural or intercultural interactions. In both types of dialog, she had a similar number of multimodal reformulations (4) and multimodal repetitions (3). See Table 13, below.

Multimodal V+G	***Dial.1 with Cf1***	***Dial.2 with Cm1***	***Dial.3 with Sf2***	***Dial.4 with Sm2***	***Total***
Repetition	2	1	0	0	3
Reformulation	0	3	1	0	4
Total	2	4	1	0	7

Table 13: Dynamic features of multimodal co-activation made by Cf2

She used slightly more multimodal repetitions and reformulations with the Chinese (6) than with the Swedish (1) interlocutors: Repetitions; 3 (i.e. 2+1) versus 0 (i.e. 0+0) and Reformulations; 3 (i.e. 0+3) versus 1 (i.e. 1+0). That is, Cf2 used slightly more multimodal co-activation in monocultural interactions (6) than in intercultural interactions (1).

With respect to the possible influence of gender, when interacting with Cf2, males used more multimodal co-activation than females (Cm1 had 2 and Cf1 had 1; Sm2 had 7 and Sf2 had 2). Cf2 used roughly the same number of multimodal repetitions with the Chinese female (2) and the Chinese male (1); however, she used slightly more multimodal reformulations with the Chinese male (3) than with the Chinese female (0). In the intercultural interactions, Cf2 used roughly the same number multimodal reformulations with the Swedish female (with a frequency of 1) as with the Swedish male (0). Cf2 did not use any multimodal repetitions with the Swedish interlocutors at all.

6. Discussion

In section 5, we have found more unimodal co-activation instances than multimodal ones (approximately 12 times as many) in the examined recordings. Possibly this indicates that co-activation in human communication is more unimodal than multimodal. We also found that unimodal gestural co-activation was twice as common as unimodal vocal-verbal co-activation. This possibly shows that co-activation in human communication is more dependent on gestures than on speech. In addition, we found that multimodality plays a relatively less important role

than unimodality for co-activation in the first encounters we have studied.

Both Chinese and Swedish participants used more unimodal vocal-verbal and gestural repetitions than unimodal reformulations in their co-activation. This may be an automatic effect of 'mirror neurons', or because in first encounters interlocutors repeat each other's vocal-verbal information, in order to confirm whether they have perceived and understood the information correctly. Both Chinese and Swedish subjects used more multimodal reformulations than multimodal repetitions, possibly because it is more difficult to repeat complex multimodal units of behavior. Unimodal behavior may be easier to repeat, especially vocal-verbal unimodal behavior; whereas, multimodal behavior is more difficult to repeat but easier to reformulate.

We found that both vocal-verbal and gestural unimodal co-activation occurred more frequently with the males than with the females when they were interacting with the Chinese female Cf2, in both mono-cultural and intercultural interactions. Specifically, we found that the males used more unimodal gestural repetition than the females, when interacting with Cf2. Possibly, this is because males are less socially elaborating than females, repeating more and reformulating less.

We have also observed what parts of speech or what parts of the body were involved in unimodal vocal-verbal or gestural co-activation. We found that nouns or noun phrases and verbs or verb phrases comprise most of the unimodal vocal-verbal co-activation, and that more than half of them have a feedback function. Possibly this is because nouns and verbs mostly provide the core of the topic being talked about, and feedback is needed for managing and keeping the interaction going. Further, we found that head, general face (especially smile, chuckle, giggle, laughter), and gaze movements are the most common unimodally co-activated gestures. This may be, because head and face are central in human interaction, so that people attend and react more to the information carried by head movements and facial expressions. For instance, they often try to be friendly in a first encounter and therefore smile or laugh, or they express emotional rapport, hesitation/uncertainty, and/ or interest through gaze movement. Again, more than 50% of the unimodal gestural co-activation has a feedback function, which indicates that giving and eliciting feedback plays a very important role in co-activation in human communication.

If we turn to features that might be specifically Chinese, Cf2 exhibited slightly more vocal-verbal and multimodal co-activation in the mono-cultural interactions than in the intercultural interactions, but more unimodal gestural co-activation in the intercultural ones (cf. table 3, above). The reason for this might be that she felt more comfortable with the other person's vocal-verbal behavior when both of them come from the same cultural and linguistic background, not least for reasons of automatic linguistic proficiency. Perhaps this makes vocal-verbal co-activation easier in mono-cultural interactions, and gestural co-activation, relatively speaking, more comfortable in intercultural interactions.

Cf2 used more unimodal gestural repetition with the same gender and more unimodal gestural reformulation with the other gender in both mono-cultural and intercultural interactions. The reason could be that it is easier to repeat gestural behavior from persons of the same gender. It may be that the more similarities interlocutors share in cultural and biological background, the more repetitions they produce.

7. Limitation of research

Our study has some limitations. First of all, since there are only two Chinese-Chinese mono-cultural and two Chinese-Swedish intercultural interactions, involving two Chinese females, one Chinese male, one Swedish female and one Swedish male, the preliminary results and conclusions are all very tentative.

Second, the results based on Cf2 may be dependent on Cf2 as an individual, and other results may be activity dependent. This necessitates further studies in the future.

Third, Cf2 was video-recorded four times. This means that Cf2 had more experience in the later recordings, and to some extent she was used to communicating with a stranger before a video camera.

Fourth, this pilot study focuses on a small number of Chinese overseas and Swedish native university students in first encounters. So it is unclear to what extent it can be regarded as repre-

senting the general Chinese features of unimodal and multimodal co-activation.

8. Conclusions

The aim of this study was to explore the following research questions: What are the features of co-activation with strangers in vocal-verbal and gestural behavior? Do interlocutors use different types of co-activation in mono-cultural and intercultural interactions? Are there any gender influences?

Because our study is small in size, below are only some suggestions and tendencies that can be seen in our data. Concerning the Chinese female participant Cf2's co-activation in mono-cultural and intercultural interactions, she had slightly more unimodal vocal-verbal and multimodal co-activation in mono-cultural than in intercultural interactions but for unimodal gestural co-activation the difference went in the other direction and since the differences, in any case, were too small to be significant, we do not really have an answer to the question of whether interlocutors use different types of co-activation in mono-cultural and intercultural interactions.

Second, Cf2 used more unimodal gestural repetitions with the same gender in both mono-cultural and intercultural interactions. She also used more multimodal repetitions with the same gender in mono-cultural interactions. This suggests that it is easier for an interlocutor to repeat gestural unimodal and multimodal behaviors when the gender of the interlocutors is the same, possibly for biological reasons. It also supports the view that the more similarities interlocutors share in cultural/ethnic, linguistic, and gender/biological background, the more co-activation is possible.

We also found some common trends for Chinese and Swedish interlocutors. First, unimodal gestural co-activation was more common than unimodal vocal-verbal co-activation, which points to easier access to gestures than to speech or to a greater role for the visual modality than for the auditory modality in co-activation. Multimodality, thus, seems to play a relatively less important role in co-activation, at least in the first encounters we have studied. Second, both Chinese and Swedish interlocutors used more unimodal vocal-verbal and gestural repetitions than unimodal reformulations, but they used more multimodal reformulations than multimodal repetitions. Some possible explanations for this could be that they are making a conscious effort at giving vocal-verbal confirmatory feedback on perception and understanding, or that they are reacting as a result of unconscious mechanical effects of 'mirror neurons'. Another possibility is that it is more difficult to repeat multimodal unit of behaviors, at least in a first encounter. These all necessitate further study.

It was also found that nouns, verbs, and feedback expressions comprised most of the vocal-verbal unimodal co-activation; head, general face (especially smile, chuckle, giggle, laughter), and gaze were the most common unimodally co-activated gestures. This may be because nouns and verbs often are centrally related to the topic, and feedback is used for managing interaction; head and face attract more attention in human interactions, and interlocutors try to be friendly in first encounters or express emotional rapport, hesitation/ uncertainty, and/ or interest through gaze movement.

Males used more vocal-verbal unimodal co-activation and more gestural unimodal repetition but less gestural unimodal reformulation than females in both mono-cultural and intercultural interactions. We speculate that the reason for this might be that males are less socially elaborating than females.

Since our data and activity variation are quite limited, further research is needed to attempt generalizations about cultural and gender differences. This pilot study can therefore mostly contribute to a general description of how people adapt to others through co-activation of vocal-verbal and gestural unimodal and multimodal behavior.

Acknowledgement:

We express our gratitude for support to the VR (Swedish Research Council) project "Återkopplingsprocesser" and to the NOS-HS (Nordic Research Council) project NOMCO. We would also like to thank Elisabeth Ahlsén for discussions and comments and Alexander Holender and Yansi Xu for reliability checking work done at the SCCIL Interdisciplinary Research Center at the university of Gothenburg (Sweden). Finally, we thank our reviewers for constructive comments.

References

Allwood, J., Ahlsén, E., Nivre, J. (1992). On the Semantics and Pragmatics of Linguistic Feedback. *Journal of Semantics* 9, 1-26.

Allwood, J. & Cerrato, L. (2003). A Study of Gestural Feedback Expressions. In P. Paggio, K. Jokinen & A. Jönsson (eds) *First Nordic Symposium on Multimodal Communication.* ISSN 1600-339X, 7-22. Copenhagen.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C. & Paggio, P. (2007). The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing. In J. C. Martin et al. (eds) *Multimodal Corpora for Modelling Human Multimodal Behaviour*. Special issue of the International Journal of Language Resources and Evaluation. Springer. Vol.41, no.3-4, pp.273–287.

Ahlsén, E. (2008). Embodiment in communication-aphasia, apraxia and the possible role of mirroring and imitation. *Clinical Linguistics and Phonetics*, April-May 2008; 22 (4-5): 311-315.

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

Arbib, M. A., Bonaiuto, J. & Rosta, E. (2006) The mirror system hypothesis: From a macaque-like mirror system to imitation. *In Proceedings of the 6th International Conference on the Evolution of Language*, 3--10.

Bhattacharya, A. & Lahiri, A. (2002). Mirror Movement in Clinical Practice. *Indian Academy of Clinical Medicine.* Vol. 3, No. 2, 177-81.

Buck, R. (1984). *The communication of emotion*. New York: Guilford Press.

Burgoon, J. K., Stern, L. A. & Dillman, L. (1995). *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge, UK: Cambridge University Press.

Cappella, J. N. (1991). The biological origins of automated patterns of human interaction. *Communication Theory*, 1, 4-35.

Farmer, S. F. (2005). Mirror movements in neurology. *Neurol Neurosurg Psychiatry* 76, 1330. Online ISSN 1468-330X.

Galegher, J. & Kraut, R. E. (1992). Computer-mediated communication and collaborative writing: media influence and adaptation to communication constraints. *CSCW '92 Proceedings of the 1992 ACM conference on Computer-supported cooperative work*, ISBN:0-89791-542-9, 155-162.

Gallese, V. & Lakoff, G. (2005). The brain's concepts: the role of the sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology*, 22 (3/4), 455–479.

Grammer, K., Allwood, J., Ahlsén, E., Kopp, S. (2008). A Framework for Analyzing Embodied Communicative Feedback in Multimodal Corpora. *JLRE (Special Issue on Multimodal Corpora)*. J.C. Martin (ed.) No. 66466.

Ivry, R. B. & Richardson, R. E. (2002). Temporal control and coordination: The multiple timer model. *Brain and Cognition*, 48, 117–132.

Kolb, B. & Whishaw, I. Q. (2003). *Fundamentals of Human Neuropsychology*. New York: Freeman.

Nivre, J. (1999) *Göteborg Transcription Standard. Version 6*.2, pp. 38. Göteborg: Göteborg University, Department of Linguistics.

Pickering, M. & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27, 169-225.

Rizzolatti, G. & Arbib M. A. (1998). Language within our grasp. *Trends in Neurosciences*, 21,188-194.

Semjen, A., & Ivry, R. B. (2001). The coupled oscillator model of between-hand coordination in alternate handtapping: A reappraisal. *Journal of Experimental Psychology: Human Perception & Performance*, 27, 251–265.

Use of other-repetitions/reformulations as feedback by foreign and Swedish physicians in medical consultations

Nataliya Berbyuk Lindström, PhD
Department of Applied IT
Chalmers and University of Gothenburg
berlinds@chalmers.se

Abstract

In medical consultation, understanding between physician and patient is essential for the quality of the care. Confidence in understanding is especially important in intercultural medical consultations as language problems and cultural differences may cause problems in interactions.

This study presents an analysis and comparison of how foreign and Swedish physicians use repetitions and reformulations of their patients' utterances in order to indicate and check understanding. The analysis is based on 63 recordings of medical consultations (34 foreign physician-Swedish patient and 29 Swedish physician-Swedish patient consultations). Activity-based communication analysis is used to analyze the material.

The results show that the foreign physicians tend to repeat and to reformulate (parts of) their patients' utterances more often than the Swedish ones. Some of the reasons are uncertainty concerning understanding, language factor and consequent increased need to check and "record" information provided by interlocutor compared to native speakers. The fact that those foreign physicians who spent the least time in Sweden produce more repetitions and reformulations may confirm the influence of language acquisition. Furthermore, the native languages of foreign physicians might also have an impact on the frequency of use of this communicative strategy.

1 Introduction

1.1 Foreign physician-native patient communication

While there is a relatively large body of research focusing on native physician - foreign patient communication, little research has been done on the opposite situation, i.e. foreign physician-native patient communication, though foreign physicians are common in many countries, such as USA (Steward, 2003, McMahon, 2004), Australia (Birrell, 2004), the United Kingdom (Swierczynski, 2002, Sandhu, 2005), and Canada (Hall et al., 2004). In the above-mentioned countries, non-native physicians represent between 23 and 28 percent of physicians (Mullan, 2005). In 2009, about 55% of all physicians who were granted medical licenses had been educated outside Sweden (Socialstyrelsen, 2009).

At this moment, few studies have yet reported on foreign physicians and their communication with patients. Such issues as differences in views on doctor-patient relationships and problems with foreign language usage, understanding dialects, colloquial speech and questioning of the quality of physicians' medical education have been raised (Berbyuk Lindström, 2008).

Successful physician-patient communication is important for quality of health care. An essential element in communication is understanding. Showing understanding is "the least one can demand from a cooperative receiver is that he acknowledges apprehension and understanding, so that the sender has a chance of knowing if he has got his information across" (Allwood, 1976). If it is not clear that the information has been understood, checking is necessary to avoid lack of understanding/misunderstanding, missing information, uncertainty, stress and anxiety. It is especially important in intercultural communication,

when language problems and cultural differences often present challenges to interactants.

In intercultural foreign physician-Swedish patient consultations, anxiety and uncertainty of the patients about the physicians' understanding of their problems often together with experiences of pain and suffering is be an unfavorable combination (Berbyuk Lindström, 2008). Thus, the physicians' expression of understanding of what their patients say and verification if they understand their patients correctly are essential factors to ensure the quality of care provided.

1.2 Aim of the study

This study focuses on analysis and comparison of foreign and Swedish physicians use of repetitions and reformulations of the utterances of their patients as a feedback tool for indicating and checking understanding during medical consultations.

2 Background

2.1 Verbal feedback in interaction

Linguistic feedback defined as "linguistic mechanisms which ensure that a set of basic requirements on communication, such as possibilities for continued contact, for mutual perception and for mutual understanding can be met" (Allwood, 2003, p.1). Allwood categorizes into simple feedback units (which consist of one word) such as *yeah* and *mm* and secondary FB units such as adjectives, adverbs, conjunctions, pronouns, verbs and nouns, which may be used for feedback purposes, but which have other important functions in the language as well, for example *good*, *certainly*, etc. Other categories comprise reduplications of simple FB units such as *yeah yeah*; deictic and anaphoric linking (often by reformulating preceding utterances), such as English *I do*, *it is*, Swedish *de e de*, *de gör ja*; idiomatic phrases such as *thank you very much*; and modal phrases such as *I think so*.

Functionally, two primary feedback (FB) functions can be distinguished: FBG (feedback giving or "pure feedback") and FBG/FBE (feedback giving and elicitation). FBG is used to indicate that one is listening to and understanding what the interlocutor says and to express attitude, for example, (dis)agreement, emotions, etc. The FBG/FBE function stands for both showing listening and understanding and checking whether one has heard and understood what the interlocutor said by eliciting a response in the form of confirmation or additional specification.

2.2 Other repetions/reformulations as feedback

Repetitions and reformulations of (parts of) interlocutors' utterances, so-called echo-backchannels (Sugito et al., 2000), allo-repetitions (Tannen, 1989), interactive repetitions/reformulations (Martinovsky, 2001) or other-repetitions (Long, 1981, Svennevig, 2004) have multiple functions in interactions. Sugito *et al*. (2000), in their analysis of Japanese informal conversations, emphasize that repeating what the other speaker says indicates willingness to interact and involvement in the interaction. Perrin *et al*. (2003, p. 1849) present a summary of the functions of repetitions such as a taking into account function, "by which a speaker indicates that what was just said by the interlocutor has been heard and interpreted" (corresponds to Allwood's pure FBG function of repetition); a confirmation request function (signaling a problem related to some aspect of the interlocutor's talk), "by which a speaker seeks confirmation or a specification of what has just been said by the interlocutor" (corresponds to Allwood's FBG/FBE function); a positive reply function, "by which a speaker expresses agreement with the preceding talk of the interlocutor"; and a negative reply function, "by which a speaker expresses disagreement with what the interlocutor has just said" (both are subcategories
of FBG).

Svennevig (2004) shows how other-repetitions are often used to display the receipt of information in interactions between native Norwegian clerks and their non-native clients, pointing out the impact of intonation on the function of repetition, showing that a plain repeat with falling intonation is a display of hearing while a repeat plus a final response particle, *ja* ('yes'), constitutes a claim of understanding. The use of rising intonation can also display emotional stance (surprise or interest) (p. 489).

Allwood (1988) points out that repetitions/reformulations are widely used by language learners as means for feedback giving and elicitation, especially early in acquisition process, since they are "a simple means of feedback giving for the learner who does not have many other means of expression" (p. 277). The use of repetitions/reformulations is observed to decrease over time; they seem to be replaced by primary feedback units. Furthermore, the native speakers in the above-mentioned study produced little repetition compared to the non-native speakers.

The use of repetitions/reformulations depends upon a number of factors, such as a particular speaker's characteristics, activity type and how common the use of repetitions/reformulations for feedback giving/eliciting is in the speaker's native language. Culture can also be a contributing factor, as Tannen points out: "for individuals and cultures that value verbosity and wish to avoid silences in casual conversation, repetition is a resource for producing ample talk, both by providing material for talk and by enabling talk through automaticity"(Tannen, 1989, p. 48).

The above-mentioned functions of repetitions and reformulations make them both relevant and interesting to investigate in the context of medical consultation. In spite of the apparent scarcity of research on repetitions/reformulations in medical context, their positive impact on communication between physician and patient cannot be overestimated. In his book on communication with patients, aimed at medical students, Bendix (1980) stresses the importance of repeating the patient's last words; among other things, this strategy can encourage the patient to become more open, help to make the issues discussed clearer, and keep both participants interested.

These outcomes are essential for the quality of care. In addition, it might be interesting to see how non-native speakers in a higher position (foreign physicians) than native speakers use this type of feedback to ensure understanding, as well as the possible influence of culture.

3 Methods

3.1 Recordings and participants

Video and audio-recordings for the study were made in health care centers and hospitals in Western Sweden between 2005-2007. The choice of the institutions was influenced by availability of the participants who agreed to participate in the study. The consultations were recorded after obtaining written consent from all involved in the recordings. No researcher was present during the consultations.

Sixty-three (63) recordings are used for this study (34 foreign physician-Swedish patient and 29 Swedish physician-Swedish patient consultations). Total recording time is about 15 hours (about 9 for intercultural and 6 for Swedish consultations). Thirteen (13) foreign and seven (7) Swedish physicians participated in the study.

The majority of foreign physicians come from Hungary (4, Hungarian group) and Iran (5, Iranian group). Other physicians are from Germany, Colombia, former USSR (Russia) and former Yugoslavia. Age range is 34-56 years.

Participant code	Age	Gender	Specialty	Years as physician		Time in Sweden (years)
				in home country	*in Sweden*	
Hungarian group						
HuD1	45	male	anesthesiology	20	1	1
HuD2	34	female		7	1	1
HuD3	36	male		9	1.5	1.5
HuD4	44	male		11	2	2
Iranian group						
IraD5	49	female	geriatrics, rehabilitation	4	10	13
IraD6	40	female	general practice	5	>1	7
IraD7	45	male	surgery	5	13.5	14
IraD8	48	male	ophthalmology	3.5	16	17
IraD9	50	female	obstetrics, gynecology	8	15	18
Mixed group						
GerD10	56	male	orthopedics	30	1	1
ColD11	39	male	surgery	2	10	12
RusD12	45	female	general practice	45	10	14
YugD13	35	female	anesthesiology	>4	>2	2

Table 1: Foreign physicians demographics

Seven Swedish physicians (5 male and 2 female), 4 surgeons and 3 general practitioners, age range 27-52 years have been involved. The patients are native Swedes, aged between 20 up to 89 years.

3.2 Transcription and coding

The recordings of the consultations were transcribed and checked (Allwood et al., 2000, Nivre et al., 2004), the communication was analyzed using activity-based communication analysis (Allwood, 2003). The transcriptions in the article are presented in the Swedish original and an English translation. In the table below, transcription conventions are presented:

Symbol	Explanation
$P, $D,	participant (patient, doctor)
[ ]	overlap brackets; numbers used to indicate the overlapped parts
/, //, ///	short, intermediate and long pause, respectively
+	incomplete word, pause within word
CAPITALS	stress
:	lengthening
< >, @ < >	comments about non-verbal behavior, comment on standard orthography, other actions
< SO: du >	SO stands for standard orthography. The dialectal forms of Swedish and incorrect forms used by the foreign physicians are commented

Table 2: Transcription conventions

An overview of corpus is presented below:

Participant categories	Number of words	Participant categories	Number of words
ICCMedConsult		***SweMedConsult***	
Consultation types: anesthesiology, gynecology, eye, general practice, rehabilitation, intensive care, orthopedics, surgery		*surgery and general practice*	
Foreign physicians	31 037	Swedish physicians	28 727
Hungarian physicians	9 352		
Iranian physicians	12 112		
Mixed physicians	9 573		

Table 3: Corpus

In the coding, I distinguish between repetitions and reformulations. The repetitions and reformulations are divided into those used for feedback giving (FBG) and those used for both feedback giving and eliciting (FBG/FBE). FBG and FBG/FBE are distinguished as follows. Repetitions/reformulations that do not evoke confirmation from the interlocutor in the next utterance are coded as FBG while those that evoke such confirmation are coded as FBG/FBE. In addition, in the case of repetitions and reformulations for FBG, falling intonation is used. When the repeated/reformulated segment is used with interrogative (rising) intonation, it is coded as FBG/FBE. When intonation is interrogative, it encourages the production of feedback from the interlocutor. However, the absence of interrogative intonation does not rule out the production of feedback in the next utterance. Therefore, sequences in which the repeated element is followed by confirmation from another speaker constitute a primary criterion for distinguishing between FBG and FBG/FBE. The repetitions and reformulations produced by the foreign and Swedish physicians were extracted from the transcriptions and analyzed. All the repetitions and reformulations are grouped on the basis of their function into FBG and FBG/FBE categories.

4 Results

4.1 Repetitions and reformulations for feedback giving (FBG)

Both foreign and Swedish physicians use repetitions and reformulations to give feedback, repeating (part of) their patients' answers to their questions to show that they listen to what their patients say. This strategy is also used to "record" new information provided by patient (e.g., a new symptom that might be worth paying attention to). Svennevig (2004) comments that such repeats often occur after statements presenting new (and often specific) information, and can therefore be called "information receipts" (p.490). Declarative intonation is used in these cases, not interrogative. Consider the example below:

	Transcription	Translation into English
$D:	m // men e hade du magblödning eller magsår eller [1 nej inget sånt]1	*m // but er did you have a gastric hemorrhage or a gastric ulcer [1 no nothing like that]1*
$P:	[1 nä nä nä]1 de har ja nog inte haft men ja har haft problem <1 me magen va // [2 att]2 ja har fått ja kan ju inte äta va som helst >1 [3 för då]3 / får ja	*[1 no no no]1 I don't think I've had that but I've had problems <1 with my stomach // [2 see]2 I've got I can't eat just anything >1 [3 because then]3*
	halsbränna å [4 å andra]4 <2 å rapar >2 väldit mycke rapningar	*/ I get **heartburn** and [4 and other]4 <2 and burp >2 a lot of belching*
@ <1 hand gesture: left hand on stomach >1 @ <2 hand gesture: left hand moving up towards the throat >2		
$D:	[2 m]2	*[2 m]2*
$D:	[3 < jaha >]3	*[3 < I see >]3*
@ < head movement: nod >		
$D:	[4 < **halsbränna** >]4	*[4 < **heartburn** >]4*
@ < head movement: nod >		
$D:	jaha // ja // och e är du allergisk mot någonting	*I see // well // and er are you allergic to anything*

Example 1: Heartburn (HuD2)

First, the physician gives feedback using *m* and *jaha* together with a head nod. However, she also nods and repeats the word *halsbränna* ('heartburn'), which constitutes more exhaustive feedback. It is also a way of "recording" a new symptom and marking a concept important for giving a diagnosis. In similar examples from the data, simple feedback items such as *jaha, ja, jaså, okej, mm*, etc., are often combined with nonverbal behavior (e.g., nod, smile, long pause, etc).

Physicians also tend to paraphrase their patients' utterances for the same purpose – to give feedback, show that they are listening and retain information delivered by the patients. Reformulations represented in the data are primarily the result of grammatical and lexical changes. For example, when a physician asks on which side the patient is feeling pain in, the patient answers *i höger* ('in the right'), which is followed by the physician's feedback, *i höger sida // okej* ('in the right side // okay'). Here, the physician reformulates the patient's utterance, adding the word *sida* ('side'), to provide feedback.

A common reformulation type in medical consultation results from a deictic shift of person, which can be explained by the influence of the activity structure: two main participants, physician and patient, are involved in interaction.

Consider the example below:

	Transcription	Translation into English
$D:	du ska opereras idag	*you will have surgery today*
$P:	**m vet [ja]**	***m [I] know***
$D:	**[vet du] m** // har du nån e problem som du vill // prata om	***[you know] m** // do you have any er problem that you want to // talk about*

Example 2: I know (HuD4)

Feedback is used to show contact, perception and understanding, as well as the speaker's attitude. The example below shows a physician who uses reformulation to give feedback and shows his agreement with the patient:

	Transcription	Translation into English
$D:	ha du haft ont i ögat nån gång	*have you ever felt any pain in your eye*
$P:	aldri de bara att / **ja ser dåligt**	*never it's just that / **I have poor eyesight***
$D:	**du ser dåligt me de ögat ja** // å så helt plötslit	**you have poor eyesight in that eye I see** // and then all of a sudden

Example 3: Poor eyesight (SweD2)

In addition to giving feedback by reformulating the patient's utterance *jag ser dåligt* ('I have poor eyesight'), the physician shows his agreement and confirms his awareness of the patient's problem.

Repetitions and reformulations are also used to express emotions such as surprise as in the example below:

	Transcription	Translation into English
$D:	hur har du [mått]	*how have you [been]*
$P:	[ja] allså nu kan ja ju tala om att ja har gått ner ungefär **tjufem kilo** i vikt / från å me förra året //	*[well] now I can tell you that I've lost about **twenty five kilos in weight** / since last year*
$D:	**tjufem kilo** / de e mycke de	***twenty-five kilos** / that's a lot*
$P:	a:	*yeah*

Example 4: Twenty-five kilos (SweD5)

The physician gives feedback of understanding and expresses his surprise about the patient's weight loss by repeating part of her utterance.

To summarize, foreign and Swedish physicians use repetitions and reformulations of their patients' utterances (often answers to the physicians' questions) for feedback purposes (i.e., to show attention and understanding, as well as to express emotions, agreement, etc. Repetitions and reformulations are also a tool used to "record" the information provided by the patients and to elicit confirmation from them.

4.2 Repetitions and reformulations for feedback giving and feedback elicitation (FBG/FBE)

In addition to using repetitions and reformulations just to give feedback, the physicians use them to simultaneously give and elicit feedback (FBG/FBE). Consider the example below from an interaction between an Iranian male physician and his Swedish patient:

	Transcription	Translation into English
$D:	i vilket öga tar du droppar	*in which eye do you take drops*
$P:	< **vänster** >	< ***left*** >
@ < hand gesture: left hand pointing at left eye >		
$D:	**vänster**	***left***
$P:	**ja**	***yeah***
$D:	e höger har du inga [droppar]	*er right you don't use [drops]*
$P:	[nej] nej // ja tar en på / moron å två på kvällen	*[no] no // I take one in / the morning and two in the evening*

Example 5: Left eye (IraD9)

The patient answers the physician's question, and the physician repeats that answer (*vänster* ['left']). The patient's next utterance is a simple feedback item *ja* ('yes'), confirming the information he has already provided, which the physician was attempting to check correct receipt of by using repetition. As we can see, the repetition here serves not only to show that the physician is listening and remaining involved, but also to check that the information has been understood correctly. The repetition in the example above does not have interrogative intonation, whereas other cases presented in the data do. As I mentioned earlier, interrogative intonation encourages the interlocutor to produce a confirmation in the next utterance. Furthermore, the feedback provided may be limited to a simple feedback unit (as above), but it can also be combined with more detailed information:

	Transcription	Translation into English
$D:	< okej > [va e de för fel]	*< okay > [what's the problem]*
$P:	[både fysist] och psykist	*[both physically] and psychologically*
$D:	mestadelen > alltså	*< mostly> that is*
$P:	**både och**	*both*
$P:	**< både och >**	*< both >*
@ < head movement: nods >		
$P:	**ja** e: < > fysist e att ja ö e ja tror ju personlien ja har inte ja har inte sett röntgenbilderna	***well** er < > physically it's that I er er why personally I think I haven't seen the X-ray pictures*
@ < hand gesture start: left hand on right shoulder >		

Example 6: Both (IraD8)

The patient states that he feels bad both physically and psychologically (*både och* ('both')). This is repeated by the physician and is followed by the patient's detailed explanation of why he feels bad (both non-verbally by putting his hand on the shoulder where the pain is localized and by expressing his anxiety).

Reformulations are also used to both give and elicit feedback. This is exemplified by an excerpt from an interaction between a Russian female physician and her male patient:

	Transcription	Translation into English
$D:	då får vi se / ja ska ta / blodtrycket för att lyssna på hjärtat // men du e duktig / du RÖR på dej / du springer till < buss+ > bussen	*let's see then / I will measure / your blood pressure to listen to your heart // but you are doing well / you EXERCISE / you run to the < bus+ > bus*
@ < cutoff: bussen/the bus >		
$P:	nä: nu // ja gå till bussen	*why now // I walk to the bus*

$D:	**du går till bussen**	***you walk to the bus***
$P:	**ja springer gör jag inte**	***yeah I don't run***
$D:	för vadå	*why*
$P:	va	*what*
$D:	varför då varför inte	*why why not*
$P:	**nä: ja orkar inte**	***no I don't have the strength***
$D:	**de du orkar inte**	***you don't have the strength***
$P:	**nä det e va vet du / det får så ont i fötterna**	***no it's you know / my feet hurt so much so then***

Example 7: Bus (RusD18)

As we can see, a misunderstanding that has occurred earlier in the conversation – the physician assumes that the patient runs to the bus whereas actually he walks – results in the physician complimenting her patient: *du e duktig / du RÖR på dej / du springer till < buss+ > bussen* ('you are doing well / you EXERCISE / you run to the < the bus+>'). When the patient denies this, saying *jag går till bussen* ('I walk to the bus'), the physician uses reformulation (deictic shift of person) with an interrogative intonation, *du går till bussen* ('you walk to the bus?'), to make sure she understands the patient correctly. The patient confirms it (*ja springer gör jag inte* ['yeah, I don't run']) and expresses his reason for not doing so (*nä jag orkar inte* ['no, I don't have the strength]) in response to the physician's question (*varför då varför inte* ['why, why not?']). Here, by repeating her patient's utterance, the physician is again checking to make sure she understands him correctly.

Both foreign and Swedish physicians use repetitions and reformulations of their patients' utterances to give feedback and make sure they have understood information correctly, eliciting confirmation from the patients.

5 Results: Quantitative analysis

The occasions when the physicians use repetitions and reformulations for FBG and FBG/FBE were counted; the numbers are expressed in parts per million (PPM). To verify the significance of differences, $\chi 2$ tests were used.

Participant category/type	**Foreign physicians**				**Swedish Physicians**			
	FBG		FBG/FBE		FBG		FBG/FBE	
Type rep/ref	rep	ref	rep	ref	rep	ref	rep	ref
Total per category	4830	1640	1579	1382	1184	627	174	313
Total rep+ref:	6470		2961		1811		487	

Table 4: Repetitions and reformulations used by physicians and patients in PPM[1]

1 PPM is determined as follows: number of occurrences of repetitions/reformulations ÷ number of tokens for the participant category (foreign physicians = 31,037 and Swedish physicians = 28,727) x 1,000,000.

The foreign physicians produce more repetitions and reformulations than the Swedish physicians for both FBG (total rep+ref FBG: 6,470 vs. 1,811, $\chi 2 = 51.92$ [df = 1], $p < .001$) and FBG/FBE (total rep+ref FBG/FBE: 2,961 vs. 487, $\chi 2 = 37.88$ [df = 1], $p < .001$).

Looking at the data for the different cultural groups, the following picture can be observed:

	Hungarian physicians				**Iranian physicians**				**Mixed group**			
Participant category/type	FBG		FBG/FBE		FBG		FBG/FBE		FBG		FBG/FBE	
Type rep/ref	rep	ref	rep	ref	rep	ref	rep	ref	rep	ref	rep	ref
Total per category/type	9078	3631	2136	2350	2310	577	1237	1237	3861	1044	1461	626
Total rep+ref	12709		4486		2887		2474		4905		2087	

Table 5: Cultural groups: repetitions and reformulations in PPM [2]

Repetitions and reformulations are used most by the Hungarian physicians, followed by the Mixed group physicians and then the Iranian physicians.

6 Discussion

The foreign physicians use more repetitions and reformulations of their patients' utterances to give and elicit feedback than the Swedish physicians. This might be related to the greater need for foreign physicians to show their understanding and check the information provided by their patients compared the Swedish physicians, as a strategy to prevent lack of understanding/misunderstanding in communication. It might also be a result of the language acquisition process, confirming what Allwood (1993a) mentions concerning the use of repetitions and reformulations by language learners to give and elicit feedback.

Both foreign and Swedish physicians use repetitions more than reformulations for FBG. However, for FBG/FBE, the foreign physicians use repetitions more than reformulations, while the

2 PPM is determined as follows: number of occurrences of repetitions/reformulations ÷ number of tokens for the participant category (Hungarian physicians = 9,352; Iranian physicians = 12,112, Mixed group physicians = 9,573) x 1,000,000.

opposite is true of the Swedish physicians. One might presume that it is more complicated to paraphrase than to simply repeat, and that the language competence factor might be reflected in the native speakers' tendency to paraphrase more than the non-native speakers. However, there are not enough data to draw any definite conclusions.

Concerning the linguistic and cultural background of foreign physicians, the fact that the Hungarian physicians and the physicians from the Mixed group, who have spent the least time in Sweden, produce more repetitions and reformulations may confirm the influence of language acquisition on the use of repetitions and reformulations. In addition, the foreign physicians' native languages, more specifically how often repetitions/reformulations are used in the foreign physicians' native languages, may influence how they use them in Swedish. Unfortunately, no linguistic studies on this issue for Hungarian, Farsi, Russian, or Bosnian are known to me, so I cannot speculate further on this issue. Concerning German and Spanish, it is worth mentioning that some data on the use of feedback (primarily concerning the use of simple FB words) in these languages (as well as Swedish, Dutch, English, French, Arabic, Finnish, Italian, Punjabi and Turkish) have been presented by Allwood (1993a). As mentioned above, Allwood points out that language learners use repetitions/reformulations for feedback, especially in the initial stages of language acquisition, with a gradual decrease for the majority of learners (but not all) as language acquisition proceeds. It is interesting that speakers who are observed not to decrease their use of repetition for feedback include Finnish and Spanish learners of Swedish, which might indicate the influence of their native languages.

Another point worth mentioning here is that the analysis of the non-native speakers' use of repetitions and reformulations was done in a context in which they are in a superior position to native speakers, which is an uncommon perspective in research. The analysis shows that non-native speakers in a superior position talking to native speakers in a subordinate position use repetitions and reformulations more than native speakers interacting with subordinates of the same linguistic (and cultural) background. In addition, a number of factors have been mentioned that might contribute to the foreign physicians using more repetitions/reformulations for feedback than the Swedish physicians. It is important to add that the fact that the non-native speakers are responsible for the interaction might lead to their using repetitions and reformulations as a more comprehensive type of feedback.

Is there anything in the data that might signal cultural differences? As has already been mentioned, the power distance in Sweden is shorter than in the countries the foreign physicians come from; thus, one can assume that a more paternalistic type of relationship between physician and patient, in which the physician has control over the interaction and core responsibility for the choice of treatment, predominates in those countries. On the contrary, the mutuality type of relationship (more common in Sweden than in the foreign physicians' home countries) presupposes informality and shared responsibility for the interaction; the physician acts as a counselor or advisor (Herlitz, 2003, Berbyuk Lindström, 2008). This difference in the view of the physician's role might result in the foreign physicians' using repetitions and reformulations a good deal in order to show their patients that they have the ability to bear responsibility for the interaction in spite of speaking a foreign language and (possibly) experiencing cultural differences. Repetitions and reformulations represent a way to provide more *exhaustive* feedback than other kinds of feedback. Repeating/reformulating (part of) what the interlocutor says is a clear and powerful way to show that one is listening to and participating in the interaction. This is essential for medical interactions in general, and intercultural medical encounters in particular.

References

Jens Allwood. 1976. *Linguistic Communication as Action and Cooperation.* Göteborg: Department of Linguistics, Göteborg University, Sweden.

Jens Allwood. 1988. *Feedback in Adult Language Acquisition (Final Report Ii).* Ecology of Adult Language Acquisition (ESF).

Jens Allwood. 2007. Activity Based Studies of Linguistic Interaction. In: *Gothenburg Papers in Theoretical Linguistics* (93), Depatment of Linguistics, Göteborg University.

Jens Allwood, Elisabeth Ahlsén, Leif Grönqvist and Magnus Gunnarsson. 2003. Annotations and Tools for an Activity Based Spoken Language Corpus. In J. van Kuppevelt and R.W. Smith, eds., *Current and New Directions in Discourse and Dialogue.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

Torben Bendix. 1980. *Din nervösa patient : Det terapeutiska samtalet: Introduktion till en undersökningsteknik du aldrig fick undervisning i*. Lund: Studentlitteratur.

Nataliya Berbyuk Lindström. 2008. *Intercultural-communication in health care - Non-Swedish physicians in Sweden.* Department of Linguistics. University of Gothenburg.

Robert Birrell. 2004. Australian Policy on Overseas-Trained Doctors. *Med J Aust* 181:635-639.

Pippa Hall, Erin Keely, Susan Dojeiji, Anna Byszewski and Meridith Marks. 2004. Communication Skills, Cultural Challenges and Individual Support: Challenges of International Medical Graduates in a Canadian Healthcare Environment. *Med Teach* 26:120-125.

Herlitz, Gillis. 2003. *Svenskar: Hur Vi Är Och Varför.* Uppsala: Konsultförl./Uppsala Publ. House.

Michael Long 1981. Native Speaker/Non-NativeSpeaker Conversation and the Negotiation of Comprehensible Input. *Applied Linguistics 4*

Bilyana Martinovsky. 2001. *The Role of Repetitions and Reformulations in Court Proceedings: A Comparison of Sweden and Bulgaria.* Göteborg: Department of Linguistics, Göteborg University.

Graham T. McMahon. 2004. Coming to America--International Medical Graduates in the United States. *N Engl J Med* 350:2435-2437.

Laurent Perrin, Denise Deshaies and Claude Paradis. 2003. Pragmatic Functions of Local Diaphonic Repetitions in Conversation. *Journal of Pragmatics* 35:1843.

David Sandhu. 2005. Current Dilemmas in Overseas Doctors' Training. *Postgrad Med J* 81:79-82.

David E. Steward. 2003. The Internal Medicine Workforce, International Medical Graduates, and Medical School Departments of Medicine. *Am J Med* 115:80-84.

Miyoko Sugito, Nagano-Madsen Yasuko and M. Kitamura. 2000. Analysis of Echo Backchannels in a Lively Multi-Speaker Conversation in Japanese. *Fonetik 2000 (The Swedish Phonetics Conference)* 129-133.

Socialstyrelsen (Swedish National Board of Health Care and Welfare). 2008. Statistics. http://www.socialstyrelsen.se/english. Retrieved 20111015

Jan Svennevig. 2004. Other-Repetitions as Display of Hearing, Understanding and Emotional Stance. *Discourse Studies* 6:489-516.

Martha Swierczynski. 2002. Induction Courses for International Doctors. *Bmj* 325:S159.

Deborah Tannen. 1989. *Talking Voices: Repetition, Dialogue, and Imagery in Conversational Discourse*. Cambridge: Cambridge Univ. Press.

Synchrony and Copying in Conversational Interactions

Kristiina Jokinen
University of Helsinki
Helsinki, Finland
University of Tartu
Tartu, Estonia
kristiina.jokinen@helsinki.fi

Siiri Pärkson
University of Tartu
Tartu, Estonia
siiri.parkson@ut.ee

Abstract

This paper describes nonverbal communication in conversations, and focuses especially on the interlocutors' synchrony and copying of each other's behaviour. Synchrony and copying indicate the speakers' cooperation with each other, and manifest in the speakers' use of the same words or similar syntactic patterns in their utterances, adjusting their intonation as well as aligning their nonverbal behaviour. We point out some repeated patterns of nonverbal communication in three-party conversations, and offer some interpretations for them.

1 Introduction

One of the fascinating aspects of human conversations is the accurate timing and coordination of the participants' communicative behaviour. Interlocutors react to each others' actions and alternate their turns in a coordinated manner, and they also tend to anticipate and follow the partner's behaviour so that their communication occurs simultaneously and can be described as synchronous activity. This kind of adaptation of the interlocutors to each other's behaviour is often called alignment (Pickering and Garrod, 2004; Katagiri, 2005). Another term that has been used to refer to synchronous behaviour is that of copying or mimicry, which can range from an unintentional copying of a fellow human to an intentional mimic performance. For instance, Caridakis et al. (2007) talk about copying the human behaviour on a virtual character and especially focus on facial expressions and their expressivity, while Mancini et al. (2007) analyze human body movements in order make the virtual character to respond to the user's expressive behaviour appropriately. In virtual agent interactions, mimicry management consists of the subtasks of perception, interpretation, planning, and animation of the expressions shown by the other person, and it is based on models that represent the user's original expressive behaviour instead of exactly duplicating this.

We can also distinguish synchrony, which functions in a more agent-centred way: although it also requires that the agent has perceived and interpreted the partner's behaviour, it also presupposes that the agent naturally exhibits similar behaviour as the partner: simultaneous behaviour results from the agent's anticipation of the partner's reaction by evaluating the partner's behaviour with respect to the agent's own goals and intentions: synchrony is unconsciously planned rather than intentionally copied from the partner's acting (cf. also Sebanz et al., 2006). The difference between mimicry and synchrony is thus related to the anticipation and coordination of communicative acts: in synchrony, the form of the action originates from the partner's intention to present something in a manner that coincides with the partner's behaviour, while in mimicry only the overt expression of the partner's behaviour is copied.

We have studied synchronous behaviour in three-party conversations and focussed especially on the participants' gestures, body posture, and head movements that occur at the same time. Synchrony can also appear between different communication modalities within a single person, e.g. when one coordinates words with beating gestures, or hand and head movements. However, this kind of intra-partner synchrony is related to the agent's own communication management and has no immediate reference to interaction with the other partner's behaviour, and we will not discuss it here.

In this paper we will focus on inter-partner synchrony, or simultaneous and reciprocal behaviour. Since it signals that the interlocutors are engaged in the interaction and can anticipate the partner's behaviour accurately, we regard syn-

chrony as an indication of the participants' co-operation with each other: the more inclined the participants are to collaborate with their partners, the more synchronous behaviour they show with one another unconsciously. Although the difference between mimicry and synchrony is small on the descriptive level, we aim to distinguish them by referring to intentionality, anticipation and coordination of the speakers' reactions. We say that in mimicry, the speaker synchronizes behaviour in order to produce an affective reaction to the partner's perceived action, but in synchrony, the speaker anticipates a particular behaviour and thus produces spontaneous cooperation, the signal of which is simultaneous similar activity among the partners.

We expect to find a difference concerning the time that it takes for the partner to produce a similar action as the agent, dependent on the time that it takes for the speaker to react. We operationalise the difference by defining the copying behaviour as synchronous activity that has a short time delay with respect to the copied behaviour (due to the time delay in perception, interpretation, planning, and production of an action): the agent copies the partner's gestures, body postures or head movements after a minimum delay of 100ms. It may be difficult to distinguish the two if the delay is a few milliseconds only, and the distinction often depends on the observer's sensitivity to observe the delay too: judgments can vary depending on whether the observer regards the timing of the actions simultaneous or not.

In this paper we describe qualitatively the type of synchrony that occurs between participants, and make a general classification between synchrony and copying by using 100ms as the minimum delay threshold for copying behaviour. We discuss a few examples of synchronous and copying behaviour and try to answer the question if it is possible to distinguish the two in naturally occurring conversations, and if so, which one is more common. In Section 2, we first describe the role and function of gestures and body movement in interactive situations and provide background about the related work. We proceed to describe the data in Section 3, and provide examples of synchrony and copying in Section 4. Finally we discuss some consequences of the work in Section 5, and draw conclusions on the type and function of such behaviour with respect to constructing shared ground in Section 6.

2 Gestures and body movement in interaction

Gestures and body movement have an important role in human communicative behaviour. They are related directly to the information flow of the interactions and they also function in an iconic manner to display the speaker's emotions, attitudes, and mutual relations. They also function on meta-discursive levels (Kendon, 2004; Jokinen and Vanhasalo, 2009), and are used to control and coordinate interaction (Allwood et al., 2007). For example, leaning forward often means interest and leaning backward withdrawing from the conversational situation. Besides displaying the interlocutor's attitudes towards the topic being discussed, body movements can also control interaction by signalling to the partner if they should stop or if they are encouraged to continue further. Such body movements are also used to fill in pauses in conversation: e.g. if the speaker does not want to take the turn, they move backwards. Often the interlocutors also change their position without intending to take the turn in the conversation. They can tacitly state that they are present and have a role in the conversation by adjusting their sitting position appropriately. It is also possible that the body movement is simply related to physical tiredness of staying in a particular position for a long time, but even in this case it can be interpreted as the partner finding the situation uncomfortable and wanting to leave.

Also gaze can control conversations. Gaze signals the speaker's focus of attention and mutual gaze is an important signal in agreeing successful turn-taking (e.g. Jokinen et al., 2009). Gaze may also signal if the speaker wishes to take a turn, or if the turn is offered to another interlocutor (in the latter case, gaze functions in a similar way as pointing, see Fig. 1).

Figure 1 Gaze as a simultaneous pointing device.

In defining communicative gestures and body movement, we follow (Kendon, 2004), who notices that there is a continuum from movements that are perceived as random gesturing to gestures that are understood as communicatively important actions. "Gesture" denotes any possible hand and body movement, but only those which are perceived as communicatively meaningful are *communicative* gestures: potentially all gesturing can be communicatively important if the interlocutors interpret it so. (Sign languages are different in that they form highly structured gesture systems which function by providing abstract representations for communication.)

It is often difficult to assign a clear unambiguous meaning to gesturing and body movements, and often this is not even possible. From the viewpoint of synchronous communication, semantic disambiguation is not necessary since it is not a particular conceptual meaning that is to be conveyed but indication of the partner's collaboration. Any movement can thus function as a starting point for joint gesturing since the partners unconsciously respond to the speaker's gesturing. The speaker also unconsciously reacts to the listener's behavior and would be interesting to study further how the speaker role (the one who speaks) and the contributor role (the one who contributes to conversation) don't necessarily coincide.

If the movements get echoed and amplified by the partner, while the speaker moves back and forth, waves their hands, etc., synchronous behaviour can start. The intuitive nature of such behaviour is often captured in the interlocutors' impressions that it is easy/difficult to talk to the partner: the interlocutors' tacit individual behaviour patterns can either amplify or diminish their joint communicative behaviour, and thus affect their experience of the interaction. To understand what contributes to synchronous behaviour and makes interaction experience pleasant, it is important to investigate how interactions continue and are built up on such movements.

This has an important consequence for synchrony: there are culturally and contextually defined gestures and gesture systems, but only spontaneously elicited gestures that the partner reciprocates can be regarded as truly synchronous in a given situation. Moreover, these gestures can be considered universal in the sense that they are recognized and produced by watching the partner and anticipating their behaviour, without any cultural influence.

Kendon (2004) points out that gestures have a clear peak or stroke, preceeded by a preparation phase, and followed by a post-phase, unlike posture shifts which often are gradual. We define gesture synchrony with respect to the start of the gesture phases: while the length of the speakers' individual preparation phases may vary, it is the timing of the peak that should coincide in their synchronous behaviour.

3 Data

Two videotaped Estonian conversations were used as the basis of our studies. The analysed conversations are altogether about 15 minutes long, and concern three participants talking about plans for a new school building and about inspection of a recently built school building. The situations are role-playing situations, where the participants have adopted the roles of an architect, a school house expert, a town government representative, and a building company representative. The situations are thematically related to each other, i.e. the second conversation is a logical follow-up meeting of the first one, and continuity is supported by two of the participants being assigned the same role in both conversations. Although role-playing may differ from actual situations, it must be noted that people always have a certain role when they are engaged in conversations. Moreover, nonverbal communication and synchrony are mostly unconscious signalling processes, and their conscious modification is not so common; thus nonverbal behaviour may not necessarily differ in role-playing and in spontaneous situations, especially if the participants are familiar with each other as in our case. Since we are not interested in the participants' institutional behaviour, but in their nonverbal communication and synchrony which mostly are unconscious signalling processes, we assume that possible differences between actual and role-playing situations are minimal concerning the purpose of our study.

For the experiment, we manually annotated the behaviour of four individual speakers in the video clips (altogether 15 minutes), and considered dialogue acts, gaze, face, head, turn-taking, feedback and emotion/attitude according to a modified MUMIN scheme (Allwood et al., 2007). Two shorter clips of the same videos were annotated by three annotators and the agreement was measured by Cohen's kappa-coefficient which varied between 40-80% depending on the element. According to the scale proposed by

Rietveld and van Hout (1993), these values correspond to moderate up to substantial agreement. The final annotation is summarized in Table 1 and the relative distribution of different verbal and nonverbal behaviours by the four speakers is shown in Figure 2.

	words	face	gesture	body	all
Sp1	45	213	150	99	462
Sp2	12	75	44	46	165
Sp3	84	242	172	160	574
Sp4	10	127	76	75	278
All	151	657	442	380	1479

Table 1. Statistics of the individual speakers and their behaviour (NV=nonverbal).

There are clear differences between the speakers: Speaker 3 speaks by far the most, and also produces most observable non-verbal communicative acts. Speaker 1 also speaks a lot and is more expressive than Speakers 2 and 4 when it comes to producing facial expressions and hand gestures. The dominance of Speakers 3 and 1 is clearly seen in their gestural and body movements: together they produce more than two thirds of all the observed face, gesture, and body movements. Speakers 2 and 4 speak the least, but differ from each other concerning non-verbal communication: Speaker 2 is the least communicative non-verbally. Synchronous behaviour often occurs between the dominant Speakers 3 and 1, too; this is to be expected as they are the most active in the dialogue.

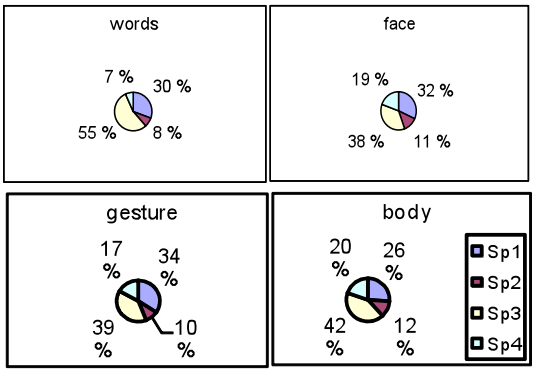

Figure 2. Nonverbal elements in each speaker's behaviour.

When looking at the relative amount of various nonverbal aspects in individual speaker behaviour (Figure 3), we notice that each speaker has majority of the observed nonverbal behaviour encoded in their face and head movement, supporting the fact that the face is an important means that accompanies speech in a visible and obvious manner. It is interesting that the least talkative participants Speaker 2 and Speaker 4 still have more face and head activity than body or hand movements, but that they use their hands relatively more than their body. This is in accordance with the hypothesis that speech and hand gesturing have an intrinsic connection (see Kendon, 2004), while body movements are not so directly related to speaking.

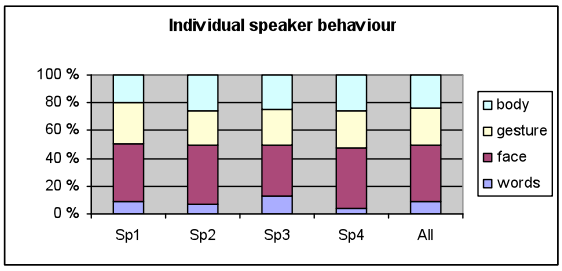

Figure 3. Nonverbal elements in each speaker's behaviour.

4 Synchrony in interaction management

Synchrony and mimicry usually take place between two participants: synchrony that would involve three or more participants seems to be rare, and no such cases appear in our data. In fact, the reason may be obvious as in multi-party dialogues the interlocutors' different roles (speaker, main recipient, side participant) affect their nonverbal behaviour (Battersby, 2011): the interlocutors with different roles react differently to the speaker, and thus it is less probable that their behaviour is synchronised. It must be noted that in two-party dialogues, synchronous situations are not clearly symmetrical either, since one of the partners usually takes the initiative.

Some examples of the synchrony observed in our data are shown in the still-shots in Figures 4-7. Most prominent cases include similar positions with hands crossed (Fig. 4), or hands leaning on the chin (Fig. 5), but also similar body posture (Fig. 6) and the partners' gaze focused on the same object. There are also several examples of beat hand gestures used to emphasize one's arguments (Fig. 7), and the partners copying the behaviour when it is their turn.

Figure 4. Hand-crossing synchrony.

Figure 5. Hand movement synchrony.

Figure 6. Body posture and gaze synchrony.

It is well-known that bursting out to laughing as well as smiling often occur synchronously in smooth conversations (Benus, 2009). Laughing can also create bonds between some participants and leave the others out and thus control the conversation. In our data, for instance, a particular speaker makes a joke which only one of the partners laughs at, and somewhat later, the same speaker makes another joke, which the other partner laughs at. The joking speaker thus seems to control the conversation, and is able to create suitable common grounds so as to engage both partners separately.

Figure 7. Hand beating synchrony.

Similarly to laughing, also nodding often occurs simultaneously, and shows the participants' co-operation and shared understanding. Nodding can also occur as a control signal which directs the participants to talk about certain issues and reach a shared conclusion.

Besides indicating the participants' excitement and reinforcing their experience (positive synchrony), synchrony also occurs when a significant change or communication problem happens in the conversation which the speakers become aware of. In order to restore conversational balance and cooperation, the speakers immediately align their behaviours. For example, when a speaker misremembers a fact (last summer was very hot) and the partner hints at misunderstanding (children do not go to school in summer), the speaker realises his mistake, and in a moment, a synchrony occurs between the speakers.

Synchronized movements often happen at the start of a new topic and at the change of the speaker. For instance, mutual gaze is an example of this. As one of the participants raises gaze to show that he is ready to take the turn, also the partner simultaneously raises gaze and provides feedback about being interested and listening.

Simultaneous gaze aversion usually also indicates the end of a topic or a sequence, and during the moments of silence, all participants look at their papers or the table. The silence can mark the time the participants need to reflect on the topic, or they simply pretend thinking and hope that someone else will take the turn (this seems to be a steady behaviour pattern especially in the conversation video among the male-only partners). However, the breaking of the silence often happens simultaneously.

Speaker's gaze towards the interlocutor can also show that the content of the talk is addressed to the listener or that the listener already has the information (or more information about the issue). In our data, the participants do not often look at each other during the discussion, but they

look at the speaker in the beginning and end of the turns, when giving feedback etc. This kind of gazing behaviour may be related to culture-specific conventions.

Finally, it is interesting that, in the conversations, copying of the partner's behaviour is more common than synchrony. Obviously copying of the partner's hand gestures, head movements, and posture shifts helps to create a common ground but it also implies that the participants can easily follow their partner's behaviour and they do this in order to harmonize with their partner. However, it seems less common to get inspired into such simultaneous activity than the synchrony definition presupposes: this would require that the synchronizing partners truly behave in a similar manner as part of their own presentation.

5 Synchrony, copying and cooperation

As mentioned earlier, we consider synchrony as a sign of cooperation: interlocutors cooperate with each other on several levels. In psycholinguistic and social interaction studies such behaviour has been much studied. We base our analysis on the hypothesis that human–human interaction is cooperative activity which emerges from the speakers' capability to act in a relevant and rational manner (Allwood et al., 2007). The basic enablements of communication, Contact, Perception, and Understanding (CPU) must hold for the interaction to proceed smoothly, and consequently, the agents' cooperation can be said to manifest itself to the extent in which the agents can interpret the partner's feedback, and provide relevant feedback on the CPU enablements. Cooperation can manifest itself as a tight collaboration in order to achieve a particular goal, or as similar behaviour patterns that occur when the interlocutors interact and start to align their behaviour with that of the partner. The agents thus constantly monitor themselves (own communication management, see Allwood, 2001) as well as each other, paying attention to the partner's activities and the communicative situation (interaction management), and if any of the enablements is unfulfilled, react to the problems.

In recent years the number of studies concerning synchrony and alignment has increased, maybe due to the new opportunities to experimentally measure and build computational models for simultaneous behaviour. For instance, Benus (2009) studied rhythmic structure of utterances such as pitch accents and syllables with a coupled oscillator model of Wilson and Wilson (2005), and found weak support for the model. They also found that backchannelling had more salient rhythmical characteristics than other turn-taking events.

In general, we can say that synchrony appears between participants who hold together, while asynchronous behaviour is typical between participants who have a contradiction (or pretended contradiction) against each other. The contradiction could be personal or caused by the participants' roles. Synchronous behaviour builds the common ground among the speakers, but we also note that synchrony can also effectively be used to control flow of communication. The speakers can elicit synchronous behaviour e.g. via jokes and nods, and thus express their own individual wishes and viewpoints which, if reinforced through the partner's copying or synchronous behaviour, can further help to achieve the task goals of the interaction itself.

6 Conclusion and future work

We started with the hypotheses that mimicry and synchrony are signs of cooperation through which the participants reinforce their mutual bonds, agreement, and belonging together. According to our analysed data we can confirm this general view: synchrony and mimicry have their own unique role during conversation, and they are signs of the participants' cooperation.

Synchrony may also have other functions. Further analysis with more data is necessary to study these functions in order to produce solid generalizations. It is also important to investigate whether the results hold for other type of conversational activity than the free chatting.

We assume that gesture management deal with the interlocutors' coordinated action of speaking and listening so that only one of the interlocutors speaks at same time. Natural conversations also contain overlaps and silences which can be signals of excitement, cooperation, or ignorance, i.e. they give feedback about the CPU and about the participants' emotional stance. Usually they are short vocalizations as the speakers take their partners cognitive capability into consideration: it is impossible to get one's message across if the speakers speak at the same time.

Relations between interlocutors in interactive situations are usually expressed directly in words but also through nonverbal behaviour. This study focuses on the patterns of nonverbal communica-

tion which can help understand relations between interlocutors, their cooperation, and alignment with each other. Studies on synchrony may be able to explain how the speakers can convey their ideas and viewpoints to their partners, and how they can reach a shared understanding of the communicative situation: by aligning their behaviours, the speakers can experience similar aspects of the situation for which they otherwise have different viewpoints.

References

Jens Allwood. 2001. The Structure of Dialog. In M. Taylor, D. Bouwhuis and F. Nel (Eds.), *The Structure of Multimodal Dialogue II*, 3–24. Amsterdam, Benjamins.

Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta and Patrizia Paggio. 2007. The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing Phenomena. In Martin, et al. (Eds.). Multimodal Corpora for Modelling Human Multimodal Behaviour. *Language Resources and Evaluation*, 41(3-4):273-287.

Stuart Battersby. 2011. Moving Together: the organization of non-verbal cues during multiparty conversation. *Ph.D Thesis*. Queen Mary University of London, UK.

Štefan Benus. 2009. Are We 'in Sync': Turn-taking in Collaborative Dialogues. 10th Annual Conference of the International Speech Communication Association, Interspeech, Special Session: Active Listening & Synchrony. Brighton, UK.

George Caridakis, Amaryllis Raouzaiou, Elisabetta Bevacqua, Maurizio Mancini, Kostas Karpouzis, Lori Malatesta and Catherine Pelachaud. 2007. Virtual Agent Multi-modal Mimicry of Humans. *Language Resources and Evaluation*, 41(3-4):367-388**.**

Phillip Glenn. 2003. Laughter in Interaction. *Studies in Interactional Sociolinguistics,* volume 18. Cambridge University Press, Cambridge, UK.

Kristiina Jokinen, Masafumi Nishida and Seiichi Yamamoto. 2009. Eye-gaze and Turn-taking. *Proceedings of the third International Universal Communication Symposium, IUCS.* ACM International Conference Proceeding Series. Tokyo, Japan.

Kristiina Jokinen and Minna Vanhasalo. 2009. Stand-up Gestures – Annotation for Communication Management. In C. Navarretta, P. Paggio, J. Allwood, E. Ahlsén and Y. Katagiri. (Eds.). *Proceedings of the NoDaLiDa workshop on Multimodal Communication - from Human Behaviour to Computational Models.* NEALT Proceedings Series, volume 6: 15-20.

Yasuhiro Katagiri. 2005. Interactional alignment in collaborat-ive problem solving dialogues. *Proceedings of the 9th International Pragmatics Conference*. Riva del Garda, Italy.

Adam Kendon. 2004. *Gesture: Visible action as utterance.* New York: Cambridge University Press.

Maurizio Mancini, Ginevra Castellano, Elisabetta Bevacqua and Christopher Peters. 2007. Copying Behaviour of Expressive Motion. *Lecture Notes in Computer Science*, volume 4418. Computer Vision/Computer Graphics Collaboration Techniques. Third International Conference, MIRAGE 2007. Proceedings. 180-191.

Martin J. Pickering and Simon Garrod. 2004. *Towards a Mechanistic Psychology of Dialogue, Behavioral and Brain Sciences*, 27:169-226.

Toni Rietveld and Roeland van Hout. 1993. *Statistical Techniques for the Study of Language and Language Behaviour*. Mouton de Gruyter, Berlin.

Natalie Sebanz, Harold Bekkering and Günther Knoblich. 2006. Joint Action: Bodies and Minds Moving Together. *Trends in Cognitive Science*, 10:70-76.

Margaret Wilson and Thomas P. Wilson. 2005. An Oscillator Model of the Timing of Turn-taking. *Psychonomic Bulletin and Review*, 12(6): 957-968.

Head movements and prosody in multimodal feedback

Max Boholm
SCCIIL (SSKKII)
Department of Applied Information Technology
University of Gothenburg
Gothenburg, Sweden
max.boholm@gu.se

Gustaf Lindblad
SCCIIL (SSKKII)
Department of Applied Information Technology
University of Gothenburg
Gothenburg, Sweden
gustaf.lindblad@sskkii.gu.se

Abstract

The study analyses the relation between words, including their prosodic features, and head movements in communicative feedback, i.e. unobtrusive vocal and gestural expressions which convey information about ability and willingness to continue, perceive, and understand, as well as attitudes and emotions. Examples are words such as *m* and *okay*, and head movements such as nods and shakes. Six recorded first acquaintance conversations in Swedish have been analyzed. Initial direction, repetition, start time, and duration of head movements has been identified by frame-by-frame video analysis. Start time, duration, F0-contour, and pitch of vocal-verbal feedback were analyzed. Main results of the study are: first, multimodal nods more frequently start before or at the same time as words, than words starting before nods. Second, nods have longer duration when produced with words than without. Third, certain words are typically associated with certain nod types, e.g. *okay* with up nods, and *m* with repeated nods. Finally, certain prosodic patterns are more associated with certain nod types, e.g. rising pitch and longer durations with single up nods, and falling or flat pitch with repeated down nods.

1 Introduction

It has often been recognised that gestures can serve to express many of the functions that are known to be expressed by prosody. For example, emphasis of words and phrases in speech can be achieved by both prosodic features and so called "batonic" gestures with hand or head (Bull and Connelly, 1985; Kendon, 2004; McNeill, 1985). Today a growing literature suggests a tight connection between prosodic features of speech and the gestures that accompany speech. This multimodal interplay is sometimes discussed under the terms of *optical phonetics* (Scarborough et al 2009) or *visual* (or *audiovisual*) *prosody* (Graf et al, 2002; Krahmer and Swertz, 2009; Munhall et al, 2004; Swertz and Krahmer, 2010). Words that are made prominent by acoustic means are often accompanied by head and eyebrow movements (Swertz and Krahmer, 2010). Graf et al (2002) report that pitch accents are strongly correlated with accompaniment of head movements. Scarborough et al (2009) found that facial movements were larger and faster with stressed words. Related to these findings that strongly suggest a co-activation of acoustic and gestural means in producing prominence of a linguistic component, Cavé et al (1996) observed a kind of audio-visual isomorphism. They found that the F0 rises were accompanied by raised eyebrows in 71% of the cases (Cavé et al, 1996). It has also been demonstrated that gestural visual cues play an important role for the perception of a word as prominent, and even that gestural accompaniment facilitate speech perception, comprehension and intelligibility as well as the experienced naturalness of Embodied Communicative Agents (see Munhall et al, 2004; Granström and House, 2005; Moubayed et al, 2010). Taking into account the interaction of lexical and prosodic features, as well as timing, have been shown to improve recognition of feedback head movements in human-computer interfaces (Morency et al 2007).

The present study analyses the relation between words, including prosody, and head movements in communicative feedback. Feedback is defined as unobtrusive vocal and gestural expressions used in communication to inform an interlocutor about the ability and willingness to (i) continue the interaction, to (ii) perceive, and

(iii) understand what is communicated, and (iv) in other ways attitudinally and emotionally react (see Allwood, 1988; Allwood et al, 1992; Allwood et al, 2007). Types of vocal-verbal feedback in Swedish include feedback words, feedback phrases, feedback clauses such as *jag förstår* ('I understand') and repetition of what the interlocutor just said (other-repetition). Feedback words, in turn fall in two sub-types (see Allwood 1988): primary and secondary feedback words. Primary feedback words are words which are used to express feedback, i.e. the basic communicative functions (i-iv) above, but which cannot be used as predicates, attributes or adverbs. Examples of primary feedback words in Swedish are *m* ('m'), *ja* ('yes'), *nä* ('no'), *jo* (contrastive 'yes') and *okej* ('okay'). Secondary feedback words are words which in addition to function as feedback can be used as predicates, attributes or adverbs (and are tentatively more commonly used with such functions). Examples of secondary feedback words in Swedish are adjectives and adverbs such as *precis* ('precisely'), *bra* ('good') and *exact* ('exactly'). These types of vocal-verbal feedback can be used alone or in combination, forming units, which in turn can have different positions in an utterance: (i) single position, i.e. constitute the entire utterance, (ii) initial position, (iii) medial position, and (iv) final position. Of these, single and initial positions are characteristic and the most common for feedback.

Examples of gestural feedback are head nods, head shakes, smiles, raised or frowning eyebrows and shoulder shrugs. Using vocal-verbal and gestural feedback in combination results in multimodal feedback, e.g. a feedback word *ja* ('yes') in combination with a nod.

The following research questions are addressed: What are the timing relations between head movements and words in multimodal feedback contributions, i.e. do head movements start before, at the same time or after words, or vice versa? What prosodic features (F0-curve, duration, pitch) of vocal-verbal feedback are found when produced with versus without accompanying head movements, and more specifically, in relation to different kinds of head movements, i.e. in terms of their initial direction (e.g. up, down), repetition (repeated, single) and duration? Which feedback words co-occur with which prosodic patterns and with which types of head movements?

2 Method

Data for analysis consist of six video and audio recordings of dyads in (spontaneous and natural) first acquaintance conversations. Audio data was recorded with individual microphones for each speaker to facilitate acoustic/prosodic analysis. Video data was recorded using a three camera set up, with one camera taking in the whole scene, and two cameras focusing on the head and torso of each speaker respectively. The subjects had never met prior to the recording session, and were instructed to get to know each other during approximately 8 minutes. All subjects except one were university students. Of the six recordings, four are male-male and two are female-female conversation. Two of the speakers take part in two conversations each (with different partners), so the empirical material comprises 10 different speakers (six males and four females), in total.

Head movements have been coded by manual frame-by-frame analysis of the video recordings, identifying the following features: (i) type of head movement: head nod (vertical movement of the head, where the chin's distance to the torso varies as the head goes up and down), head shake (horizontal movements of the head, turning the head from side to side), head tilt (vertical and horizontal movements, tilting the head from side to side) or other; (ii) initial direction (in the case of head nods): up or down; (iii) repeated or single movement; and (iv) start time, end time and duration. Each frame of the recording is 40 milliseconds (ms), hence measures of time for head nods are measured with a level of detail of 40 ms.

Vocal-verbal feedback were analyzed using Praat and Audacity, identifying starting point, duration, general shape of the F0-curve and mean frequency of F0. Measurements of the average pitch of the highest and lowest 30 ms portions of the F0 curve were also taken for every utterance.

All analyzable cases of vocal-verbal feedback were categorized as one of eight different F0-curve types. The types were: complex, complex-down, complex-up, down, down-up, flat, up, and up-down. Up and down are to be interpreted as rising and falling pitch respectively. The categorization of vocal-verbal feedback was straightforward for most cases, as the shapes of the F0-curves clearly fell into one category or another. A statistical relation was used to decide if a curve was flat: any curve where the difference between the highest and the lowest 30 ms portions was less than 5% was deemed to be flat,

since such a small difference in pitch would not be audibly noticeable. The complex-up and complex-down categories were used for curves that had a general rising or falling shape, but with some irregularities of one kind or other. The complex category was used for curves that were judged not to fit into any of the other categories (22 out of 618 analyzed instances).

Certain values where derived from these measurements. An average pitch value of the F0 was calculated for every speaker (*Speaker Pitch*), based on the average of all of that speaker's vocal-verbal feedback. Subsequently, the average pitch of every vocal-verbal feedback unit was compared to the Speaker Pitch to describe its relative pitch (*Frequency Deviation*). For vocal-verbal feedback with a non-flat F0-curve, the difference between the highest and the lowest average was calculated as a percentage value (*Frequency Difference*).

108 out of 703 cases of vocal-verbal feedback were not analyzable in all prosodic dimensions, and 23 of these were not analyzable for any prosodic qualities at all. The most common reason for a unit not being analyzable is that sound from the other speaker is bleeding in to the signal, thus masking it. Instances of unanalyzed or partly analyzed vocal-verbal feedback were still used in cases where the affected prosodic dimensions were not of interest.

All pitch data should be fairly accurate within ±1 Hz. Duration data should be accurate within ±10 ms. The margin of error for comparative timing data is about ±40 ms or ±1 frame. Because the audio was recorded separately from the video, to ensure good audio quality, the two data streams had to be synchronized post recording. As the video is the master time track, the accuracy of timing relations is only as good as the time resolution of the video.

3 Results

3.1 General observations

Since the feedback system involves both vocal-verbal and gestural means, as well as different possible combinations of them, a variety of feedback types are possible. Based on the type of vocal-verbal component of the feedback contribution, if any, and the type of head movement, if any, the feedback types in Table 1 have been identified.

Head nods are by far the most common head movement used for feedback in the analyzed material, where 534 feedback head nods have been identified, while only 20 instances of other head movements with feedback function (e.g. shakes and tilts). Feedback head nods are more often co-produced with words (393 instances), than without words (141 instances): 74% vs. 26%. Inversely, vocal-verbal feedback is also more likely to be produced with feedback nods (393 instances, versus 290 instances produced without), but the difference is not as large: 58% vs. 42% (59% vs. 41% when including other head movements).

Vocal-verbal feedback (FB)	**Head movement**			**Tot.**
	Nod	Other	None	
Single primary FB word	297	15	227	539
Series of primary FB words	43	3	27	73
Combo of primary FB word(s) & secondary FB words (adverbs)	23	2	21	46
Combo of primary FB word(s) & other-repetition	4	0	9	13
FB clause	8	0	1	9
Other vocal-verbal (without FB word)	9	0	0	9
Single secondary FB word (adverb, adjective)	4	0	1	5
Primary FB word and OCM word	1	0	4	5
Other-repetition	2	0	0	2
Combo of secondary FB word (adverb) & other-repetition	1	0	0	1
Combo of primary FB word, secondary FB word (adverb) & other-repetition.	1	0	0	1
No vocal-verbal feedback (silence)	141	0	-	141
Total	534	20	290	844

Table 1. Number of instances of combinations of different kinds of vocal-verbal feedback and feedback head movements.

The two most common vocal-verbal forms of feedback are single primary feedback words, e.g. *m* ('m'), *ja* ('yes'), *nä* ('no'), *jo* (contrastive 'yes') and *okej* ('okay'), and primary feedback words used in series, e.g. *ja okej* ('yeah okay') and *ja ja* ('yeah yeah') (self-repetition). Furthermore primary feedback words are found in combination with secondary feedback words (e.g. adverbs), other-repetition and words for own communication management (OCM), e.g. *eh ja* ('um yeah'). Feedback that has a vocal-verbal component but lack a primary feedback

word altogether is uncommon. All these vocal-verbal types of feedback have initial positions of utterances, i.e. are followed by some non-feedback part, or they constitute the entire utterance themselves.

In addition to the contributions presented in Table 1, there are four cases of contributions in the recorded conversations which consist of a nod but where the vocal-verbal component is impossible to hear. None of these cases are considered for analysis below.

Due the meager data on other feedback head movement than head nods (only 20 instances in total), results on duration and timing focus only on nods (sect. 3.2 and 3.3). Also the comparison between head movements and prosodic features of speech will mainly, but not exclusively, focus on nods (sect. 3.5).

3.2 Timing of nods and words

In multimodal feedback contributions including nod and words (n=393), the nod can start before, after or at the same time as the word starts. That nod and word(s) start and/or end at exactly the same time is very unlikely, even though there are indeed two instances where this has been observed. This is of course subject to the level of detail of measurement, which in this study comes down to the marginal of a frame (40 ms). Consequently, in almost all cases the nod starts before the word(s) or the word starts before the nod, where the former being slightly more common than the latter (195 vs. 150 instances). However, since it is quite common that the nod and the word(s) start within a 120 ms time span,[1] i.e. almost at the same time, the following relations can be differentiated:

a) Nod starts more than 120 ms before word(s) (115 instances; 29% of the cases)
b) Nod and word(s) start within 120 ms span (146 instances; 37% of the cases)
c) Nod starts more than 120 ms after word(s) (instances 86; 22% of the cases)
d) In 46 cases the timing relation is unknown due to lack of reliable measurements (12% of all instances)

A majority of type c are produced with a gap (53 instances), i.e. the word both start and end before the nod starts. Less common, there are also 23 instances of gaps in the case of type a. This raises questions about the multimodality of such cases, and it is here argued that multimodality is a question of perceiver interpretation; that two communicative behaviors in different modalities belong together as a multimodal unit cannot be reduced to a simple question of co-occurrence in time.

3.3 Duration of nods

Feedback nods are longer when co-produced with words (multimodal), than when produced without words, see Table 2.

Nod type	With Words (MM)			Without words			MM nods are:
	M	n	St.d	M	n	St.d	
Repeated down nod	1201	153	724	940	84	469	28% longer
Single down nod	330	33	105	273	6	96	21% longer
Repeated up nod	1229	108	790	1028	40	420	20% longer
Single up nod	511	99	290	422	11	196	21% longer
Total	961	393	723	896	141	472	

Table 2. Mean duration in milliseconds (ms) and standard deviation of feedback nods in relation to co-production with words (multimodal, MM), or not.

Table 2 shows that for all nod types the nods which are produced with words are 20-28% longer in mean duration than those produced without words. As we shall see below, words do not have longer duration when they are accompanied by head movements, in general.

3.4 Head movement types and feedback words

A majority of the contributions under consideration here contain one or several primary feedback words, in different ways, e.g. as a single constituent of the vocal-verbal feedback, in series or together with secondary feedback words or other-repetition. (Exceptions are, for instance, contributions that as a vocal-verbal feedback part only consist of secondary feedback word or other-repetition, see Table 1).[2] Primary feedback words differ both in the extent that they do co-

[1] The 120 ms (three video frames) time span is chosen because it is larger than the error margin of the synchronization of video and audio, while still being an almost unnoticeable delay for a human observer.

[2] There are 684 contributions in the empirical material which contain at least one primary feedback word.

occur with head movements, and the types of head movements (and nod types) they do co-occur with.

First, considering the five most common feedback words, which all are primary, the prevalence of accompanying nods differ. The five most common feedback words in the material are: *ja* ('yes/yeah') (382 instances), *m* ('m') (148 instances), *okej* ('okay') (78 instances), *nä* ('no') (55 instances), and *jaha* ('yes, I see') (23 instances). The feedback word *ja* ('yes/yeah') is equally common together with head movement as without any. The word *nä* ('no') is slightly more common without head movements than with. The words *m* ('m'), *okej* ('okay') and *jaha* ('yes, I see') are more common together with head movement than without. See Table 3.

FB words	n	Without head movement	With head movement
ja	382	50%	50%
m	148	23%	77%
okej	78	24%	76%
nä	55	55%	45%
jaha	23	35%	65%

Table 3. The extent that the five most common feedback (FB) words *ja* ('yes/yeah'), *m* ('m'), *okej* ('okay'), *nä* ('no') and *jaha* ('yes, I see') are multimodal with respect to head movements.

Second, looking closer at the types of head movements that accompany these five words, further differences emerge, see Table 4.

FB words	Rep. down nod	Sing. down nod	Rep. up nod	Sing. up nod	Other head movem.
ja (n=191)	31%	10%	27%	29%	2%
m (n=114)	56%	10%	29%	4%	2%
okej (n=59)	15%	2%	34%	46%	3%
nä (n=25)	24%	0%	8%	40%	28%
jaha (n=15)	0%	0%	13%	80%	7%

Table 4. The relation between the five most common feedback (FB) words *ja* ('yes/yeah'), *m* ('m'), *okej* ('okay'), *nä* ('no') and *jaha* ('yes, I see') and different kinds of head movements in multimodal feedback.

The word *m* ('m') is strongly associated with repeated nods. For *m* ('m') co-production with repeated down nods and repeated up nods constitute 85% of its uses in contributions with head movement. Of the five words, *m* ('m') is the word which is strongest associated with repeated down nods. Both *okej* ('okay') and *jaha* ('yes, I see') are strongly associated with (single and repeated) nods which have an initial upward direction: 80% of *okej* ('okay') and 93% of *jaha* ('yes, I see'), which are produced with head movements, are produced with single or repeated up nods. Of the five words, *jaha* ('yes, I see') is the word which is strongest associated with single up nods (80%). The word *nä* ('no') is the only of the five words which is common together with other head movements than nods. The most common kind of head movement in question here is the head shake. It should however be noted that *nä* ('no') is more common with nods than with shakes. This results is to be interpreted in relation to the affirmative use of *nä* ('no') in response to utterances which contain negation (see e.g. Allwood et al 1992). When the word *ja* ('yes/yeah') is co-produced with head movements, it is overwhelmingly used with head nods, but lacks any strong association with a particular type of nod.

3.5 Head movements and prosody

This section discusses the prosodic features of word duration and pitch in relation to head movements in multimodal feedback. Above, feedback nods were found to have longer duration when accompanied by words, see section 3.3. Turning to the duration of vocal-verbal feedback, the trend that "multimodal is longer" does not seems to apply; cf. Allwood and Cerrato (2003) who found that feedback words were 20-40% longer when produced with head movements. Table 5 shows the mean duration of the five most common feedback words, in cases where these feedback words alone constitute the vocal-verbal feedback of a contribution, including all cases when this feedback is only a part of a contribution as well as constituting the whole contribution (see "Single primary FB word" of Table 1).

In cases of *m* ('m') and *okej* ('okay') as single feedback words, the difference in mean duration of them being co-produced with nod or not is minimal (only 1% difference in the case of *m* and 3% difference in the case of *okej*). The word *nä* ('no') as a single feedback word is slightly longer in duration when produced with head movement, than when it is produced without (9% longer), while *ja* ('yes') is slightly longer in duration when produced without head movement than with (15% longer). Of these single feedback words, only *jaha* ('yes, I see') is considerably longer when produced with head movement, than

without head movement (56% longer), but note that the instances of *jaha* ('yes, I see') are quite few.

Single FB word	n	With head movement			Without head movement		
		M	n	St.d	M	n	St.d
ja	273	217	132	75	250	141	137
m	120	225	93	68	222	27	57
okej	47	305	40	145	313	7	114
nä	32	255	13	121	233	19	94
jaha	11	397	7	122	255	4	55

Table 5. Mean duration of the five most common feedback (FB) words as the only vocal-verbal feedback part of a contribution, in relation to co-production with head movement, or not.

This suggests that feedback words not are longer when they are accompanied by head movements, than when they are not, at least not when considering head movements in general. However, when turning to more specific head movements, namely different nod types, a slightly different pattern emerge. Diagram 1 shows the mean duration of the five most common feedback words, as single feedback words, in relation to their accompaniment of single up nods, repeated up nods, repeated down nods, single down nods, and no nod at all.

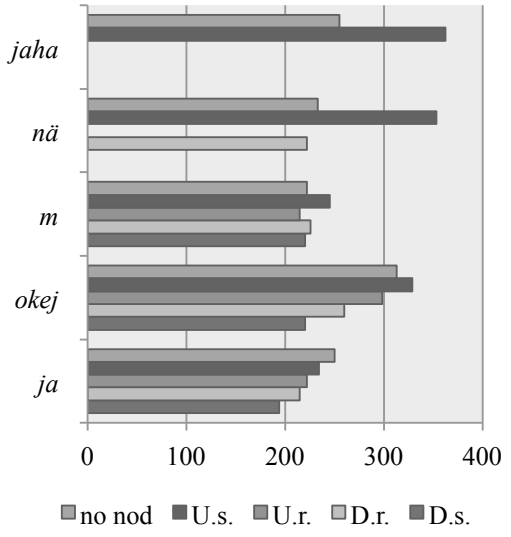

Diagram 1. Mean duration (ms) of the five most common single feedback words when co-produced with down-single (D.s.), down-repeated (D.r.), up-repeated (U.r.), up-single (U.s.), and no nod.

For all the words in Diagram 1, except for *ja* ('yes'), all have the longest mean duration when produced with single up nods. (It should be noted that *m* ('m') is uncommon with single up nods, see Table 4, so the mean duration of *m* ('m') with single up nod is based on only two instances.) So even though there is no consistent finding that words are longer when they are produced with head movements, than when they are not, it seems to be the case that feedback words are typically longer when they are produced with single up nods, than when they are not. That *jaha* ('yes, I see') have longer duration when produced with head movement (in general), see Table 5, should be understood in relation to its high co-occurrence with single up nods, see Table 4, and the observation that feedback words tend to be longer when produced with single up nods. Diagram 1 also shows a contrast between words accompanied by up-single nods and down single nods. All multimodal words are longest with up-single nods, while at least in the case of *ja* ('yes') and *okej* ('okay'), the shortest words are produced with down single nods. The duration of the word seems to vary systematically depending on what head movement accompanies the word. Another prosodic feature that also shows evidence of such systematic variation is the feature of pitch, which will be discussed below.

F0 contour	n	With head movement		Without head movement	
		n	%	n	%
Flat	183	116	63	67	37
Down	151	105	70	46	30
Up	123	51	41	72	59
Up-down	56	24	43	32	57
Complex-down	31	21	68	10	32
Down-up	26	14	54	12	46
Complex	32	21	66	11	34
Complex-up	16	13	81	3	19
Measuring error	55	24	44	31	56
Total	673	389		284	

Table 6. F0 contours of vocal-verbal feedback in relation to the accompaniment of head movement or not.

The most common F0 contours of the material are flat, down/falling and up/rising. These three types differ in their distribution with and without the accompaniment of head movement, see Table 6. Vocal-verbal feedback which have a flat or falling F0 contour are more common together

with head movement, than without, while vocal-verbal feedback having a rising F0 contour are more common without the accompaniment of head movement.

Looking at associations of different kinds of head movements and different prosodic features, we find a number of differences.

There seems to be a general trend that single upwards nods tend to co-occur with more stressed vocal-verbal feedback. Also repeated downward nods often co-occur with less stressed vocal-verbal feedback. This is shown in several prosodic dimensions.

Nod type	Freq. diff.	Freq. diff. ≥ 20%	Freq. dev.
Repeated down-nod	15%	21%	-3%
Single down nod	16%	37%	-6%
Repeated up nod	17%	28%	-2%
Single up nod	26%	54%	3%
None	18%	31%	12%

Table 7. Mean pitch measures of vocal-verbal feedback in relation to nod types. Freq. diff. ≥ 20% are the percentage of instances that have a frequency difference larger or equal to 20%.

Nod type	Mean duration of words (ms)
Repeated down-nod	223
Single down- nod	207
Repeated up-nod	239
Single up-nod	289
None	252

Table 8. Mean duration of single feedback words with different nods types (in milliseconds)

Nod type	Low saliency F0		High saliency F0	
	n	%	n	%
Repeated down-nod	104	76%	33	24%
Single down-nod	21	69%	12	31%
Repeated up-nod	67	69%	30	31%
Single up-nod	42	49%	44	51%
None	123	48%	131	52%

Table 9. Saliency of F0 contours of vocal-verbal feedback in relation to nod types. Low saliency F0 = flat, down, and complex-down. High saliency F0 = up, up-down, down-up, complex-up, and complex.

On average, vocal-verbal feedback co-occurring with single up nods have more prominent pitch features, such as more pitch variation (Frequency difference), in general higher pitch compared to mean speaker pitch (Frequency deviation). These single feedback words also have longer duration on average (Table 8), as well as a tendency to have more salient F0-curve characteristics (i.e. rising pitch at some point) (Table 9). As increased and/or rising pitch and increased duration are all considered to be typical prosodic features of stress, this suggests that the single up nod type is more likely to be co-produced with stressed vocal-verbal feedback than the other nod types are.

4 Summary and discussion

The results of this study are summarized below:

- Head nods are by far the most common type of head movement used for feedback (in Swedish first acquaintance conversations).
- In communicative feedback, words and nods are more frequently used in combination (multimodal), than on their own.
- Multimodal nods more frequently start before or at the same time as words, rather than words starting before nods.
- Nods have longer duration when produced with words (multimodal) than without, but words, however, are not typically longer when multimodal.
- Vocal-verbal feedback having a flat or falling F0 contour are more common together with nod, than without, while vocal-verbal feedback having a rising F0 contour are more common without the accompaniment of nod.
- Certain feedback words (*m* ('m') and *okej* ('okay')) are more often produced with nods, than others (*ja* ('yes') and *nä* ('no')).
- Furthermore, certain feedback words are typically associated with certain nod types, most prominently *okej* ('okay') and *jaha* ('yes, I see') with up nods, and *m* ('m') with repeated (down) nods.
- Single up-nods tend to occur with vocal-verbal feedback that have more salient prosodic features, while repeated down nods tend to occur with vocal-verbal feedback that have less salient prosodic features.

These results do to some extent differ from previous findings. First, Allwood and Cerrato (2003) found that feedback words were 20-40%

longer when produced with head movements, than without. This pattern was not confirmed for our data (see Diagram 1 and Table 8). (Note that nods are longer when they are multimodal, than when they are not.) Also, to be predicted from previous research on "visual prosody" is that emphasis in speech is likely to be produced with head movements (Graf et al, 2002; Swertz and Krahmer, 2010). Again, we do not find any unequivocal evidence for this pattern here. For example, feedback words do not typically have longer duration (Table 8), nor are they more salient (Table 9) when produced with nods, *per se*, than when they are produced without. Feedback words do have salient prosodic features and longer duration with some nod types, i.e. single up nods and to some extent repeated up nods, but not with other nod types, i.e. down nods. We therefore suggest that, how word and head movement are co-produced in multimodal feedback seems to be dependent on the *type* of word and the *type* of nod forming the multimodal contribution, rather than their co-production, *per se*.

Acknowledgments

This work is funded by the Swedish Research Council (VR) and the Nordic council (NOS-HS NORDCORP). We wish to thank Jens Allwood, Karl Johan Sandberg and Axel Olsson, as well as the anonymous reviewer for constructive criticism and suggestions.

References

Allwood, J. (1988) Om det svenska systemet för språklig återkoppling. In: P Linell, V. Adelswärd & P. A. Pettersson (ed.) *Svenskans beskrivning* 16, vol. 1. Linköping: Tema kommunikation, Linköpings universitet.

Allwood, J. and 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, pp. 7-22.

Allwood, J., Nivre, J. & Ahlsén, E. (1992) On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9(1): 1-26.

Allwood, J., Kopp, S., Grammer, K., Ahlsén, E., Oberzaucher, E. & Koppensteiner, M. (2007) The analysis of embodied communicative feedback in multimodal corpora: a prerequisite for behavior simulation. *Language Resources and Evaluation*, 41(3-4): 255-272.

Bull, P. and Connelly, G. (1985) Body movement and emphasis in speech. *Journal of Nonverbal Behavior*, 9(3): 169-187.

Cavé, C., Guaïtella, I., Bertrand, R., Santi, S. Hralay, F. & Espesser, R. (1996) About the relationship between eyeborw movements and F0 variations. In: H. T. Bunnell & W. Idsardi *Procedings of ICSLP*, Philadelphia, PA, USA, pp. 2175-2178.

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*.

Granström, B. and House, D. (2005) Audiovisual representation of prosody in expressive speech communication. *Speech Communication*, 46: 473-484.

Kendon, A. (2004). *Gesture: Visible Action as Utterance.* Cambridge: Cambridge University Press.

McNeill, D. (1985). So you think gesture are nonverbal? *Psychological Review, 92*, 350-371.

Morency, L.-P., Sidner, C., Lee, C., and Darrell, T. (2007) Head gestures for perceptual interfaces: The role of context in improving recognition. *Artificial Intelligence*, 171(8-9):568-585.

Moubayed, S. A., Beskow, J. and Granström, B. (2010) Auditory visual prominence: from intelligibility to behaviour. *Journal on Multimodal User Interfaces*, 3(4): 299-311

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T. and Vatikiotis-Bateson, E. (2004) Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science*, 15(2): 133-137.

Scarborough, R., Keating, P., Mattys, S. L., taehong, C. and Alwan, A. (2009) Optical phonetics and visual perception of lexical and phrasal stress in English. *Language and Speech*, 51(2-3): 135-175.

Krahmer, E. and Swertz, M. (2009) Audiovisual prosody – introduction to the special issue. *Language and Speech*, 52(2-3): 129-133.

Swerts, M. and Krahmer, E. (2010) Visual prosody and newsreaders: Effects of information structure, emotional content and intended audience on facial expressions. *Journal of Phonetics*, 38: 197-206.

Feedback and gestural behaviour in a conversational corpus of Danish

Patrizia Paggio
University of Copenhagen
Centre for Language Technology
paggio@hum.ku.dk

Costanza Navarretta
University of Copenhagen
Centre for Language Technology
costanza@hum.ku.dk

Abstract

This paper deals with the way in which feedback is expressed through speech and gestures in the Danish NOMCO corpus of dyadic first encounters. The annotation includes the speech transcription as well as attributes concerning shape and conversational function of head movements and facial expressions. Our analysis of the data shows that all communication modalities, i.e. head, face and eyebrows, contribute to the expressions of feedback, with repeated nods and smiles as the most frequent feedback gesture types. In general, the use of nods as feedback gestures in our data is comparable to what earlier studies have found for other languages, but feedback is also often expressed by other head movements and smiles.

1 Introduction

Head movements and facial expressions play an important function in face-to-face interaction. In particular, many authors have observed that head nods are an important means of expressing what we here call feedback, i.e. unobtrusive behaviour that has the purpose of either giving or eliciting signs of *contact*, *perception*, *understanding* and *agreement* or *disagreement* (Allwood et al., 1992).

Dittmann and Llewellyn (1968), for instance, focus on nodding by listeners, and find that nods occur together with brief feedback responses more often than predicted by chance. Yngve (1970) and Duncan (1972) consider head nods as examples of backchannels, i.e. feedback signals given by the listener without trying to take the floor. Hadar et al. (1985) monitor head movements in five subjects during conversation, and find that agreeing is one of the functions head movements are associated with (the others are wanting to take the turn and aligning with the interlocutor's stressed syllables and pauses). Maynard (1987) studies head nods in dialogues between Japanese speakers. The most frequent function is found to be feedback by listeners, but speakers also nod a lot in different contexts. An interesting observation in this study relates to the culture-specificity of gesturing: the Japanese nod with an average frequency of 5.57 seconds (in other words, one nod for every 5.57 seconds), while Americans do so with an average of only 22.5 seconds. McClave (2000), in a qualitative study of head movements in dialogues between two pairs of American speakers, observes that head movements occur together with a whole array of functions and senses, one of which is linked to what she calls backchanneling requests: the speaker nods to ask the listener for feedback, and the listener in turn nods.

Head movements have also been studied in relation to Scandinavian languages, of which Danish, which is targeted in this paper, is an example. It has been observed that 70% of all head movements in a subset of the Swedish GSLC corpus (Nivre et al., 1998) are related to feedback, and that most of these are nods and up-nods (Cerrato, 2007).

While there is a whole body of research on facial expressions as vehicles of emotional response (Hager and Ekman, 1983; Busso and Narayanan, 2007), less attention has been given to the role played by facial expressions with respect to conversational feedback. Smiles and laughter as signals of feedback are studied for instance by Allwood and Lu in this volume and Lu et al. (Under publication), who find that in first encounter situations, both Chinese and Swedish speakers use smiles and chuckles to give feedback.

In previous work (Jokinen et al., 2008), we studied facial expressions and head movements in Danish and Estonian dialogues, and noticed significant interdependences between non-verbal expressions and communicative functions. Nods

often indicate feedback, while head movements sideways or up-down together with gaze are related to turn-taking. In Paggio and Navarretta (2010) and Navarretta and Paggio (2010) we looked at the relation between head movements and facial expressions on the one hand, and the dialogue act functions of linguistic feedback expressions on the other and showed that head gestures, where they occur, contribute to the semantic interpretation of feedback expressions in a significant way.

Here we present empirical evidence from a multimodal corpus of Danish first encounters of how head movements and facial expressions are used in conversational Danish as signals of feedback giving and eliciting. We start by explaining how the corpus was collected in Section 1. We then describe the annotation categories and procedure used in Section 2. In Section 3 we provide quantitative measures of the annotated data. In Section 4 we briefly discuss how the corpus can be used in machine learning studies of multimodal behaviour and conclude.

2 Corpus collection

The Danish NOMCO corpus is one of a number of multimodal corpora in Swedish, Danish, Finnish and Estonian that have been collected and annotated within the Nordic NOMCO project (Paggio et al., 2010). The aim of the project is to provide comparative annotated multimodal data in the Nordic languages and, based on these data, to investigate how speech and gestures together are used to express feedback, turn taking, sequencing and information structure.

The Danish first encounter corpus consists of 12 dyadic interactions of a duration of approximately 5 minutes each, in which subjects who have not met before try to get to know each other. The participants were six males and six females, all native speakers of Danish aged between 21 and 36, either university students or people with a university education. They did not know each other beforehand, and were not acquainted with the purpose of the recordings. The videos were recorded in the TV studio of the Faculty of Humanities at the University of Copenhagen. The subjects are standing in from of each other and are recorded by three different cameras. The speech is recorded through microphones attached to the ceiling. For each dialogue, two versions were produced, one showing a long shot of the two participants facing each other, the other combining two mid shots taken from different angles into a split video. The two views are shown in Figure 1.

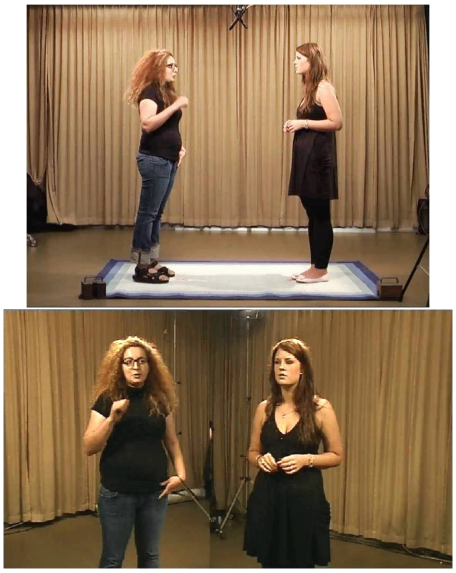

Figure 1: Recordings from the Danish NOMCO dialogues: total and split views

A questionnaire was given to the participants to collect information on how they experienced the conversations. They were asked to rate their experience along a number of parameters concerning their emotional state and the interaction itself. The results indicate that the subjects were not too affected by the artificial setting even though they were aware of it. In particular, since the scores for perturbedness, tenseness and awkwardness were all below average, we consider the corpus a relatively valid exemplification of natural interaction. For a more detailed analysis of the questionnaire results, see Paggio and Diderichsen (2010).

3 Annotation categories and procedure

3.1 Orthographic transcription

The first step in the annotation process was to produce an orthographic transcription of the audio signal. This was done using Praat (Boersma and Weenink, 2009). The transcription includes word boundaries as well as word stress, indicated by a "," before the stressed vowel. Pauses are represented by a "+", and filled pauses glossed with English words, e.g. *laugh, breath* or expressions such as *øh*. The Praat transcriptions were then imported into the ANVIL tool (Kipp, 2004), which was

used for gesture annotation. In ANVIL, sentence boundaries, in the front of an attribute *boundary true*, were added to the transcription based on the occurrence of pauses as well as on syntactic criteria. Furthermore, topic and focus were identified in each sentence, and the attributes *topic true* and *focus true* were added to the corresponding words according to the methodology described in (Paggio, 2006a; Paggio, 2006b). In short, *topic* indicates the presupposed entity about which the sentence predicates something new, while *focus* indicates non-presupposed information.

Word token	topic	focus	boundary
+	false	false	*true*
jeg	*true*	false	false
hedder	false	*true*	false
H,anne	false	*true*	false
+	false	false	*true*

Table 1: Topic and focus annotation example

In Table 1 we show in table format the assignment of topic, focus and clause boundary attributes to the utterance *jeg hedder Hanne* (lit: I call Hanne, or "My name is Hanne") from one of the NOMCO dialogues. Boundaries are placed together with the pauses that precede and follow the sentence, *jeg* (I) refers to topic, i.e. the entity about which the sentence predicates something new, whilst *hedder Hanne* (lit: call Hanne), which contains the only stressed word, is the focus, i.e. the new information.

3.2 Gesture annotation scheme

Gestures in the NOMCO data are annotated with a subset of the attributes defined in the MUMIN annotation scheme (Allwood et al., 2007). The MUMIN scheme is a general framework for the study of gestures in interpersonal communication that has been applied to multimodal data in several languages within the context of the Nordic MUMIN network (www.cst.dk/mumin). It concerns facial expressions, head movements, hand gestures and body posture, and it provides attributes for shape as well as function.

The attributes for the annotation of gesture shape used in this study are shown in Table 2. The granularity of the annotation categories is deliberately coarse in that we only want to be able to distinguish different communicative functions rather than provide precise morphological descriptions.

The functional annotation features in MUMIN concern feedback, turn management and sequencing. In this study, however, only feedback attributes will be considered. They are shown in Table 3.

Modality	Attribute	Value
Head	HeadMovement	Nod, Jerk, HeadForward, HeadBackward, Tilt, SideTurn, Shake Waggle, HeadOther
	HeadRepetition	Single, Repeated
Face	GeneralFace	Smile, Laugh, Scowl, FaceOther
	Eyebrows	Frown, Raise, BrowsOther

Table 2: Shape Annotation Features for Head and Face

Attribute	Value
Basic	ContactPerceptionUnderstanding (CPU), BasicOther
Direction	FbGive, FbElicit, FbGiveElicit, FbUnderspecified
Agreement	Agree, NonAgree

Table 3: Functional annotation of feedback gestures

The *Basic* attribute has two possible values: *ContactPerceptionUnderstanding (CPU)* indicates that participants are willing and capable of interacting, perceiving and understanding what is being communicated (Allwood et al., 1992); *BasicOther* is used if one of the above dimensions, e.g. understanding, appears to be lacking (this does not occur in the current corpus, thus only CPU is used) If *Basic* is coded, a value for the *Direction* attribute has to be chosen, too. We distinguish between i. *FeedbackGive*, where the listener gives feedback (often called backchannelling), ii. *FeedackElicit*, where the speaker appears to be eliciting feedback from the listener, iii. a combination of both values, and iv. an underspecified value. Finally, a feedback gesture may express agreement or disagreement towards a statement, for which the scheme foresees the two values *Agree* and *NonAgree*.

In addition to the shape and function attributes, for each gesture a relation with the corresponding speech expression, if one such exists, is also annotated by means of a link. The link can point to a speech segment uttered by the person producing the gesture (by means of the attribute *MMRelationSelf*), or to a speech segment in the interlocutor's vocal stream (by means of the attribute *MMRelationOther*).

3.3 Gesture annotation procedure

Three annotators, all of them students of linguistics, created the annotation. To ensure reliability, they received an initial training where they all worked together coding the same video. Then a second video was coded by each of them separately. The results were discussed and corrected, and a set of written guidelines were developed based on these discussions. In this preliminary exercise, the *Cohen's kappa* (Cohen, 1960) figures obtained were on average for the three pairs of coders in the range 0.5-0.6 for face attributes and 0.6-0.8 for head movements. Considering the fact that the agreement measure calculated in ANVIL reflects agreement of segmentation as well as labelling, these figures are quite satisfactory.

Each of the remaining videos was subsequently annotated by one of the coders and corrected by the other. Disagreements were again discussed and evened out. If the two coders still could not agree, a third annotator made the final decision. Throughout this process, the guidelines were continually improved with examples and explanations. After having annotated five videos following this procedure, we repeated the inter-coder agreement exercise between the two annotators who had shown most disagreement the first time, and noted an improvement of about 10% for both face and head gestures.

To annotate facial expressions and head movements according to this procedure takes on average 2 hours per minute per speaker including discussions and subsequent corrections.

4 Data analysis

So far, nine of the twelve videos have been annotated and analysed. The total duration of this annotated material is 3027 seconds, in other words 50 minutes and 45 seconds. The length of the individual annotated clips varies from about 140 seconds to about 360. The total number of word tokens (including filled pauses) is 10800. The total number of gestures identified is 3391.

4.1 Gesture frequency

Table 4 shows how gestures are distributed according to the three major shape attributes. Note that the *Eyebrows* gestures listed here are those occurring without a concomitant general facial expression like *Smile* or *Laughter*. Head movements are also coded with a value for repetition. The distribution is 1714 *Single* movements and 579 *Repeated* ones. Head movements constitute the majority of the gestures, and most of them are single movements.

FaceGeneral		**Eyebrows**		**HeadMovement**	
Smile	499	Raise	263	Nod	520
Laughter	198	Frown	85	Tilt	388
FOther	45	BOther	3	SideTurn	328
Scowl	5			HForward	264
				Shake	257
				HBackward	200
				HOther	148
				Jerk	122
				Waggle	66
Face total	747	Brows total	351	Head total	2293

Table 4: Gesture types in the Danish NOMCO corpus

Type	**No**	**sec/g**	**g/w**	**g/sec**
All gestures	3389	0.89	0.31	1.12
Head	2291	1.32	0.21	0.76
Nods	520	5.82	0.05	0.17
Face	747	4.05	0.07	0.25
Eyebrows	351	8.62	0.03	0.12

Table 5: Gesture type frequency

In Table 5 we show the frequency counts for some of the most frequent gesture types. The second column shows the raw counts, the third one the proportion of seconds per gesture, the fourth one the proportion of gestures per word, and the last one the proportion of gestures per second. The proportion of seconds per gesture allows us to compare with the findings in the already mentioned study by Maynard, where it is claimed that Japanese speakers make a nod every 5.5 seconds. The figure for Danes is one nod every 5.6 seconds, which is very similar. This seems to show that Danes and Japanese behave similarly as far as nodding is concerned - at least in the sense that they nod with similar frequencies. However, the subjects in Maynard's study already knew each other, so the datasets are not directly comparable. Moreover, we have not looked at dimensions concerning the amplitude or velocity of the nods, where differences may indeed arise. A discussion of how differences in gestural behaviour can be couched in the perspective of cultural diversity can be found in Paggio and Navarretta (2011).

An interesting issue is how much individual difference can be observed in a corpus which is trying to model culture-specific behaviour in a certain communication situation, or activity. Table 6

Modality	Average No	SD
Face	61.00	26.80
Head	127.28	34.61

Table 6: Number of facial expressions and head movements: average and standard deviation

shows the average number of facial expressions (this time including eyebrows) and head movements together with standard deviation figures. The variation is especially large in the case of facial expressions, suggesting that one should be cautious in generalising from these data, and that more data should be added to the corpus to provide a more reliable basis for quantitative studies of facial expressions. A question that we will investigate further in these data is whether the deviation in gesture production is dependent on the amount of speech produced by the gesturer and by the interlocutor.

4.2 Gesture and feedback

Out of the total 3391 gestures identified in the corpus, 1594 (47%) have been annotated with the *Basic CPU* feedback feature. This means that on average, there is a feedback gesture either by the speaker or by the interlocutor every 0.3 seconds. Is this what one should expect? In order to answer the question, it may be useful to compare with other corpora in Danish or similar corpora in other languages.

Corpus	No g	g/w	FB g	No w	FB w
NOMCO	3391	0.3	47%	10,800	0.06%
DanPASS	264	0.05	21%	5,556	7.00%

Table 7: Feedback in the NOMCO and DanPASS corpora

We can start by looking at feedback in the DanPASS dialogues, which are part of a corpus of spoken Danish (Grønnum, 2006) in which two speakers have to solve a map-task. The subjects sit in separate studies without being able to see each other, and they talk through headsets. Given the very different settings as well as the different genres (map-oriented dialogue vs free conversation), we would expect more feedback words (*yes*, *no*, and similar) and less feedback gestures in DanPASS as opposed to NOMCO. We have used a small sub-set of this corpus (8 videos) for earlier studies, where head movements and facial expressions were annotated following the same methodology as in NOMCO. In Table 7 we show how this sub-corpus compares with the NOMCO data on a number of parameters. As expected, the number of gestures by word is in general much lower in DanPASS, and the proportion of gestures that are used for feedback is also lower. We have not conducted an analysis of the functions of the remaining gestures, we can only guess that they may have a turn taking or focusing function. Finally, the percentage of feedback words is as expected much higher in DanPASS compared to NOMCO. Participants in a task-oriented dialogue that cannot see each other need to check mutual understanding and grounding by using feedback words.

Gesture	No	%
Nod Repeated	250	0.16
Smile	248	0.16
Nod Single	134	0.08
Tilt	125	0.08
Raise	117	0.07
Shake	112	0.07
HeadBackward	110	0.07
HeadForward	99	0.06
Jerk	92	0.06
Laughter	91	0.06
SideTurn	84	0.05
Frown	40	0.03
HeadOther	40	0.03
FaceOther	32	0.01
Waggle	20	0.01
Total	1594	1

Table 8: Feedback distribution in the Danish NOMCO corpus

While it is easy to see that NOMCO is different from a map-task dialogue with respect to gestural behaviour in general, and to gestural feedback in particular, it is not so straightforward to compare it with similar corpora in different languages. The NOMCO project is working on a comparison between Danish, Swedish and Finnish data. Here, we will hold the Danish NOMCO data against earlier findings on the use of nods as feedback signals in Japanese and Swedish and Japanese.

Table 8 shows how feedback gestures in the Danish NOMCO corpus are distributed among different gesture types. Head movements are in general the preferred feedback modality. In fact, about 67% of the head movements (as opposed to 47% of all movements) is used to express feedback. This is similar to the results obtained by Cerrato (op.cit.) for Swedish. If we look at specific movement types, nods are by far the most common type. We have seen that nods occur roughly as often in our corpus as in the Japanese data studied by May-

nard (op.cit.), i.e. every 5-6 seconds. In the Danish data, in 54.61% of the cases, nods are used to express feedback. In the Japanese data, Maynard claims that nods are used as feedback signals in almost 50% of the cases (other functions mentioned in this study are turn shifts, emphasis and clause boundary marking). Thus, the Danish and Japanese data also seem similar on this dimension, although, as already pointed out, these comparisons should be taken with due caution because not all aspects are kept equal in the two corpora.

In general we can conclude that the use of head movements and facial expressions as feedback signals in the NOMCO corpus confirms earlier findings concerning the pervasiveness of the phenomenon as well as the frequent use of nods as feedback signals. However, our data also show that other head movements, such as tilts, shakes and head-backward movements, are often used to express feedback. Finally, to conclude this section, the data also allow us to see which of the feedback directions is the most frequent. In 77% of the cases feedback is given, in 20% it is elicited, and in 3% of the cases both directions seem present at the same time.

5 Conclusion

The analysis of feedback in a multimodally annotated Danish corpus of first encounters shows that both speech and gestures (in the present study head movements and facial expressions) are used, alone or in combination, to give and elicit feedback. The most frequently used feedback-related gestures in the data are head movements, especially repeated and single nods, confirming preceding studies of multimodal feedback. However in our corpus also other types of head movement and various facial expressions have been recognised to have a feedback-related function.

Comparing feedback expressions in this corpus and in a map-task corpus we found that feedback was expressed more frequently with gestures in the former, and verbally in the latter. These results are not surprising given the nature and the settings of the two corpora.

The analysis of the annotated data also indicates that there is a large individual variation in the frequency with which the interaction participants used gestures to express feedback. This is especially true for facial expressions. In future we will investigate the relation between individual frequency of speech and gesture production in the NOMCO data. Furthermore, future work still related to the study of feedback will also comprise the comparison of feedback expression in first encounters corpora in two other Scandinavian languages for which these corpora have been collected and annotated.

While the focus of this study has been on gestural feedback, the Danish NOMCO corpus of first encounters provides the means to investigate the interaction of speech and gestures with respect to a number of conversational functions, especially turn taking and information structure. The rich functional annotation of gestures will be analysed against the focus and topic tags but also in comparison with automatically extracted prosody features. Finally, we also plan to annotate hand gestures to provide a comprehensive analysis of the multimodal behaviour in the corpus.

Acknowledgements

We would like to acknowledge our partners in the NOMCO project Elisabeth Ahlsén, Jens Allwood and Kristiina Jokinen, as well as the annotators Sara Andersen, Josephine B. Arrild, Anette Studsgård and Bjørn Wesseltolvig. The NOMCO project, the full name of which is "Multimodal Corpus Analysis in the Nordic Countries", is funded by the NOS-HS NORDCORP programme, see http://sskkii.gu.se/nomco/. The work of the Danish group in the project is also funded by the Danish Research Council for the Humanities, see http://cst.dk/vkk/uk/.

References

Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9:1–26.

Paul Boersma and David Weenink. 2009. Praat: doing phonetics by computer (version 5.1.05) [computer program]. Retrieved May 1, 2009, from http://www.praat.org/.

Carlos Busso and Shrikanth S. Narayanan. 2007. Interrelation between speech and facial gestures in emotional utterances: A single subject study. In *IEEE Transactions on Audio, Speech and Language Processing*, volume 15, pages 2331–2347, July.

Loredana Cerrato. 2007. *Investigating Communicative Feedback Phenomena across Languages and Modalities*. Ph.D. thesis, Stockholm, KTH, Speech and Music Communication.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Allen Dittmann and Lynn Llewellyn. 1968. Relationship between vocalizations and head nods as listener responses. *Journal of Personality and Social Psychology*, 9.

Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23:283–292.

Nina Grønnum. 2006. DanPASS - a Danish phonetically annotated spontaneous speech corpus. In Nicoletta Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, and D. Tapias, editors, *Proceedings of the 5th LREC*, pages 1578–1583, Genoa, May.

Uri Hadar, T. J. Steiner, and F. Clifford Rose. 1985. Head movement during listening turns in conversation. *Journal of Nonverbal Behavior*, 9(4):214–228, December.

Joseph C. Hager and Paul Ekman. 1983. The Inner and Outer Meanings of Facial Expressions. In J. T. Cacioppo and R. E. Petty, editors, *Social Psychophysiology: A Sourcebook*, chapter 10. The Guilford Press, New York.

Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2008. Distinguishing the communicative functions of gestures. In *Proceedings of the 5th MLMI*, LNCS 5237, pages 38–49, Utrecht, September. Springer.

Michael Kipp. 2004. *Gesture Generation by Imitation – From Human Behavior to Computer Character Animation*. Boca Raton, Florida: Dissertation.com.

Jia Lu, Jens Allwood, and Elisabeth Ahlsén. Under publication. A study on cultural variations of smile based on empirical recordings of Chinese and Swedish first encounters. In Dirk Heylen, Michael Kipp, and Patrizia Paggio, editors, *Proceedings of the workshop on Multimodal Corpora at ICMI-MLMI 2011*, Alicante, Spain, Nov.

Senko Maynard. 1987. Interactional functions of a nonverbal sign: Head movement in Japanese dyadic casual conversation. *Journal of Pragmatics*, 11:589–606.

Evelyn McClave. 2000. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32:855–878.

Costanza Navarretta and Patrizia Paggio. 2010. Classification of feedback expressions in multimodal data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 318–324, Uppsala, Sweden, Juli 11-16.

Joakim Nivre, Jens Allwood, Jenny Holm, Dario Lopez-Kästen, Kristina Tullgren, Elisabeth Ahlsén, Leif Grönqvist, and Silvana Sofkova. 1998. Towards Multimodal Spoken Language Corpora: TransTool and SyncTool. In *Proceedings of the Workshop on Partially Automated Techniques for Transcribing Naturally Occurring Speech at COLING-ACL '98*, Montreal, Canada, August.

Patrizia Paggio and Philip Diderichsen. 2010. Information structure and communicative functions in spoken and multimodal data. In Peter Juel Henriksen, editor, *Linguistic Theory and Raw Sound*, volume 49 of *Copenhagen Studies in Language*, pages 149–168. Samfundslitteratur.

Patrizia Paggio and Costanza Navarretta. 2010. Feedback in head gesture and speech. In Kipp et al., editor, *Proceedings of LREC-2010*, pages 1–4, Malta, May 17.

Patrizia Paggio and Costanza Navarretta. 2011. Head movements, facial expressions and feedback in Danish first encounters interactions: a culture-specific analysis. In C. Stephanidis, editor, *Universal Access in Human-Computer Interaction. Users Diversity. Proceedings of 6th International Conference, UAHCI 2011, Held as Part of HCI International 2011*, pages 583–590, Orlando, FL, USA, July. Springer.

Patrizia Paggio, Jens Allwood, Elisabeth Ahlsén, Kristiina Jokinen, and Costanza Navarretta. 2010. The NOMCO multimodal nordic resource - goals and characteristics. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Patrizia Paggio. 2006a. Annotating information structure in a corpus of spoken Danish. In *Proceedings of the 5th International Conference on Language Resources and Evaluation LREC2006*, pages 1606–1609, Genova, Italy.

Patrizia Paggio. 2006b. Information structure and pauses in a corpus of spoken Danish. In *Conference Companion of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 191–194, Trento, Italy.

Victor Yngve. 1970. On getting a word in edgewise. In *Papers from the sixth regional meeting of the Chicago Linguistic Society*, pages 567–578.

Unimodal and Multimodal Feedback In Chinese and Swedish Mono-cultural and Intercultural Interactions (a pilot study)

Jia Lu
University of Gothenburg & Chalmers
Göteborg, Sweden
jia.lu@gu.se

Jens Allwood
University of Gothenburg & Chalmers
Göteborg, Sweden
jens@ling.gu.se

Abstract

Communicative feedback in human-human and human-computer interaction is of interest to both language and ICT researchers. In this study, unimodal and multimodal feedback, produced by Chinese and Swedish interlocutors, has been investigated in four Chinese-Chinese, four Swedish-Swedish, and eight Chinese-Swedish informal dyadic video-recorded dialogs. We are investigating two issues: First, what are the typical unimodal and multimodal feedback expressions used by Chinese and Swedes in mono-cultural interactions? Second, what type of feedback do they use when they speak English in inter-cultural interactions? On the basis of our investigation, we describe similarities and differences between Chinese and Swedish participants in using unimodal and multimodal feedback.

Key Words:

Feedback, gestural/vocal-verbal, unimodal/ multimodal, Chinese, Swedish, mono-/inter-cultural interaction

1 Introduction

In this paper, communicative feedback refers to unobtrusive vocal and bodily expressions, which are used to give and elicit information concerning contact, perception, understanding, and emotional/attitudinal reactions to messages from interlocutors. There are a number of previous studies on feedback within the area of Interactive Communication Management (ICM) (Allwood, 2008), analyzing the functions of feedback, describing various ways of producing feedback (Clark & Schaefer, 1989), analyzing affective aspects of feedback (Navarretta, Paggio & Jokinen, 2008; Poggi & Merola, 2003), or exploring the relation between gestural and vocal-verbal feedback in either human-human or human-computer interaction (Allwood, Ahlsén, & Nivre, 1992; Cerrato & Skhiri, 2003). This paper is a pilot study on investigating features of unimodal and multimodal feedback expressions in Chinese and Swedish mono-cultural and intercultural interactions.

2 Purpose

The main purpose of this study is to investigate two issues. First, what are the typical unimodal and multimodal feedback expressions used by Chinese and Swedish communicators in mono-cultural interactions? Second, what feedback expressions are used when they communicate in English in an intercultural setting?

3 Data and Method

The study is based on four Chinese-Chinese, four Swedish-Swedish, and eight Chinese-Swedish video-recordings of face-to-face dyadic dialogs. Four Chinese and four Swedish participants took part in the recordings. The languages used are Chinese, Swedish, and English respectively. The subjects are university students studying in Sweden, and their task is to get acquainted with each other. In order to eliminate as much as possible the influence of factors like prior acquaintance and physical environment, strangers who had no earlier acquaintance were filmed by three video cameras (left-, center-, and right-posited) in a standing position. Each video recording lasts approximately seven to ten minutes, and the entire conversation is analyzed in this study. Information concerning the length of time and the number of words of each transcription is presented in Table 1. Our data was transcribed and checked according to the GTS (Göteborg Transcription Standard) version 6.2 (Nivre, 1999) and manually annotated according to the MUMIN multimodal coding scheme for feedback (Allwood, Cerrato, Jokinen, Navarretta & Paggio, 2007).

Recording	Time length (min.)	No. of words
Chi-chi 1	07:49	1608
Chi-chi 2	06:45	1475
Chi-chi 3	07:12	1571
Chi-chi4	06:30	1432
Total of CN-CN	27:36	6086
Chi-swe 1	11:44	2070
Chi-swe 2	07:56	1380
Chi-swe 3	09:04	1309
Chi-swe 4	10:29	1555
Chi-swe 5	08:11	1122
Chi-swe 6	06:52	983
Chi-swe 7	06:08	943
Chi-swe 8	04:44	678
Total of CN-SE	64:47	10040
Swe-swe 1	06:29	1294
Swe-swe 2	07:01	1604
Swe-swe 3	08:10	1889
Swe-swe 4	08:14	1908
Total of SE-SE	29:54	6695

Table 1: Time length and number of words in the analyzed recordings. In Swedish, words were operationalized as a sequence of graphs between two spaces occurring in transcribed utterances while in Chinese, we used verbal units that have traditionally been regarded as words. CN = Chinese and SE = Swedish

Inter- and intra-coder reliability checking was done between six Chinese and Swedish transcribers and annotators. First one Chinese and two Swedish transcribers/annotators coded a sample of 100 occurrences together in order to establish a common procedure that was used by all transcribers. Each transcription was transcribed as well as coded by one person and then checked by two other persons.

4 Analysis and Results

We will now first present the results concerning the Chinese and Swedish mono-cultural interactions and then turn to the intercultural ones, ending with a summary and comparison of feedback used by Chinese and Swedish in the three types of interactions.

4.1 Feedback in Chinese and Swedish Mono-cultural Interactions

As we can see from Table 2, Swedish interlocutors use more feedback of all types than Chinese interlocutors. In the table, the frequency column provides the number of feedback units of a specific type. A unit can contain more than one contiguous word or gesture or be multimodal with a combination of a word and a gesture, so that e.g. '*ja ja'* ('yes yes') or a '*ja'*+nod is counted as a unit. The per word column is derived by dividing the total number of vocal words in the CN-CN or SE-SE recordings by the total number of feedback units of a particular type in the same recordings. The per minute column is derived similarly by dividing the total number of minutes for the CN-CN and SE-SE recordings by the number of feedback units of a particular type. Thus, Table 2, for instance, shows us that there are 139 vocal-verbal feedback units in the CN-CN recordings and that on an average, there are 5.08 such units per minute and 2.28 units per 100 words.

Modality	Chinese			Swedish		
	Freq.	Per 100 words	Per min.	Freq.	Per 100 words	Per min.
VFB only	139	2.28	5.08	307	4.59	10.27
GFB only	59	0.97	2.16	145	2.17	4.85
Unimodal total	198	3.25	7.24	452	6.75	15.12
VFB+GFB	226	3.71	8.26	267	3.99	8.93
Total	424	6.97	15.50	719	10.74	24.05

Table 2: The use of feedback in four Chinese and four Swedish mono-cultural interactions (GFB= gestural feedback, VFB= vocal-verbal feedback)

4.1.1 Unimodal Gestural FB in Chinese and Swedish Mono-cultural Dialogs

As can be seen in Table 3, below, the most common unimodal gestural feedback expressions in the Chinese-Chinese interactions are nods, smile, gaze sideways, and single nod. Over and above this, there are many unimodal gestural feedback expressions that occur only once or twice. These are lumped together as 'others' in the table.

Unimodal GFB expression	Raw freq.	Per 100 words	Per min.
Nods	18	0.30	0.66
Smile	9	0.15	0.33
Gaze sideways	6	0.10	0.22
Single Nod	3	0.05	0.11
Others (freq.≤2)	23	0.37	0.84
Total	59	0.97	2.16

Table 3: Chinese unimodal gestural FB types,[1] in four mono-cultural Chinese dialogs

In Excerpts 1, 2, and 3 below, we exemplify how nods, smile, and gaze sideways are used by the Chinese subjects to express feedback functions which are coded using the abbreviations C, P and

[1] In this study, unimodal gestural feedback refers to gestural feedback without vocal-verbal accompaniment.

U[2]. Besides this, many feedback expressions also have emotional/attitudinal functions which are coded with the abbreviations E/A, e.g. friendliness and hesitation in Excerpts 2 and 3.

Excerpt[3] 1: (example of Chinese unimodal FB nods)

Original transcription	**English correspondence**
Cf2: <1 | > 1 <2 dui >2	$Cf2: <1 | > 1 <2 right >2
$@ <1 GFB general face: laughter; CPUE/A friendliness/agreement >1 @ <2 VFB; CPUE/A agreement >2	
$Cf1: <1 | >1	$Cf1: <1 | >1
@ < **GFB head: nods; CPU** >	

Excerpt 2: (example of Chinese unimodal FB smile)

Original transcription	**English correspondence**
$Cm2: <1 dui dui dui >1 <2 da san >2 <3 ying gai shi ran hou /// >3	$Cm2: <1 right right right >1 <2 but >2 <3 should be and then /// >3
@ <1 VFB; CPU confirmation >1, <1 GFB head: nod; CPU >1 @ <2 VFB; CPU confirmation >2 @ <3 GFB general face: smile; CPUE/A friendliness >3, <3 head move slightly to the left >3	
$Cf1: < | >	$Cf1: < | >
@ <**GFB general face: smile; CPUE/A friendliness**>	

Excerpt 3: (for Chinese unimodal gaze sideways)

Original transcription	**English correspondence**
$Cm1: ni ke yi xuan ze hao duo zhong lei you furniture dui ba hai you wang ye she ji hai you dong hua she ji	$Cm1: you have many options there are furniture and web design as well as flash or animation design
$Cm2: < | >	$Cm2: < | >
@ < **GFB gaze: sideways**; **CPUE/A hesitation** >	

Nods, smile, single nod, and up-nods are the most common unimodal gestural feedback expressions in the Swedish-Swedish dialogs (cf. Table 4, below). They are sometimes used to express CPU, or CPU with agreement or amusement (see Excerpts 4, 5 and 6).

Unimodal GFB	**Freq.**	**Per 100 words**	**Per min.**
nods	76	11.35	2.54
smile	24	3.58	0.80
single nod	9	1.34	0.30
up-nods	7	1.04	0.24
eyebrow raise	4	0.59	0.13
eyebrow frown	4	0.59	0.13
head shakes	3	0.45	0.10
gaze sideways	3	0.45	0.10
others (freq.≤2)	15	2.31	0.51
Total	145	21.7	4.85

Table 4: Unimodal gestural FB in four Swedish mono-cultural dialogs

Excerpt 4: (example of Swedish unimodal GFB nods)

Original transcription	**English correspondence**
$K: De{t} beror ju på så mycke{t} på vem man < hamnar me{d} också om man trivs me{d} dom sådär >	$K: It also depends so much on who you < end up with if you're happy with them and stuff >
@ **< GFB head: S nods; CPU agreement >**	

Excerpt 5: (example of unimodal Swedish GFB smile)

Original transcription	**English correspondence**
$K:…där föräldrarna skulle skriva under att vi fick e1dricka ett glas vin <3 elle{r} ett / glas cider elle{r} en öl <4//>4 <5 e1 de{t} stoppades>5 >3	$K: … where the parents would sign a paper that we could eh drink a glass of wine <3 or a / glass of cider or a beer <4 // >4 <5 eh it was stopped >5 >3
@ **<3 GFB general face: J smile; CPUE/A amusement >**3 @ <4 general face: chuckle >4 @ <5 GFB eyebrows: J raise; CPUE/A surprise >5	

Excerpt 6: (for Swedish unimodal GFB up-nods)

Original transcription	**English correspondence**
$S: … å0 så / <2 sa han att han behövde svenskar >2 <3//så då>3	$S: … and then / <2 he said that he needed swedes >2 <3 // so then >3
@ <2 GFB head: L nods; CPU >2 @ <3 **GFB head: L up-nods; CPU >**3, <3 head start: nods >3	

4.1.2 Unimodal Vocal-verbal FB in Chinese-Chinese and Swedish-Swedish Dialogs

The most frequent vocal-verbal FB expressions in Chinese mono-cultural dialogs are 'dui' ('right'), 'a:' ('ah:/ yeah'), 'en' ('yes/ right/ ok'), and 'a' ('ah/ yes') (see Table 5). 'Dui' ('right'), 'a' ('ah:/yeah'), and 'en' ('yes/right/ok') are used to express CPU, and sometimes to confirm or

[2] CPU refers to willingness/ability to continue (C), perceive (P) and understand (U) the communicated information.

[3] The excerpts in this paper are extracted from the transcriptions of the studied data. In GTS, $ identifies a speaker. Angular brackets < > indicate the scope of a comment, and the number identifies a corresponding comment. The symbol @ initiates the corresponding comment. The number of slashes (/, //, ///) indicate length of a pause. Curled brackets{ } contains letters of the written word form that were not pronounced in the spoken form. < | > indicates a pause where communicative gestures are inserted. Colon : indicates prolongation of a sound. FB = feedback, VFB = vocal-verbal feedback, GFB = gestural feedback. CPUE/A = contact, perception, understanding, emotion/ attitude (see CPU in Footnote 3).

agree ‘yes, you are right’ (cf. Excerpts 7, 8, and 9).

VFB	Translation	Freq.	Per 100 words	Per min.
dui	right	21	0.35	0.77
a:	ah:/ yeah	12	0.20	0.44
en	yes/right/ok	10	0.16	0.37
a	ah/ yes	8	0.13	0.29
others (freq.≤2)		88	1.44	3.21
Total		139	2.28	5.08

Table 5: Unimodal vocal-verbal FB used in four Chinese mono-cultural dialogs

Excerpt 7: (example of Chinese unimodal ‘dui’)

Original transcription	English translation
$Cm2: ta men ke neng /// ta men ying gai ye kao lv na ge ba ///	$Cm2: they may /// they should also think about that I think ///
$Cm1: <1 **dui** >1 ...	$Cm1: <1 **right** >1 ...
@ <1 **VFB; CPU agreement** >1...	

Excerpt 8: (example of Chinese unimodal ‘a’)

Original transcription	English translation
$Cm1: na ni shao shu min zu	$Cm1: then you are from minority nationality
$Cf2: <1 **a** >1 <2 meng zu >2	$Cf2: <1 **yes** >1 <2 Mongolian >2
@ <1 **VFB; CPU confirmation** >1 @ <2 comment: answer to the question >2	

Excerpt 9: (example of Chinese unimodal ‘en’)

Original transcription	English translation
$Cm1: … jia zhang ke neng you yi xie wen ti	$Cm1: …our parents may have some problems
$Cf2: <1 **en** >1 <2 ni shi na li ren >2	$Cf2: <1 **yes** >1 <2 where are you from >2
@ <1 **VFB; CPU** >1 @ <2 eliciting >2, <2 eye brow raise >2	

The most common Swedish unimodal vocal-verbal feedback expressions are ‘{j}a’ (‘yeah’), ‘m’ (‘uhu’), ‘nä’ (‘no’), ‘okej’ (‘ok’), and ‘ja’ (‘yes’) (see Table 6). As can be seen from Excerpts 10 and 11, ‘{j}a’ (‘yeah’) and ‘m’ (‘uhu’) can be used to express CPU with agreement or hesitation.

VFB & ‘*translation*’	F.	Per 1000 words	Per min.
{j}a ‘*yeah*’	80	11.95	2.68
m ‘uhu’	45	6.72	1.51
nä ‘*no*’	14	2.09	0.47
okej ‘*ok*’	12	1.79	0.40
ja ‘*yes*’	11	1.64	0.37
hja ‘*yes*’	9	1.34	0.30
jo ‘*yes*’ (disagreement w. negative statement)	6	0.90	0.20
{j}a: ‘*yeah*’	6	0.90	0.20
m: ’uhu’	6	0.90	0.20
oj ‘*whoops-wow-really?*’	5	0.75	0.17
{j}a jo ‘*yes-I agree*’	4	0.60	0.13
{j}a {j}a ‘*yeah yeah*’	3	0.45	0.10
ja elle{r} hu{r} ‘*yes is that not right*’	3	0.45	0.10
Others (freq.≤2)	103	15.42	3.44
Total	307	45.90	10.27

Table 6: Swedish Unimodal vocal-verbal FB

Excerpt 10: (Use of the Swedish unimodal vocal FB word '{j}a')

Original transcription	English correspondence
$K: de{t} tror ja{g} e0 väldi{g} klokt $S: < **{j}a** >	$K: i think that's very wise $S: < **yeah** >
@ < **VFB; CPUE/A agreement >**	

Excerpt 11: (Use of the Swedish unimodal vocal FB word 'm')

Original transcription	English correspondence
$S: ja{g} vill e1 komma in hä{r} // så $L: < **m** >	$S: i want to eh be get in here // so $L: < **uhu** >
@ < **VFB; CPUE/A thoughtful/ hesitation >**	

4.1.3 Multimodal Feedback in Chinese and Swedish Mono-cultural Interactions

The multimodal vocal-verbal plus gestural feedback expressions used in the Chinese and Swedish mono-cultural interactions are shown in Table 7. The most common multimodal feedback units used by the Chinese speakers are ‘en’ (‘yes/right/ok’) +nods, laughter [4], ‘a’ (‘ah/ yes’)+nods, ‘en’ (‘yes/ right/ok’)+nod, and chuckle[5]. Instances of ‘a’ (‘ah/yeah’)+nods and ‘en’ (‘yes/right/ok’)+nods are presented in Excerpt 12. These multimodal feedback units are

[4] Laughter is regarded as one multimodal unit, consisting of sound and facial gesture.
[5] Chuckle is also treated as a multimodal unit.

primarily used to express CPU, and sometimes, in addition, with confirmation or agreement.

VFB & *translation*	GFB	F.	Per 100 words	Per min.
en *'yes/right/ok'*	nods	30	0.49	1.10
laughing	laughter	16	0.26	0.58
a 'ah/ yes'	nods	15	0.25	0.55
en 'yes/right/ok'	nod	8	0.13	0.29
chuckling	chuckle	6	0.10	0.22
a 'ah/ yes'	nod	4	0.07	0.15
dui 'right'	nods	4	0.07	0.15
a: 'ah:/ yeah'	nods	3	0.05	0.11
e 'eh'	smile	3	0.05	0.11
Others (frequency⩽2)		137	2.24	5.00
Total		226	3.71	8.26

Table 7: Multimodal feedback used in four Chinese mono-cultural dialogs (F.=raw frequency, w=word, m=minute)

Excerpt 12: (Chinese multimodal feedback units 'a' ('ah/yes')+nods and 'en' ('yes/right/ok')+nods)

Original transcription	**English correspondence**
$Cf1: … ni men ke neng zai er lou ba shi bu shi // $Cf2: <1 **a** /// >1 wo men ying gai jiu yi qian jiu zong zai si lou ran hou /// wo ying gai <2 zhe bu shi suan di er nian ma >2	$Cf1: … you are on the second floor aren't you // $Cf2: <1 **yes** /// >1 before we used to be on the second floor and then /// I should be <2 this is my second year so >2
@ <1 **VFB; CPU confirmation >1, <1 GFB head: nods; CPU confirmation** >1 @ <2 eliciting >2	
$Cf1: < **en** >	$Cf1: < **yes** >
@ < **VFB; CPUE/A agreement >, < GFB head: nods; CPUE/A agreement R** >	

The most common multimodal feedback units in the Swedish-Swedish dialogs (cf. Table 8) are: 'm' ('uhu')+nods, chuckle, {j}a ('yeah')+ nods, and {j}a ('yeah')+up-nods. Examples are given in Excerpts 13, 14, and 15.

Excerpt 13: (for Swedish multimodal FB unit 'm'+nods)

Original transcription	**English correspondence**
$S: nä men de{t} gick bra så men e1 vi va{r} verkligen oj: // $L: < **m** >	$S: no but it went well so eh we were really like wo:w // $L: < **okay** >
@ **<VFB; CPUE/A empathy>, <GFB head: nods; CPU>**	

VFB expression		GFB expression	Raw Freq.	Per 1000 words	Per min.
Swedish	**Translation**				
m	uhu	nods	20	2.99	0.67
chuckle	(chuckle)	chuckle	14	2.09	0.47
{j}a	yeah	nods	13	1.94	0.44
{j}a	yeah	up-nod	10	1.49	0.33
{j}a	yeah	nod	9	1.34	0.30
m	uhu	up-nod	8	1.19	0.27
laughter	(laughter)	laughter	7	1.05	0.23
ja	yes	nod	7	1.05	0.23
{j}a	yeah	up-nods	6	0.90	0.20
m	uhu	up-nods	5	0.75	0.17
m	uhu	nod	4	0.60	0.13
okej	okay	up-nod	4	0.60	0.13
{j}a	yeah	smile	3	0.45	0.10
{j}a	yeah	tilt	3	0.45	0.10
{j}a okej	yeah okay	nods	3	0.45	0.10
ja	yes	nods	3	0.45	0.10
mhm	uhuh	up-nods	3	0.45	0.10
Others (frequency⩽2)			145	21.66	4.86
Total			267	39.90	8.93

Table 8: Multimodal feedback used in four Swedish mono-cultural dialogs

Excerpt 14: (Swedish multimodal unit '{j}a'+up-nod)

Original transcription	**English correspondence**
$L: … de{t} e1 blir kontor då för dig eller	$L: … it'll eh be the office for you then right
$J: < **{j}a** >	$J: < **yeah** >
@ **< VFB; CPUE/A confirmation >, < GFB head: up-nod; CPUE/A confirmation R >**	

Excerpt 15: (Swedish multimodal unit '{j}a'+nods)

Original transcription	**English correspondence**
$S: ja{g} vill komma … $K: <1 **{j}a** >1 då e0 de{t} svårt <2 | >2	$S: i want to come … $K: <1 **yeah** >1 then it's hard <2 | >2
@ <1 **VFB; CPU** >1, <1 **GFB head: nods; CPU** >1 @ <2 general face: chuckle >2	

4.2 Feedback in Chinese-Swedish Intercultural Interactions

Below, we present the unimodal and multimodal feedback expressions used by four Chinese and four Swedish participants in eight Chinese-Swedish intercultural interactions.

Modality	Chinese			Swedish		
	F.	Per 1000 words	Per min.	F.	Per 1000 words	Per min.
VFB only	203	20.22	3.13	138	13.79	2.13
GFB only	165	16.43	2.55	178	17.73	2.75
Unimodal total	368	36.65	5.68	316	31.47	4.88
VFB+GFB	250	24.90	3.86	354	35.26	5.46
Total	618	64.54	9.54	670	66.73	10.34

Table 9: Chinese and Swedish uses of feedback in eight intercultural interactions (F.= frequency)

Table 9 shows that the Swedish participants, overall, in the intercultural dialogs, give more feedback than the Chinese participants (670–618). Specifically, the Swedes give more multimodal feedback and slightly more unimodal ges-

tural feedback, while the Chinese give more unimodal vocal-verbal feedback.

4.2.1 Unimodal Gestural FB in Chinese-Swedish Intercultural Interactions

The Swedes used slightly more unimodal gestural feedback than the Chinese in their intercultural interactions (see Table 10). The most frequent unimodal gestural feedback expressions used by both Chinese and Swedish speakers were: nods, single nod, smile, and up-nod. They are used to express CPU, or CPU with confirmation, agreement, or other emotions[6] (see Excerpt 16).

Chinese				Swedish			
GFB	**F.**	**Per 1000 words**	**Per min.**	**GFB**	**F.**	**Per 1000 words**	**Per min.**
nods	89	8.86	1.37	nods	117	11.65	1.80
nod	20	1.99	0.31	nod	12	1.20	0.19
smile	18	1.79	0.28	up-nods	10	1.00	0.15
up-nod	11	1.10	0.17	smile	9	0.90	0.14
head shakes	4	0.40	0.06	up-nod	8	0.80	0.12
head tilt	4	0.40	0.06	eyebrow raise	3	0.30	0.05
up-nods	3	0.30	0.05	Others (F.≤2)	19	1.88	0.3
others(F.≤2)	16	1.59	0.25				
Total	165	16.43	2.55	**Total**	178	17.73	2.75

Table 10: Unimodal gestural FB in Chinese-Swedish intercultural interactions (F.=frequency)

Excerpt 16: (for (co-activated) unimodal up-nod)

$Cf2: i also co{me} from // in+ inner mongolia yeah (you know)
$Sf2: <1 mhm >1 <2 | >2
@ <1VFB; CPUE/A surprise/interest>1
@ <2**GFB head: up-nod; CPUE/A surprise/interest** R>2, <2**GFB head: L up-nod**; CPU>2

4.2.2 Unimodal Vocal-verbal FB in Intercultural Interactions

The Chinese participants used more unimodal vocal-verbal feedback than the Swedish in the Chinese-Swedish dialogs. The most common unimodal vocal-verbal feedback expressions used by both Chinese and Swedish participants are: 'yeah', 'okay', and 'm', expressing CPU, or CPU with agreement (see below Table 11).

[6] Emotions and attitudes of feedback expression, such as surprise, politeness, embarrassment, uncertainty, certainty, amusement, happiness, agreement, disagreement, and so on, have been found and coded in our data. However, in the present study, only a few of them are presented in the examples.

Chinese				Swedish			
VFB	**F.**	**Per 1000 words**	**Per min**	**VFB**	**F.**	**Per 1000 words**	**Per min**
yeah	60	5.98	0.93	yeah	36	3.59	0.56
okay	25	2.49	0.39	m	17	1.69	0.26
m	14	1.39	0.22	okay	15	1.49	0.23
yes	9	0.90	0.14	ah	7	0.70	0.11
uhu	7	0.70	0.11	Others (F.≤5)	63	6.32	0.97
yeah yeah yeah	6	0.60	0.09				
Others (F.≤5)	82	8.16	1.25				
Total	203	20.22	3.13	**Total**	138	13.79	2.13

Table 11: Unimodal (English) vocal FB words used by Chinese and Swedish participants in Chinese-Swedish interactions (F.=frequency)

4.2.3 Multimodal Feedback in Chinese-Swedish Intercultural Interactions

In the Chinese-Swedish interactions, the Swedish participants used more multimodal feedback than the Chinese. The Chinese participants used chuckle and laughter to express CPU with amusement or friendliness, 'yeah'+ nod and 'yeah'+nods to express CPU or CPU with confirmation or agreement, as the most common multimodal feedback units; while, the Swedish participants used 'yeah'+nods, 'm'+ nods, and chuckle most frequently (see Table 12).

Chinese				Swedish			
VFB+GFB	**F.**	**Per 1000 words**	**Per min**	**VFB+GFB**	**F.**	**Per 1000 words**	**Per min**
chuckle	28	2.79	0.43	yeah+nods	45	4.48	0.69
yeah+nod	23	2.29	0.36	m+nods	25	2.49	0.39
yeah+nods	17	1.69	0.26	chuckle	18	1.79	0.28
laughter	10	1.00	0.15	m+up-nods	9	0.90	0.14
okay+nods	9	0.90	0.14	yeah+nod	9	0.90	0.14
mhm+nod	8	0.80	0.12	yeah+up-nods	8	0.80	0.12
okay+nod	7	0.70	0.11	okay+up-nod	7	0.70	0.11
mhm+nods	6	0.60	0.10	yeah+up-nod	7	0.70	0.11
Others (F.≤5)	142	14.13	2.19	laughter	6	0.60	0.09
				m+up-nod	6	0.60	0.09
				Others (F.≤5)	214	21.30	3.30
Total	250	24.90	3.86	**Total**	354	35.26	5.46

Table 12: Multimodal FB units used by Chinese and Swedish in the Chinese-Swedish interactions (F.=frequency)

5. Discussion

Feedback in the Chinese and the Swedish monocultural interactions is discussed first, followed by the Chinese-Swedish intercultural interactions.

5.1 Mono-cultural Interaction

We have already seen (cf. Table 2) that the Swedish participants, in the mono-cultural interactions, used all types of feedback expressions more than the Chinese participants. They used unimodal feedback more than twice as many times as the Chinese participants both gesturally and vocal-verbally (with a frequency of 307 compared to 139 and 145 to 59) (Table 2), and they also used slightly more multimodal feedback expressions than the Chinese (267 to 226). This clearly suggests that the Swedish participants use both more unimodal and multimodal feedback than the Chinese in the mono-cultural first acquaintance dialogs.

If we turn to similarities, both Chinese and Swedish participants used nods, single nod, and smile as the most common type of unimodal gestural feedback to express CPU in mono-cultural interactions, sometimes with an additional function of confirmation or other emotional/ attitudinal functions such as agreement or/and friendliness. Another similarity is that both Chinese and Swedish participants used chuckle as the most frequent type of multimodal feedback. Possibly, this is because both Swedes and Chinese want to show friendliness and agreement, in a first encounter.

Regarding differences, the Swedish participants used up-nods very often in mono-cultural interactions, while the Chinese participants rarely used this in Chinese-Chinese dialogs. The Chinese participants gazed sideways very frequently to express hesitation or uncertainty in the mono-cultural interactions, probably because of the insecurity or uncertainty that they may feel in a first acquaintance dialog. The Swedish participants did not gaze sideways as much as the Chinese in mono-cultural dialogs. This might be because gazing sideways is not used to express hesitation or uncertainty in Swedish communication, or because the Swedish participants felt more secure when they were filmed for this project in Sweden.

Concerning vocal-verbal feedback, Chinese 'dui' ('right' in English), 'a:' ('ah:/ yeah'), 'en' ('yes/ right/ ok'), 'a' ('ah/ yes'), and Swedish '{j}a' ('yeah'), 'm' ('yes/I agree'), 'nä' ('no'), 'okej' ('okay'), and 'ja' ('yeah') are the most common unimodal vocal-verbal feedback expressions used by Chinese and Swedish participants in mono-cultural interactions. Regarding multimodal feedback, the Chinese participants used 'en' ('yes/right/ok')+nods, laughter, 'a' ('ah/yes')+nods, and 'en' ('yes/ right/ ok')+nod as the most common multimodal feedback units, while 'm' ('uhu')+nods, '{j}a' ('yeah')+nods, and '{j}a' ('yeah')+up-nods are the most frequent Swedish multimodal units.

5.2 Intercultural Interaction

In the Chinese-Swedish intercultural interactions, Chinese participants used more unimodal vocal-verbal feedback than Swedes (203 compared to 138, see Table 9). However, the Swedish participants used slightly more unimodal gestural and more multimodal feedback expressions than the Chinese (178 to 165, and 354 to 250). Overall, Chinese participants seem to increase their feedback in the intercultural situation, while the Swedes decrease theirs.

Regarding similarities, the most frequent unimodal gestural feedback for both Chinese and Swedish participants are: nods, nod, smile and up-nod. However, as we have already noted, Chinese did not use up-nod at all in their mono-cultural interactions, but used this gesture in the intercultural interactions. This change is probably due to the adaptation and co-activation with the Swedish interlocutors. Chinese and Swedish participants both used 'yeah', 'okay', 'm' as the most common unimodal vocal-verbal feedback, and chuckle and 'yeah'+nods as the most common multimodal feedback.

Concerning differences, in the intercultural interactions, besides chuckle and 'yeah'+nods, the Chinese participants used laughter and 'yeah'+nod as the most frequent multimodal feedback; whereas, for the Swedish participants 'm'+nods was the most common.

Thus, both Chinese and Swedish participants showed more similarities in intercultural interactions than in mono-cultural interactions. Probably, this is because they were mutually influencing each other, and co-activation, therefore was possible.

6. Conclusions

This paper primarily addresses two questions, i.e. what are the typical unimodal and multimodal feedback expressions used by Chinese and Swedish speakers in mono-cultural interactions, and what expressions do they use when communicating in English in intercultural interactions.

In mono-cultural interactions, we found that Swedish participants used more unimodal and multimodal feedback than Chinese participants. In these interactions, both Chinese and Swedish participants used nods, single nod, and smile as the most common unimodal gestural feedback, and chuckle as the most frequent type of multimodal feedback. Concerning unimodal gestural feedback, gaze sideways is typical of Chinese feedback behavior, and up-nod(s) are typical of Swedish behavior. Chinese 'dui' ('right' in English), 'a:' ('ah:/ yeah'), 'en' ('yes/ right/ ok'), 'a' ('ah/ yes'), and Swedish '{j}a' ('yeah'), 'm' ('yes/I agree'), 'nä' ('no'), 'okej' ('okay'), and 'ja' ('yeah') are the most common unimodal vocal-verbal feedback expressions. Besides chuckle, Chinese participants used 'en' ('yes/ right/ ok')+nods, laughter, 'a' ('ah/ yes')+nods, and 'en' ('yes/ right/ ok')+nod as the most common type of multimodal feedback, and Swedes used 'm' ('yes-I agree')+nods, '{j}a' ('yeah')+nods, and '{j}a' ('yeah)+up-nods most frequently.

In the Chinese-Swedish intercultural interactions, possibly because of second language interference, Chinese participants used more unimodal vocal-verbal feedback than the Swedish participants. However, the Swedish participants used more multimodal feedback and slightly more unimodal gestural feedback than the Chinese. Regarding similarities, both the Chinese and Swedish participants most frequently used the following types of unimodal gestural feedback; nods, single nod, smile, up-nod, and types of unimodal vocal-verbal feedback; 'yeah', 'okay', 'm', and multimodal feedback; chuckle and 'yeah'+nods.

Besides chuckle and 'yeah'+nods, the Chinese participants used laughter and 'yeah'+nod, while the Swedish participants used 'm'+nods as the most frequent multimodal feedback.

Finally, we note that since the size of this study is relatively small, it still necessitates further study.

Acknowledgement:

We would like to thank VR (The Swedish Research Council), NOS-HS (The Nordic Research Council for Humanities and Social Sciences), and Elisabeth Ahlsén, Alexander Holender, Karl Johan Sandberg, and Yansi Xu at the SCCIL Interdisciplinary Research Center, University of Gothenburg. We also thank our reviewers for valuable comments.

References

Allwood, J. (2008). Dimensions of embodied communication - towards a typology of embodied communication. *Wachsmuth, Ipke: Lenzen, Manuela & Knoblish, Günther (eds.).Embodied Communication in Humans and Machines*. *Oxford University Press*. pp. 257-281.

Allwood, J., Ahlsén, E., Nivre, J. (1992). On the Semantics and Pragmatics of Linguistic Feedback. *Journal of Semantics* 9, 1-26.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C. & Paggio, P. (2007). The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing. In J. C. Martin et al. (eds) *Multimodal Corpora for Modelling Human Multimodal Behaviour*. Special issue of the International Journal of Language Resources and Evaluation. Springer. Vol.41, no.3-4, pp.273–287.

Cerrato, L. & Skhiri, M. (2003). Analysis and measurement of communicative gestures in human dialogues. *Proceedings of AVSP 2003,* 251-256. France.

Clark, H. & Schaefer, E. (1989). Contributing to Discourse. *Cognitive Science,* 13, 259-294.

Navarretta, C., Paggio, P & Jokinen, K. (2008). Distinguishing the communicative functions of gestures. In A. Popescu-Belis & R. Stiefelhagen (eds.) *Proceedings of 5th Joint Workshop on Machine Learning and Multimodal Interaction*, Utrecht, September 2008, Springer, 38-49.

Nivre, J. (1999) *Göteborg Transcription Standard. Version 6*.2, pp. 38. Göteborg: Göteborg University, Department of Linguistics.

Poggi, I. & Merola, G. (2003). Multimodality and Gestures in the Teacher's Communication. In *Gesture-Based Communication in Human-Computer Interaction*. *5th International Gesture Workshop, GW 2003,* Genova, Italy. pp. 405-406. Springer Berlin: Heidelberg.

Observations on Listener Responses from Multiple Perspectives

Iwan de Kok
Human Media Interaction
University of Twente
i.a.dekok@utwente.nl

Dirk Heylen
Human Media Interaction
University of Twente
heylen@utwente.nl

Abstract

In this paper we present three studies that investigate the individual differences in nonverbal listening behavior. Besides collecting a corpus of listener responses, we asked people to watch a video of a speaker and indicate where they would produce a listener response. Also we asked people to judge the appropriateness of listener responses that we generated using a virtual human. The combination of the multiple perspectives collected in these studies provides us with a rich data set in which different types of response opportunities are distinguishable. There are moments where there is *high agreement* between these multiple perspectives that a listener response is appropriate or inappropriate, moments where a listener response is *controversial* and moments neither a response was given nor a response was judged inappropriate (*neutral*). We will show that different contextual characteristics can be used to discriminate these response opportunities. Observations show relations to sentence structure, conversational structure and proximity of earlier responses.

1 Introduction

In a conversation humans highly coordinate their behavior to transfer information from one to another. In this interaction not only the behavior of the speaker guides the conversation, but the responses from the listener to the contributions of the speaker do so as well (Yngve, 1970; Kraut et al., 1982; Bavelas et al., 2000). These listener responses can take the shape of nonverbal behaviors such as head nods, head shakes and facial expressions, and verbal expressions, such as "hmmm" and "yeah". The function of these listener responses is to signal the state of mind of the listener towards the speaker, conveying whether the contributions of the speaker are attended to, understood, agreed upon and/or affective attitudes towards the contributions (Allwood et al., 1992; Clark, 1996).

Our interest in this behavior comes from the goal to build embodied conversational agents which can interact as if they are a human. A model of these listener responses is one of the components needed to achieve the same kind of coordinated interaction as humans have. A challenge in the achievement of this goal is the optional characteristic of listening behavior, which causes high variation in the type, timing and amount of listener responses between individuals. One missed opportunity for a listener responses will not immediately break the interaction, but the total absence of this behavior will. The question is which moments are essential to respond to as a listener and which ones can be passed up. And what are the characteristics of the moments where listener responses is inappropriate?

In this paper we will present three studies that capture the individual differences in nonverbal listening behavior by combining multiple (positive and negative) perspectives. In the first study a corpus is recorded with three listeners in parallel interaction with the same speaker, which gives us three positive perspectives on appropriate moments for listening behavior. In the second study we collect extra positive perspectives on appropriate listening behavior through the parasocial consensus sampling method. In the third and final experiment we collect multiple (negative) perspectives on *in*appropriate behavior by generating listening behavior and let participants judge the appropriateness of each individual listener response. By combining the data of these three studies some moments stand out by either *high agreement* between multiple perspectives (positive or negative), *controversial* perspectives on the appropriateness (positive and negative responses at the same mo-

ment) or *neutral* moments (neither positive nor negative responses). We end the paper with a discussion of these types of moments in our data and with recommendations based on our observations to improve the state-of-the-art of predictive models for listener responses.

2 Study 1: Parallel Recording

In the first study we recorded a corpus aimed at capturing the variation and similarities in listening behavior between people. In traditional corpora to study nonverbal listening behavior an interaction between two people is recorded. The listening behavior in reaction to the speaker is regarded as the ground truth. However another individual placed in the same interaction will most likely not act in the same way. He/She will provide listener responses at different times or use different type of listener responses.

By collecting multiple perspectives we are able to analyze the optionality of listener responses. Our hypothesis is that by combining multiple perspectives one can find moments where a response is given in all perspectives, moments where a response is given in some perspectives and moments where no response is given at all. In the first case, it is probably mandatory for a virtual agent to produce a response, in the second case it might be optional and in the third case it seems better to avoid giving a response. The following section explains the experiment resulting in the recording of the MultiLis corpus in which multiple listeners are recorded in interaction with the same speaker.

2.1 Procedure

The MultiLis corpus (de Kok and Heylen, 2011b) is a Dutch spoken multimodal corpus of 32 mediated face-to-face interactions totaling 131 minutes. Participants (29 male, 3 female, mean age 25) were assigned the role of either speaker or listener during an interaction. In each session four participants were invited to record four interactions. Each participant was once speaker and three times listener.

What is unique about this corpus is the fact that it contains parallel recordings of three individual listeners in interaction with the same speaker, while each of the listeners was tricked into believing to be the sole listener. The speakers saw only one of the listeners, believing that they had a one-on-one conversation. All listeners were placed in a cubicle and saw the speaker on the screen in front of them. The camera was placed behind an interrogation mirror, positioned directly behind the position on which the interlocutor was projected. This made it possible to create the illusion of eye contact.

To ensure that the illusion of a one-on-one conversation was not broken, interaction between participants was limited. Speakers and listeners were instructed not to ask for clarifications or to elicit explicit feedback from each other, so no turn-switching would take place. The speaker received a task of either watching a short video clip before the interaction and summarizing it to the listener, or learning a recipe in the 10 minutes before the interaction and reciting it to the listener. The listener needed to remember as many details of what the speaker told as possible, since questions about the content were asked afterwards.

2.2 Annotation

The recordings of each listener were annotated by one annotator on listening behavior. Each listener has her/his own (perspective on) listening behavior. To study the variety and similiarities in these perspectives one annotator grouped simultaneous listener responses in reaction to the same context. We call the timeframe they span from the first response to that context to the last response the *response opportunity*. Thus, response opportunity can be defined as the window of opportunity to provide a response to a specific context in an interaction.

2.3 Results

The MultiLis corpus contains 2796 listener responses. These listener responses are reactions to 1735 response opportunities. Of these responses opportunities 1142 have one response, 456 have two responses and 128 have responses from all three listeners.

Figure 1 represents a segment of 48 seconds from one of the interactions. It shows the distribution of response opportunities in this segment. On the horizontal axis time is represented. The response opportunities in these 48 seconds found in the MultiLis corpus are indicated with as magenta bars. The height of these bars represent the amount of recorded listeners that gave a response at this response opportunity.

The segment is taken from an interaction where agreement between listeners is quite high. In this

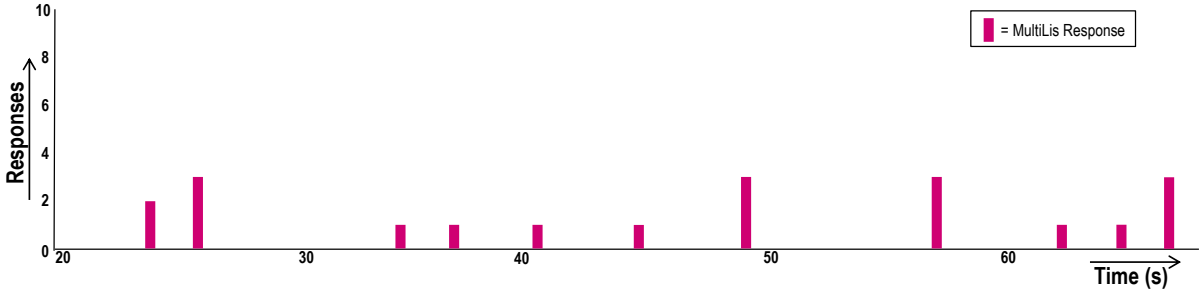

Figure 1: Sample of the distribution of responses in the MultiLis Corpus.

segment there are four response opportunities with three listener responses, one with two listener responses and six with one listener response. No listener has performed a listener response at all these response opportunities. This illustrates that with this corpus we have a more complete view of all the opportunities for a listener response.

In the remainder of the paper we will take a closer look at this segment. We will see how new perspectives correlate with the recorded behavior, how the response opportunities correlate with inappropriate moments and what the characteristics of these response opportunities are. Does the speaker explicitly elicit the listener responses at response opportunities with high agreement or are there other causes?

3 Study 2: Parasocial Consensus Sampling

In the previous study we recorded a corpus where three listeners listened and responded to the same speaker. What if we had more listeners? We would get an even more complete view of all the opportunities for a listener response. There may still be moments that all three listeners have passed up, while a listener response would still be appropriate. The discrimination between mandatory response opportunities, option response opportunities and inappropriate moments to provide a response would also be more clear.

With the Parasocial Consensus Sampling method (Huang et al., 2010b) this is actually possible. In this method multiple participants watch the video recording of the speaker and they indicate through a keyboard when they would give a listener response. We have used this method to collect 8 new (PCS-)perspectives for a subset of the MultiLis corpus.

3.1 Procedure

The collection of PCS perspectives is performed on 8 interactions from the MultiLis corpus. Ten months after the original MultiLis experiments we reinvited 6 of the original listeners in these 8 interactions to collect their PCS perspectives for the same interactions. While watching and listening to the 3 recordings of the same speakers they listened to earlier, they gave responses through the keyboard. Each time they would give a listener response they were instructed to press the spacebar of the keyboard.

Furthermore we invited 10 new participants to collect their PCS perspectives to these interactions. Each of these participants gave their PCS perspectives on 4 interactions. Thus, for each of the 8 interactions, we have 3 original listener perspectives and 7 or 8 PCS-perspectives. From these perspectives there are 5 perspectives from the new participants and 2 or 3 perspectives from the original listeners, depending on whether one of them was the speaker in that interaction or not.

3.2 Results

The 8 interactions used in this study contain 347 response opportunities of which 202 with one response, 98 with two responses and 47 with three responses as identified using the annotations of the three listeners in the corpus. Adding the new PCS perspectives increases the amount of response opportunities identified to 582 response opportunities. The distribution of the amount of responses to each response opportunity is shown in the histogram in Figure 3.

Most response opportunities have only a few responses, but there are still 15 response opportunities with 9 responses, 3 with 10 responses and 3 with 11 responses. We will take a closer look at these response opportunities in Section 5.

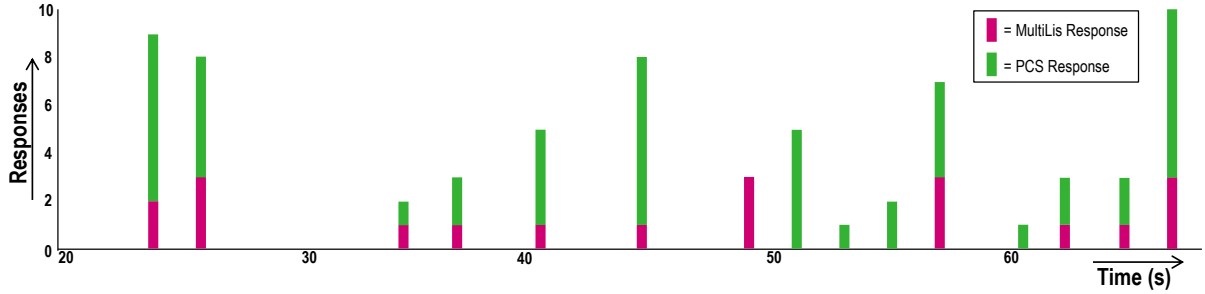

Figure 2: Sample of the distribution of responses in the MultiLis Corpus and PCS responses.

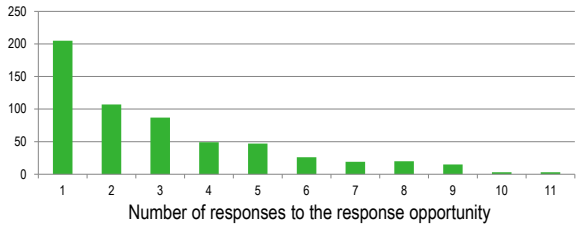

Figure 3: Histogram of the number of (MultiLis and PCS) responses to each response opportunity.

Figure 2 represents the same 48 seconds from the previous study. In green the responses from the collected PCS-perspectives are added to the responses from the MultiLis corpus. The participants provided a PCS-response to almost all the response opportunities found in the previous study with the exception of the response opportunity just before 50 seconds. Interestingly this response opportunity was responded to by each listener in the MultiLis corpus. Furthermore there are 4 new response opportunities of which one was responded to by 5 participants.

4 Study 3: Individual Perceptual Evaluation

With the previous two studies we have compiled a more complete picture of the response opportunities in the interactions than a traditional corpus does by collecting multiple (positive) perspectives. We have identified 582 moments where giving a listener response is appropriate according to at least one individual. Does this mean that every other moment is an inappropriate moment to give a listener response? And are listener responses given at these moments appropriate according to everyone?

To answer these questions we use the Individual Perceptual Evaluation method. In this method we generate virtual listening behavior in reaction to a recorded speaker and let participants judge for each generated listener response, whether this response was appropriate or not. We thus collect a negative perspective on listener responses, which tells us the inappropriate timing of listener responses. In the following we will explain the method and the used stimuli in more detail.

4.1 Stimuli

We presented subjects with clips of a speaker from the MultiLis corpus in interaction with a virtual listener, animated using the BML realizer Elckerlyc (van Welbergen et al., 2010). We used the same 8 interactions as in the previous study. The virtual listener performs only head nods (and everytime the same head nod). The timing of the head nods is based on the multiple perspectives from the previous studies.

182 head nods are generated at appropriate times and 90 head nods are generated at *not*-appropriate times according to these perspectives. The appropriately timed head nods (or *at-head-nods*) are performed at the times where at least 4 perspectives agreed that this is an appropriate time to provide a listener response. The 90 *not*-appropriately timed head nods (or *between-head-nods*) are placed in the biggest gaps between the *at-head-nods*. Within these biggest gaps they are placed in the biggest gap between the moments where at most 3 perspectives agreed to be an appropriate time to provide a listener response.

4.2 Procedure

We invited 8 participants to watch the interactions between the speaker and the virtual listener. They were asked to judge each head nod on appropriateness. When a head nod was inappropriate according to their judgment they pressed the spacebar on a keyboard (a *yuck response*). The participant had the option to replay the video.

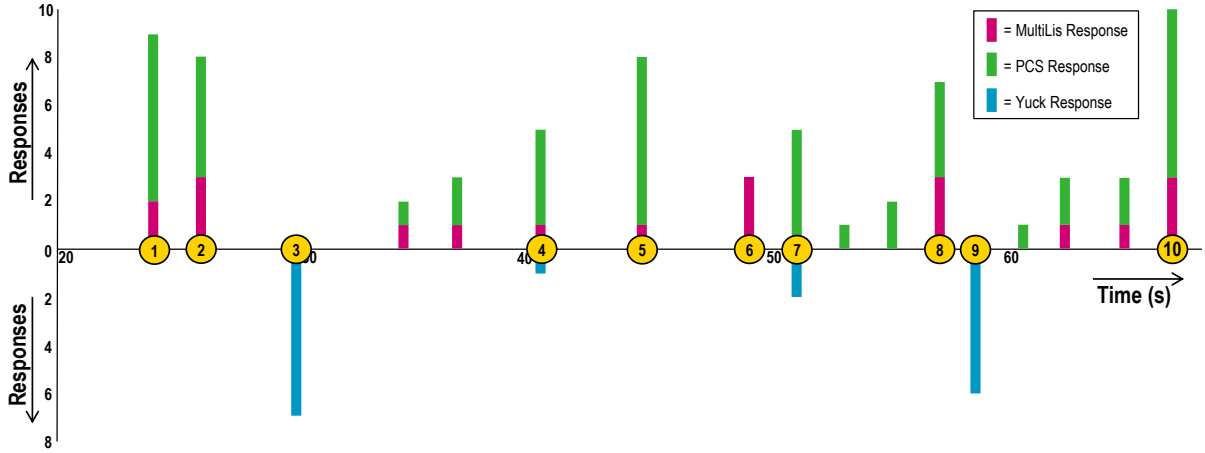

Figure 4: Sample of the distribution of responses in the MultiLis Corpus, PCS responses and the yuck responses. The numbers in the yellow circle correspond to the transcript in Table 1.

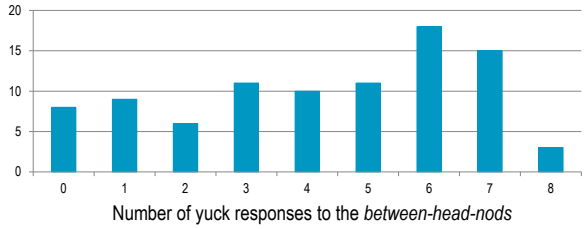

Figure 5: Histogram of the number of yuck responses to each *between-head-nod*.

4.3 Results

On average each participant judged 53 out of 272 head nods as inappropriate, for a total of 424 yuck responses. 42 yuck responses were in reaction to *at-head-nods* and 382 were in reaction to *between-head-nods*. The 42 yuck response in reaction to *at-head-nods* were in reaction to 29 individual *at-head-nods*. 4 of these *at-head-nods* were yucked 3 times, 5 were yucked 2 times and the other 22 were yucked once.

Figure 5 shows the histograms of the 379 yuck responses in reaction to the *between-head-nods*. For each of the 90 generated *between-head-nods* we counted the amount of yuck responses. Most of the *between-head-nods* (56 out of 90) get yucked by at least half of the participants. There were 3 *between-head-nods* which were yucked by each participant. There were 8 *between-head-nods* which were found appropriate by each participant, even though in the previous experiment none of the participants gave a response at that time.

Figure 4 represents the same 48 seconds from the previous studies. Now we added the yuck responses below the previous responses as negative responses. Note that only a head nod was generated and evaluated at response opportunities with at least four MultiLis or PCS responses. Moments 3 and 9 where the only generated *between-head-nods* in this segment. So, there were no head nods generated at *not*-appropriate times that nobody judged as inappropriate.

5 Discussion

In the previous studies we have collected positive responses (in the first two studies) and negative responses (in the last study). Combining these responses gives us three type of moments in our data. These types are *high agreement* (positive or negative), *controversial* and *neutral* moments. The *high agreement* moments have either positive *or* negative responses, the *controversial* moments have positive *and* negative responses and *neutral* moments have neither positive *nor* negative responses. In the following section we take a closer look at these type of moments. We do this by presenting several transcriptions of these moments and discussing the timing of the responses in relation to the context.

We first take a look at the response opportunities with *high agreement*; moments where most perspectives agree these are appropriate or inappropriate moments to provide a listener response. For this we take a look at the segment in Figure 4 and see what actually happens in the interaction. This segment is taken from an interaction where the speaker recites a recipe for risotto with mushrooms. In this segment the speaker is halfway through the ingredient list. The transcript is pre-

Table 1: Transcript of the segment displayed in Figure 4. The numbers in the rightmost column correspond to the response opportunities with the same number in Figure 4.

19.1 - 20.8	twee eetlepels	two tablespoons		
20.9 - 22.3	olie. Dus één liter	oil. So one liter		
22.5 - 23.3	twee en twee	two and two		
			24.1	**1**
24.3 - 25.2	olijfolie	olive oil		
25.3 - 25.8	natuurlijk	of course		
			25.9	**2**
27.9 - 28.7	euhm	euhm		
29.6 - 30.1	je hebt	you've got	29.9	**3**
30.3 - 32.9	verder voor de seasoning	furthermore for the seasoning		
33.4 - 34.5	één teentje knoflook	one clove of garlic		
35.6 - 36.4	één ui	one onion		
37.7 - 40.1	euh twee stengels bleekselderij	euh two sticks of celery		
			40.1	**4**
42.0 - 42.8	euh tijm	euh thyme		
42.9 - 43.9	één handjevol tijm	one handful of thyme		
			44.4	**5**
46.4 - 49.0	en natuurlijk euh heel veel paddestoelen	and of course a lot of mushrooms	49.0	**6**
49.1 - 50.1	500 gram	500 grams		
51.0 - 51.6	en	and	51.0	**7**
51.9 - 52.8	euhm	euhm		
53.2 - 54.4	natuurlijk de rijst	of course the rice		
55.2 - 57.0	400 gram rijst	400 grams rice		
			57.3	**8**
57.8 - 58.0	dus je hebt	so you've got		
58.4 - 61.9	euh 500 gram paddestoelen	euh 500 grams mushrooms	58.6	**9**
	400 gram rijst	400 grams rice		
62.4 - 65.3	en 100 gram parmezaanse kaas	and 100 grams parmesan cheese		
	dus in totaal	so in total		
65.4 - 65.7	mooi	nicely		
66.0 - 66.5	één kilo	one kilo	66.5	**10**

Table 2: Transcript of the most controversial response opportunity in the collected data, with 6 positive responses (3 MultiLis and 3 PCS) and 3 negative yuck responses.

29.0 - 31.1	het moment dat hij boven komt, euhm	the moment he arives at the top, euhm		
31.6 - 32.1	oh wacht	oh wait		
32.3 - 32.9	helemaal verkeerd	that's wrong		
			33.3	**11**

Table 3: Transcript of a neutral response opportunity where no positive and no negative responses are recorded.

30.5 - 34.1	euh, volgende list moet ie verzinnen	euh, he has to come up with a new trick		
	hij gaat vanaf	he goes from		
34.6 - 35.4	euh	euh		
35.5 - 36.1	een tegenoverliggend gebouw	an opposing building	36.1	**12**
36.1 - 40.8	via allemaal lijnen	across all those cables		
	die daar gespannen zijn	that are spanned there		

sented in Table 1. The numbers in the rightmost column correspond to the response opportunities with the same number in Figure 4. The *high agreement* moments in this segment are 1, 2, 5, 8 and 10 (positive), and 3 and 9 (negative).

The response opportunities 1 and 10 both are in reaction to a summarizing statement. Both statements summarize the previous ingredients with a mnemonic device to help them memorize the ingredients by summarizing the numbers mentioned (1) or by adding up the weights to a round figure (10). Beside the verbal cues, the speaker also makes iconic gestures to accompany the summarizing statements.

The other three *high agreement* response opportunities in this segment (2, 5 and 8) are all in reaction to a refining statement in which a previously mentioned ingredient is more precisely described: the oil is specified as being olive oil (2), the amount of the thyme is specified (5) and the precise weight of the rice (8). The other ingredients (like the garlic and onion) are also acknowledged with a listener response by some, but agreement between individuals is much lower in these cases (see the unnumbered response opportunities in Figure 4).

The moments with *high agreement* in negative yuck responses (3 and 9) are both mid sentence. They are not placed near or after the end of a grammatical clause, which is identified as a cue by Dittman and Llewellyn (Dittmann and Llewellyn, 1968), but instead are placed during or directly after the theme of the sentence. So, no new information has been mentioned by the speaker yet (rheme) and the listener response is premature. Furthermore, moments with *high agreement* in negative yuck responses are moments after long silences of at least 2 seconds, moments in between the article and the noun, and moments shortly (within 1.5 seconds) following another listener response.

An interesting case are the moments 6 and 7. The listeners in the corpus respond to "mushrooms", while the PCS responses are in reaction to the refining statement "500 gram". According to a previous study PCS responses are on average 220 ms slower (de Kok and Heylen, 2011a). Since the pause between the two statements is very short (a little over 100 ms), this delay would cause the PCS-er to place the PCS response during the "500 gram" statement. Instead they wait until the refining statement is finished. However, the faster responses from the listeners do not interfere with this statement and are made before the refining statement is started. Response opportunity 7 is a *controversial* moment since it is also yucked by two individuals. This is probably due to the timing, which is synchronous to the start of the word "and".

Besides response opportunities 4 and 7 there are other *controversial* response opportunities in the corpus. The most controversial moment has 6 positive responses (3 MultiLis and 3 PCS) and 3 negative yuck responses. The transcript of this moment is presented in Table 2. In this segment the speaker corrects himself. An acknowledgment from the listener through a listener response is valid according to six perspectives. The recorded listeners all responded to this moment, however two of them did not respond with a head nod, but with a polite smile (the speaker also smiles at this moment). However, the generated virtual agent in study 3 only performs a head nod. So it is likely that the response opportunity is not yucked because of the timing, but because of the type of listener response displayed.

Another reason for controversy in the corpus is that two response opportunities in quick succession (within 2 seconds) are individually regarded as good response opportunities (at least 4 positive response to each opportunity in the first two studies), but when generating a listener response at both moments in the third study, the second listener response gets yucked by some.

The last category of responses are the *neutral* responses. These are responses which are generated as *between-head-nods* in Study 3 at moments they received no positive responses in the first two studies. However, in the third study they were not seen as inappropriate responses and thus not yucked. In Table 3 one of these moments is transcribed. The head nod is placed mid sentence, not during a pause. The complete statement is not yet finished. However, it is placed directly after a vital piece of information within this statement ("an opposing building"), which is emphasized by the speaker and memorized after a short hesitation. A confirmation of this piece of information is appropriate according to Study 3 even though no other perspectives previously provided a response there. There are 7 *neutral* moments in our data (see Figure 5). In 5 of these moments the listener response

is placed mid sentence after a vital piece of information as in the previous example. In the other two cases the listener response is placed between sentences.

6 Conclusion

In this paper we have illustrated individual differences in nonverbal listening behavior. The combination of the multiple perspectives collected in these studies has provided us with a rich data set in which different types of response opportunities are distinguishable. There are moments where there is *high agreement* between these multiple perspectives that a listener response is appropriate or inappropriate, moments where a listener response is *controversial* and moments neither a response was given nor a response was judged inappropriate (*neutral*).

Analysis of the context of the different type of response opportunities has shown different contextual characteristics that should help discriminating these response opportunities. Observations have shown relations to sentence structure (listener responses before (part of) the rheme is completed are considered inappropriate), conversational structure (listener responses in reaction to summarizing or refining statement are more appropriate) and proximity of earlier responses (producing two similar listener responses in close succession is considered inappropriate).

So far these characteristics are not used in state-of-the-art predictive models for the timing of listener responses (Morency et al., 2010; de Kok et al., 2010; Huang et al., 2010a). We feel that, in order to push these predictive models beyond the state-of-the-art, these characteristics should be taken into account. An obstacle towards the use of these characteristics, is the absence of real-time recognition systems of these characteristics on output of speech recognition software, such as theme and rheme discrimination within sentence and classification of statements and their relation to earlier statements.

References

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

Janet Beavin Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79(6):941–952.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Iwan de Kok and Dirk Heylen. 2011a. Appropriate and Inappropriate Timing of Listener Responses from Multiple Perspectives. In *Intelligent Virtual Agents*, pages 248–254. Springer.

Iwan de Kok and Dirk Heylen. 2011b. The MultiLis Corpus - Dealing with Individual Differences of Nonverbal Listening Behavior. In Anna Esposito, Antonietta Esposito, Raffaele Martone, Vincent C. Müller, and Gaetano Scarpetta, editors, *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces: Theoretical and Practical Issues*, pages 374–387. Springer Verlag.

Iwan de Kok, Derya Ozkan, Dirk Heylen, and Louis-Philippe Morency. 2010. Learning and Evaluating Response Prediction Models using Parallel Listener Consensus. In *Proceeding ICMI-MLMI '10 International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*. ACM Press.

Allen T. Dittmann and Lynn G. Llewellyn. 1968. Relationship between vocalizations and head nods as listener responses. *Journal of personality and social psychology*, 9(1):79–84.

Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. 2010a. Learning Backchannel Prediction Model from Parasocial Consensus Sampling : A Subjective Evaluation. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 159–172.

Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. 2010b. Parasocial Consensus Sampling: Combining Multiple Perspectives to Learn Virtual Human Behavior. In *Proceedings of Autonomous Agents and Multi-Agent Systems*, Toronto, Canada.

Robert E. Kraut, Steven H. Lewis, and Lawrence W. Swezey. 1982. Listener responsiveness and the coordination of conversation. *Journal of Personality and Social Psychology*, 43(4):718–731.

Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. 2010. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, 20(1):70–84.

Herwin van Welbergen, Dennis Reidsma, Zsófia M. Ruttkay, and Job Zwiers. 2010. Elckerlyc - A BML Realizer for continuous, multimodal interaction with a Virtual Human. *Journal on Multimodal User Interfaces*, 3(4):271–284.

Victor H. Yngve. 1970. On getting a word in edgewise. In *sixth regional meeting of the Chicago Linguistic Society*, volume 6, pages 657–677.

Speaker Clustering in Multi-party Conversation

Masafumi Nishida, Yuki Ishikawa, Seiichi Yamamoto

Graduate School of Engineering, Doshisha University, Kyoto 610-0321, Japan

{mnishida,seyamamo}@mail.doshisha.ac.jp, dtl0720@mail4.doshisha.ac.jp

Abstract

Speech feature variations are mainly attributed to variations in phonetic and speaker information included in speech data. If these two types of information are separated from each other, more robust speaker clustering can be achieved. We propose a speaker clustering method using principal component analysis transformation by separating speaker information from phonetic information, under the assumption that a space with large within-speaker variance is a "phonetic subspace" and a space with small within-speaker variance is a "speaker subspace". We carried out comparative experiments of the proposed method with conventional methods based on Bayesian information criterion and Gaussian mixture model in an observation space. The experimented results showed that the proposed method can achieve higher clustering accuracy than conventional methods.

1 Introduction

In automatic interaction management, it is important to improve interactions by making interaction smooth and natural, and be able to elicit and to provide communicative signals that allow the user to take the turn. Recently there has been growing interest in the automatic analysis of conversational data so as to further our understanding of human-human communication and multimodal signaling of social interactions. Due to advance technology, it is possible to study communicative behavior and social signaling patterns using automatic analysis techniques. Besides speech and speaker recognition, also motion capture and gesture recognition technology can be used, while the development in eye-tracker technology allows us to study gaze behaviour in an objective manner.

Chen et al. (2009) investigated combining verbal with nonverbal cues (i.e., hand gesture and eye gaze) to detect floor control shifts in multi-party meetings. Jokinen et al. (2010) showed that eye-gaze is an important cue in deciding turn-taking: the use of eye-gaze information improves classification accuracy of turn-taking significantly, compared with the use of only speech features or dialogue acts. Battersby (2011) studied interactions with a motion tracker device, and points out that the speaker's gesturing behavior differs from that of the addressees, and that head and hand movements are also different between primary and secondary addressees.

In this paper, we focus on speaker clustering based on speaker recognition technique in multi-party conversations. Speaker clustering is a technique for clustering utterances from the same speaker, and is useful for retrieving the utterances of a specific speaker and for improving automatic speech recognition performance based on speaker adaptation of the acoustic model. Speaker clustering has been studied mainly for broadcast news audio, multi-party conversations, and telephone conversations (Tranter and Reynolds, 2006) (Reynolds and Torres-Carrasquillo, 2005).

In previous studies, Chen et al. (1998) presented a maximum likelihood approach for acoustic change detection; the detection of a turn is based on the Bayesian information criterion (BIC), a model selection criterion well-known in statistics. Furthermore, Cheng et al. (2010) proposed three divide-and-conquer approaches for BIC-based speaker segmentation. The three approaches are used to detect speaker changes by recursively partitioning a large analysis window into two sub-windows and recursively verifying the merging of two adjacent audio segments using ΔBIC, a widely adopted distance measure of two audio segments. Iso (2010) proposed a method for representing a speech segment with a vector of Vector quantization (VQ) code frequencies by using a cosine between two vectors as their similarity measure. The clustering is done using a spectral clustering algorithm with cluster number estimation based on an eigen structure of the similarity matrix. Nishida et al. (2005) proposed a flexible framework in which an optimal speaker model (GMM or VQ) is automatically selected based on the BIC and according to the amount of training data available. Reynolds et al. (1998) presented the cross likelihood ratio (CLR), and Le et al. (2007) presented the normalized cross likelihood ratio (NCLR) and the advantages of using it in a speaker diarization system.

For speaker identification and verification, Nishida et al. (2001) proposed a method based on a statistical speaker model (GMM) in the "speaker subspace" which is created using all speech data projected to the speaker subspace where the phonetic information is suppressed. The speech data include two types of information, phonetic and speaker. Phonetic information is attributed to the phonetic features in speech data, and speaker information is attributed to the speaker features in speech data. In particular, phonetic information varies depending on the speech data. Therefore, if these two types of information are separated from each other, robust speaker recognition can be achieved.

Conventional speaker-clustering methods do not distinguish between phonetic and speaker information. We propose a speaker clustering method based on a statistical speaker model (GMM) in the "speaker subspace", which is created using all speech data projected to the speaker subspace where the phonetic information is already suppressed. In speaker clustering, we believe that our method is effective in separating speaker from phonetic information because the variance in duration of each segment enlarges variation of phonetic information in the segment more in comparison with speaker identification and verification. We carried out speaker clustering experiments with three methods. The first method was a hierarchical agglomerative clustering method based on the BIC in an observation space. The second method was a hierarchical clustering method based on CLR using GMM in an observation space. The third method is the proposed method based on GMM in the speaker subspace obtained from an observation space. Our proposed method clusters using the CLR.

The remainder of this paper is organized as follows: Section 2 explains speaker clustering based on GMM in speaker subspace, Section 3 describes our speaker clustering experiments and section 4 concludes the paper.

2 Speaker Clustering based on GMM in Speaker Subspace

2.1 Separation of phonetic and speaker subspaces

We describe a separation method of phonetic and speaker information. The speech feature variation is mainly caused by the variation in the phonetic information included in speech data. This insight enables the separation of the phonetic and speaker information based on this variance. Principal component analysis (PCA) is conducted to locate each speaker's speech data of phonetic information in a subspace constructed using the principal component axes (lower order axes), and speaker information in a complementary subspace constructed using the higher order axes. We call the subspace with the large variation constructed using the lower axes "phonetic subspace", and the subspace with the small variation constructed using the higher axes "speaker subspace".

A sequence of speech data $\left\{x_t^{(s)}\right\}\left(t = 1,2,\dots,N^{(s)}\right)$ of a segment s is observed in an n-dimensional observation space. Its mean vector $\bar{x}^{(s)}$ and covariance matrix $R^{(s)}$ are then computed from the training data as follows:

$$\bar{x}^{(s)} = \frac{1}{N}\sum_{t=1}^{N} x_t^{(s)} \tag{1}$$

$$R^{(s)} = \frac{1}{N}\sum_{t=1}^{N}(x_t^{(s)} - \bar{x}^{(s)})(x_t^{(s)} - \bar{x}^{(s)})^T \tag{2}$$

The covariance matrix R can be composed of eigenvectors and a matrix of eigenvalues as follows:

$$R^{(s)} = \Phi^{(s)}\Lambda^{(s)}\Phi^{(s)T}, \tag{3}$$

where $\Lambda^{(s)}$ is a diagonal matrix whose diagonal components are eigenvalues $\lambda_i^{(s)}$ $(i = 1,\dots,k,\dots n)$ of $R^{(s)}$, and $\Phi^{(s)}$ is a matrix whose columns are eigenvectors $\varphi_i^{(s)}$ $(i = 1,\dots,k,\dots n)$ of $R^{(s)}$.

The eigenvalues $\lambda_i^{(s)}$, which are obtained by eigenvalue decomposition, represent a variance in the eigenvectors $\varphi_i^{(s)}$, which are orthonormal bases. In this study, a space constructed by eigenvectors corresponding to the largest eigenvalues up to k numbers is the phonetic subspace, which represents the phonetic information. A space constructed by $(n-k)$ eigenvectors corresponding to the remaining small $(n-k)$ eigenvalues is the speaker subspace, which is complementary to the phonetic subspace. The speaker subspace represents the speaker information. Consequently, the input speech can be separated into phonetic and speaker information by projecting both type of information to the speaker and phonetic subspaces, respectively.

2.2 Speaker clustering based on projection to speaker subspace

Clustering ideally produces one cluster for each speaker in a conversation and assigns all segments from each speaker to a single cluster. Gaussian mixture models are trained using the speech data projected to the speaker subspace for each segment.

The Mel-frequency cepstral coefficient (MFCC) is commonly used in speaker recognition and is obtained from the log filter-bank amplitudes using a discrete cosine transform (DCT). However DCT is not designed to transform a space by taking into account data distribution as well as correlation of feature parameters. In this study, we used PCA instead of DCT to diagonalize a data covariance matrix and decorrelate the feature parameters of the log filter-bank amplitudes. This PCA, which we used instead of DCT for signal processing, can also construct respective speaker subspace.

A sequence of speech data $\left\{x_t^{(s)}\right\}$ of a segment s observed in an n-dimensional observation space is projected to the speaker space by using Eq. (4) and the speaker model (GMM) is trained in the speaker subspace by using the projected speech data.

$$\hat{x}_t^{(s)} = P^{(s)T}(x_t^{(s)} - \bar{x}^{(s)}) \tag{4}$$

The orthogonal matrix $P^{(s)}$ has columns that are higher order eigenvectors $\varphi_i^{(s)}(i = k, \ldots, n)$, which were obtained with PCA for the segment. Figure 1 shows an example of the projection to the speaker subspace. The speaker subspaces of segments *A* and *B*, shown with rectangles, are respectively denoted by P_A and P_B. The regions enclosed by ellipses indicate the speech data. The speaker subspace is a space constructed by axes whose variance is small. Therefore, after projecting the speech data of segments *A* and *B* to each speaker subspace, a within-speaker variance becomes smaller than that in an observation space, leaving a fixed between-speaker variance.

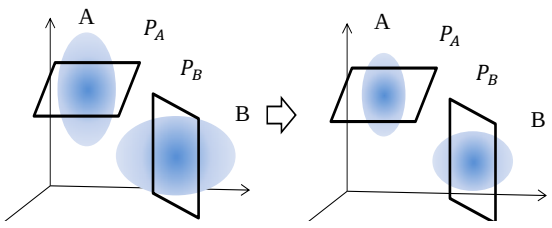

Figure 1: Projection to speaker space

Figure 2 shows a conceptual diagram of the projected phonetic subspace. The orthonormal basis vector $\varphi_1^{(s)}$ configures the phonetic subspace, and the orthonormal basis vectors $\varphi_2^{(s)}$ and $\varphi_3^{(s)}$ configure the speaker subspace. The input feature vector x_t can be divided into phonetic vector $x_{phoneme}^{(s)}$ and speaker vector $x_{\text{speaker}}^{(s)}$ by using Eqs. (5) and (6), respectively. $x_{phoneme}^{(s)}$ shows the phonetic vector projected to the phonetic subspace, and $x_{\text{speaker}}^{(s)}$ shows the speaker vector projected to the speaker subspace.

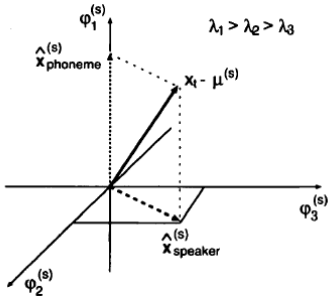

Figure 2: Phonetic vector and speaker vector

$$x_{phoneme}^{(s)} = \sum_{i=1}^{k} (x_t - \bar{x}^{(s)}, \quad \varphi_i^{(s)})\varphi_i^{(s)} \tag{5}$$

$$x_{\text{speaker}}^{(s)} = \sum_{i=k+1}^{n} (x_t - \bar{x}^{(s)}, \quad \varphi_i^{(s)})\varphi_i^{(s)} \tag{6}$$

A common approach used in speaker clustering is hierarchical agglomerative clustering with a CLR consisting of the following steps:

1. Form one cluster from each segment.
2. Construct a speaker subspace in the segment by performing PCA.
3. Project speech data in the segment to the speaker subspace by using Eq. (4).
4. Construct a statistical speaker model (GMM) in the respective speaker subspaces.
5. Compute the CLR as pair-wise distances between each cluster (Reynolds et al. ,1998). The CLR d_{ij} for clusters i and j is given by Eq. (7).

$$d_{ij} = \log\frac{P(X_i|\lambda_i)}{P(X_i|\lambda_j)} + \log\frac{P(X_j|\lambda_j)}{P(X_j|\lambda_i)} \tag{7}$$

$$\log P(X_i|\lambda_j) = \frac{1}{n_i}\sum_{k=1}^{n_i} \log P(x_{ik}|\lambda_j)$$

where X_i is a segment of cluster i, x_{ik} is its kth frame feature of the segment, n_i is the number of frames of a segments, λ_i is the parameters of GMM for cluster i, and $\log P(X_i|\lambda_j)$ is the average log likelihood of the segment of cluster i given by model λ_i.

6. Merge the closest pairs of clusters, if the minimum distance between the clusters is smaller than the threshold θ.

7. Update distances of remaining clusters to form a new cluster by using the unweighted pair-group method using arithmetic averages (UPGMA) (Sneath and Sokal, 1973) by Eq. (8).

$$d(r,s) = \frac{1}{n_r n_s}\sum_{i=1}^{n_r}\sum_{j=1}^{n_s} dist(x_{ri}, x_{sj}), \quad (8)$$

where r and s are the cluster number, n_r and n_s indicate the number of segments in each cluster, and $dist(x_{ri}, x_{sj})$ is obtained by Eq. (7).

8. Iterate steps 5-7. The clustering process finishes if all distances between clusters are not smaller than the threshold θ.

3 Experiments

3.1 Experimental Setup

We used corpus of spontaneous Japanese (CSJ) as evaluation data. The CSJ consists of 3302 talks (662 hours, 1417 speakers) collected from academic conference presentations and extemporaneous speeches (Maekawa, 2003). The talks are segmented into utterances at every pause of longer than 300 milliseconds. We chose utterances of multiple speakers randomly from the CSJ to make the test sets as close to actual multi-party conversations as possible. We used five test sets (1-5), each of which consisted of five speakers. The duration of an utterance ranged from 30 to 70 seconds. In addition, we also used another five test sets (6-10), each of which consisted of 10 speakers. The duration of an utterance ranged from 20 to 50 seconds. The duration of one speaker's total speech was about 100 seconds. There are not overlapping utterances in the test tests. Table 1 lists the detail of each test set.

The speech data was sampled at 16 kHz, analyzed with an analysis window size of 25 ms with 10-ms overlap, and parameterized into 24 cepstral coefficients obtained using a 24-channel Mel-frequency spaced filter-bank.

Table 1: Details of each test set

Test set No.	Number of speakers	Number of segments	Total segments time (min)
1	5	55	44.5
2	5	57	45.1
3	5	59	44.4
4	5	58	44.8
5	5	55	45.0
6	10	177	95.0
7	10	181	93.7
8	10	183	93.5
9	10	174	81.4
10	10	171	91.5

We carried out speaker clustering experiments with three methods: The first method was a hierarchical agglomerative clustering method based on BIC in an observation space with 24 dimensional MFCC parameters. The second method was a hierarchical clustering method based on the CLR using GMM in an observation space with 24 dimensional MFCC parameters. The third method was the proposed method based on GMM in the speaker subspace obtained from an observation space with 24 channel log filter-bank amplitudes. Our method clustered using CLR.

The clustering results were aligned with the ground truth speaker labels to measure their accuracy based on the diarization error rate (DER) (Iso, 2010):

$$\text{DER} = \frac{T_{miss} + T_{wrong}}{T_{ref}}, \quad (9)$$

where T_{miss} is the total length of segments not aligned with the speaker labels, T_{wrong} is the total length of segments aligned with the wrong speaker labels, and T_{ref} is the total length of all segments in a test set. We also calculated the purity metric (Iso, 2010):

$$\text{Purity} = \frac{T_{pure}}{T_{ref}}, \quad (10)$$

where T_{pure} is the total length of the speaker label, which is the longest utterances for each cluster.

3.2 Experimental results

Table 2 lists the clustering results for test sets 1-5, and Table 3 lists the clustering results for test sets 6-10. The parameter α for the BIC is the turning parameter, MN indicates the number of mixtures of the GMM, and SD for the proposed method indi-

cates the dimensions of the speaker subspace. To investigate the phoneme -dependency of each eigenvector axis, we compared 20 combinations of dimensions with 1-20th, 1- 21st, 1-22nd, 1-23rd, 1-24th, 2-20th, 2-21st, …, and 4-24th eigenvectors.

Table 2: Clustering results for the test sets 1-5

	DER(%)	Purity(%)	Parameter
BIC	8.8	90.5	$\alpha = 1.5$
GMM	10.1	89.4	MN = 2
Proposed method	6.8	92.2	MN = 4 SD = 2–21

Table 3: Clustering results for the test sets 6-10

	DER(%)	Purity(%)	Parameter
BIC	10.8	87.9	$\alpha = 1.2$
GMM	12.8	86.4	MN = 4
Proposed method	7.1	92.2	MN = 4 SD = 2–21

Tables 2 and 3 show that the proposed method obtained a higher clustering accuracy than that obtained with the conventional methods based on the BIC and GMM, for both groups of test sets. Test sets 5-10 contained five speakers and test sets 6-10 contained 10 speakers. Therefore, the proposed method can obtain high clustering accuracy with a variation in the number of speakers.

Figures 3 and 4 show the relation between clustering accuracy and the number of mixtures for the conventional GMM and the proposed method for test sets 1-5 (Fig. 3) and 6-10 (Fig.4). The optimal number of mixtures of the GMM varies because GMM of two mixtures is best for test sets 1-5 and GMM of four mixtures is best for test sets 6-10. However, the optimal number of mixtures of the proposed method does not depend on the number of speakers.

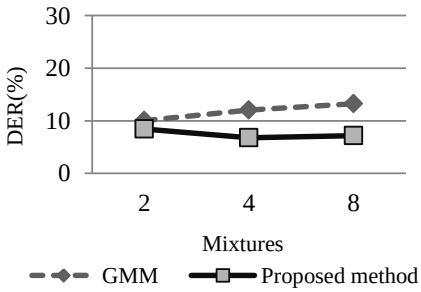

Figure 3: DER in each mixture for test sets 1-5

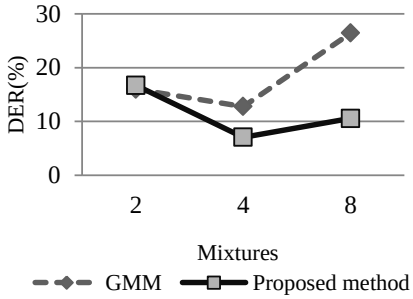

Figure 4: DER in each mixture for test sets 6-10

A preliminary experiment, showed that the first axis of PCA should not be used for configuring the low-dimensional axes of the speaker subspace in the proposed method. Therefore, Fig. 5 shows the DER when the higher-dimensional axes of the speaker subspace are reduced. The number of mixtures is four for all cases.

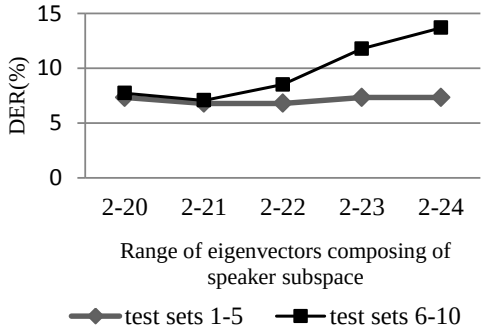

Figure 5: DER in various ranges of eigenvectors composing the speaker subspace

As clearly shown in the Fig. 5, the best DER was obtained when SD was 2-21 for both test sets. However, in each test set, the best DER varied by the dimensions of the speaker subspace because the variation in utterance lengths was large. Therefore, we will study how to select the optimal dimensions of the speaker subspace by considering the variability of phoneme in speech data.

The average number of clusters with the BIC was 5.0, the GMM was 5.8, and the proposed method was 5.6 for test sets 1-5. The standard deviation was 0.71 for the BIC, 1.30 for the GMM, and 0.89 for the proposed method. For test sets 6-10, the average number of clusters was 11.6, 13.8, and 14.0, for the BIC, GMM and proposed method, respectively. The standard deviation by the BIC was 0.55, the GMM was 2.56 and the proposed method was 1.22. The proposed method used a threshold for the CLR to stop the clustering process. For future work, we will use BIC as a stopping criterion of cluster-

ing for the proposed method to improve the estimation accuracy of the number of speakers.

4 Conclusions

We proposed a speaker-clustering method using a GMM trained in speaker subspace using speech data projected to the speaker subspace. The proposed method used PCA transform to construct the speaker subspace.

From the results of the speaker clustering experiments, the DER with the BIC was 8.8% for test sets 1-5 and 10.8% for test sets 6-10, that with the CLR using a GMM was 10.1% for test sets 1-5 and 12.8% for test sets 6-10, and that with the proposed method was 6.8% for test sets 1-5 and 7.1% for test sets 6-10. Therefore, the proposed method obtained a higher speaker clustering accuracy than that with the conventional methods. The experiments also demonstrated that separating the phonetic and speaker subspaces using PCA was effective.

For future work, we will evaluate the proposed method on the National Institute of Standards and Technology (NIST) databases to demonstrate its generality. It is also necessary to study how to select the optimal number of dimensions of the speaker subspace. Moreover, we will study on speaker clustering for test data included overlapping utterances.

References

Lei Chen and Mary P. Harper. 2009. *Multimodal Floor Control Shift Detection*, Proc. ICMI-MLMI.

Kristiina Jokinen, Kazuaki Harada, Masafumi Nishida, and Seiichi Yamamoto. 2010. *Turn-alignment Using Eye-gaze and Speech in Conversational Interaction*, Proc. Interspeech, pp.2018-2021.

Stuart Battersby. 2011. *Moving Together: the Organization of Non-verbal Cues During Multiparty Conversation*, PhD Thesis.

Sue E. Tranter and Douglas A. Reynolds. 2006. *An Overview of Automatic Speaker Diarization Systems*, IEEE Transactions on Audio, Speech, and Language Processing, Vol.14, No.5, pp.1557-1565.

Douglas A. Reynolds and Pedro A. Torres-Carrasquillo. 2005. *Approaches and Applications of Audio Diarization*, Proc. ICASSP, Vol.5. pp.953-956.

Scott Chen and Ponani Gopalakrishnan. 1998. *Speaker Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion*, Proc. DARPA Broadcast News Transcription and Understanding Workshop, pp.127-132.

Shih-Sian Cheng, Hsin-Min Wang, Hsin-Chia Fu. 2010. *BIC-based Speaker Segmentation Using Divide-and-conquer Strategies with Application to Speaker Diarization*, IEEE Transactions, Vol.18, pp.141-157.

Kenichi Iso. 2010. *Speaker Clustering Using Vector Quantization and Spectral Clustering*, Proc. ICASSP, pp. 4986 – 4989.

Masafumi Nishida and Tatsuya Kawahara. 2005. *Speaker Model Selection Based on the Bayesian Information Criterion Applied to Unsupervised Speaker Indexing*, IEEE Transactions on Speech and Audio Processing, Vol.13, No.4, pp. 583-592.

Douglas A. Reynolds, Elliot Singer, Beth A. Carlson, Gerald C. O'Leary, Jack J. McLaughlin, and Marc A. Zissman. 1998. *Blind Clustering of Speech Utterances based on Speaker and Language Characteristics*, Proc. ICSLP, pp.3193-3196.

Viet-Bac Le, Odile Mella, and Dominique Fohr. 2007. *Speaker Diarization using Normalized Cross Likelihood Ratio*, Proc.Interspeech, pp.1869-1872.

Masafumi Nishida and Yasuo Ariki. 2001. *Speaker Recognition by Separating Phonetic Space and Speaker Space*, Proc. EUROSPEECH, Vol. 2, pp. 1381-1384.

Peter Sneath and Robert R. Sokal. 1973. *Numerical Taxonomy*, W. H. Freeman and Company.

Kikuo Maekawa. 2003. *Corpus of Spontaneous Japanese: Its Design and Evaluation*, Proc. ISCA & IEEE Workshop on SSPR, pp.7-12.

Close your eyes…and communicate

Laura Vincze
Roma Tre University
Rome, Italy
laura.vincze@gmail.com

Isabella Poggi
Roma Tre University
Rome, Italy
poggi@uniroma3.it

Abstract

In this work we present a typology of eye closings and their possible meanings based on a taxonomy of communicative signals. The two types of eye closing we investigate here are blinks and eye-closure. Our aim is to prove that these social signals may be communicative and bear subtle but important meanings.

1. Introduction

Facial communication is a widely studied field, where on the one side, research is carried out on single parts of the face, like eyes or mouth, and on the other side, on face as a whole. This paper focuses on a single part of the face, eyes, specifically on two types of eye closing, *blinks* and *eye closure*, trying to interpret the possible meanings of these two signals. Numerous studies have been devoted to gaze. Gaze has been studied in many of its social and communicative functions (Kendon and Cook 1969; Argyle and Cook 1976), mainly in connection with greeting and flirting behaviour (Kendon 1973), conversational manoeuvres like turn-taking (Duncan 1974, Goodwin 1991) and backchannel (Heylen 2005, Maatman et al. 2005). Eyebrows also received attention from scholars (Ekman 1979, Eibesfeldt 1972, Pelachaud and Prevost 1994, Costa and Ricci Bitti 2003) who studied eyebrows behaviour as a signal fulfilling social and emotional but also syntactic and conversational functions. Researchers' interest was attracted also by blinks. Blinks' occurrences have been studied during cognitive tasks such as reading, memorizing and lying (Zuckerman et al. 1981; De Paulo and Kirkendol 2003; Leal and Vrij 2008). As far as we know, there have been no attempts to investigate the *meanings* borne by blinks and eye closure.

2. Gaze semantics

This paper is meant to contribute to the detailing and specifying of the lexicon of gaze (Poggi (2007). According to Poggi (2007), it is possible to single out a list of signal/meaning pairs for the features and movements of the eye region (eyebrows, eyelids, eyes, eye sockets). Moreover, according to how these features are combined, changes occur in meaning (much like with morphemes of verbal languages). Specific gaze behaviours were analyzed in detail, like eyebrow frown and eyebrow raising (Poggi 2007) and eyelids positions (Poggi et al. 2010 a). These studies have proved the semantic richness of gaze, by stressing that eyes convey much more than simply turn-taking and backchannel, emotions and some basic information like the topic/comment distinction, and that not only gaze direction should be studied, but also many other features of eyes and their behaviour.

3. Closing eyes. An observational study on blink and eye-closure during debates

In this paper we investigate the gaze behaviours of *closing the eyes*. As for any analysis of body (potentially) communicative behaviour, we must first distinguish between the signal (the set of physical features of the eyelids, their muscular actions and their physiological state in closing the eyes) and its goal.

On the signal side, we distinguish two types of eye closing: *blink* versus *eye-closure*. Both signals share a common feature, complete eye closing of both eyes, that distinguishes them from the *wink*, a unilateral eyeclosing usually conveying complicity or furtive agreement (Vincze and Poggi forthcoming). But they

differ in one major feature: the *duration* of the closing. By *blink* we mean, following Ekman and Friesen (2002), a quick closing of the eyes and return to eyes open, while by *eye-closure* we refer to a longer eye closing than in a blink, sometimes further characterized by a higher tension in the eyelids.

As to the goal of these signals, often eye features and behaviours do not have a communicative goal, so we distinguish *non-communicative* cases, that only have biological goals (like soaking the eye), and *meaningful* cases, in which either the Sender of that eye feature or behaviour had the goal of communicating some meaning *(communicative cases)*, or simply a potential Receiver can acquire information *(informative cases)*.

Within *non-communicative* blinks (at least from what results from our observation, see below) we count at least two cases: 1. the "physiological" blink, that merely fulfils the physiological need of keeping a standard level of eye humidity, and 2. the blink of a stuttering person: a person having problems in pronouncing a word may blink when engaging in the production of that word, while repeating its first syllable. From our observation it results that a *non-communicative* blink is generally rapid and single (not repeated), while a communicative blink is in general constituted by a series of rapid blinks. Repetition is not a sufficient condition to interpret blinks as meaningful, since due to idiosyncratic differences some people tend to blink more frequently; but in general repetition is necessary to consider a blink as communicative.

Also the *eye-closure* can be either *meaningful* (*communicative or informative)* or *non-communicative*. Typical non-communicative instances of eye-closure are while sleeping. But apart from this case, unlike blinks, which in their vast majority are physiological and non-communicative, all cases of eye-closure performed while speaking or listening may be, or definitely are, communicative.

4. Eye closing as a communicative behaviour

When blinks and eye-closures are communicative, we can analyze them on both the signal and the meaning side.

To describe the signal, we refer to some of Hartmann's et al. (2002) expressivity parameters: eyelid tension, velocity, duration and repetition. These parameters help us distinguish between a non communicative blink and a communicative one, and between a communicative blink and a communicative eye-closure.

a. Communicative vs. non-communicative blink. Here the relevant parameter is *repetition*: as mentioned, a physiological blink is generally single, while the communicative one is generally rapid and repeated.

b. Communicative vs. non-communicative eye-closure. To distinguish a communicative eye-closure from a non-communicative one, *duration* may be significant, but also the context in which the eye-closure appears is relevant: in a debate it is much less likely (if not impossible) for a non-communicative eye-closure (sleep) to appear, while in a relaxed, familiar situation this may sometimes occur.

c. Communicative blink vs. communicative eye-closure. We can distinguish an item of blink from one of eye-closure mainly based on the parameter of *duration*, but also *repetition* and *tension* can be pertinent.

A communicative blink and a communicative eye-closure generally differ in that a communicative blink is repeated, brief, very rapid, and therefore not tense (there is no pressure by the eyelids), while a communicative eye-closure is single, longer, with eyelids going down slowly and the upper eyelid often pressed against the lower one. During emphatic eye-closure (see Sect. 7.4), the eyebrows may be raised as well, therefore causing a tightening of the upper eyelid.

Tension is connected to *duration*. A blink is so fast that it cannot involve tension. If one has the time to press the upper eyelid against the lower one, it is not a blink anymore, but an eye-closure. So whatever closing of the eyes is long and tense, is an eye-closure.

5. Corpus and method

Our corpus is composed of six political debates of roughly 40 minutes each from the Canal 9 Corpus (available on the SSPNet website sspnet.eu).

To distinguish between communicative and non communicative eye behaviour, we first viewed the six debates. When an eye closing occurred, we focused on the concomitant

verbal message delivered by the person performing the eye closing or, when the sender of the eye behaviour was the listener, the verbal message produced by the present speaker. Based on the signal and the parallel verbal message, we attributed a possible meaning to each eye behaviour.

6. Analysis of a gaze item

To analyze eye behaviour we built the annotation scheme of Table 1, based on the principles of Poggi (2007). Column 1 contains the time in the video; columns 2 and 3 contain a description, respectively, of the verbal and body behaviour; col. 4, the goal or meaning of the behaviours in columns 2 and / or 3. For the verbal behaviour described in col. 2, its goal is by definition a communicative goal, while for the action written in col. 3 the goal to be written in col. 4 may be either a communicative goal (for the communicative blinks and eye-closures) or not (for those behaviours in which the Agent does not intend to have the other Agent know something). The goal in col. 4 and col. 5 is phrased as a sentence in the first person. Column 5 is there because a communicative action, besides its direct goal, may aim at one or more supergoals, i.e. some information to be inferred by the Addressee; so in col. 5 we write the possible supergoal of the actions in col.3. Finally, in col. 6 we classify the goal of col. 4 (or the supergoal of col. 5, when there is one) in terms of the taxonomy of meanings illustrated in the following sections.

Table 1. shows the analysis of one item of communicative eye-closure and one of non-communicative blink. In the first instance the sender of the signal is the listener, Mr. Freysinger, who performs an eye-closure during the moderator's turn. Through his head shake he communicates that the answer to the moderator's question is 'No', while the rest of his body behaviour, eye-closure accompanied by raised eyebrows, communicates that not only it is not so, but whoever believes such a thing is a fool.

The second item analyzed in Table 1. is a case of non-communicative blink. The Speaker Mr. Gabul has difficulty in pronouncing the polysyllabic word 'municipalité' and stutters while pronouncing its first syllable (*"Mu-municipalité"*). The blink, performed while pronouncing the first syllable, accompanies the effort of uttering the syllable and is not communicative, as the Speaker has no intention to communicate to the listeners that he is striving to correctly pronounce the word.

7. Types of eye closing

Based on our analysis of the above corpus of debates, and in some cases on everyday life observation, four main categories of eye closing can be singled out, grouped on the basis of their meaning (or their non-meaningful goal) and not of the signal.

7.1. Non-communicative eye closing behaviours.

Non communicative blinks

a) The most common type of blink in our corpus is the non-communicative physiological blink: a rapid eye-closing aimed at soaking the eyes.

b) Another type of non-communicative blink is the above-mentioned blink of a stuttering person.

c) A third type are blinks performed during startle reactions. According to Ekman and Friesen (2007), startle is a reflex, quite similar to the emotion of surprise, but differing from it for both expressive behaviour and underlying emotional state. Generally, in the startle reflex rapid repeated blinks are produced, the head may go backwards and there is a "leap up" of the entire body. In surprise, instead, depending on its intensity, we may raise eyebrows, open eyes widely and even perform a jaw drop, but not necessarily blinks, though startle blinks may come as the most intense reaction of surprise.

While, as we will see later (ex. 5), repeated blinks may be a communicative signal of acted surprise, a startle blink, provided it is spontaneous and not acted, although repeated, is not communicative: the Sender does not want to communicate his startle reaction to the others.

Biologically, the rapid closing of eyes in both startle and surprise might be functional to protect eyes from a potential sudden blow, thus fulfilling an instinctive self-defence function. This might be why among ancient Romans being able not to blink in front of danger was considered a cue of braveness for gladiators (see Plinius, quoted by Fornès

Pallicer and Puig Rodrìguez-Escalona, 2011). But non-communicative blinks of self-defence can also occur when the blow or injury is of a symbolic, not physical kind – for example, when receiving an insult or other unexpectedly severe offence. Here is a such case of self-defence blink (that, based on contextual cues, looks probably spontaneous, not intentionally mimicked):

(1) Gabul*: C'est vrais que les citoyens se demandent pourquoi ça va si long à Sion lorsque dans les autres municipalités qui ont beaucoup moins de moyens financiers, ça se passe* ***beaucoup plus vite****.*
(It's true that citizens wonder why in Sion it takes so long to resolve things, while in the other town halls, which have much fewer financial means, things are **much faster**). (in bold the words parallel to the gaze signals under analysis).

While the journalist Gabul is harshly criticizing the Vice-Mayor Feferler, and precisely during the phrase *beaucoup plus vite* (much faster), the latter performs rapid and repeated blinks, expressing his instinctive defence from this, albeit symbolic, attack.

Non communicative eye-closures

a) The most common example of non communicative eye closure – while sleeping – cannot be found in a debate.

b) A quite common type of non-communicative eye closure is while laughing. During laughter one may sometimes close eyes for a longer duration than in a blink. In collaborative and not competitive debates, a higher percentage of smiles and laughter are exchanged among the participants. In the closing of one debate in our corpus, where participants try to find solutions against the brain drain of young graduates from the Canton of Valais, one of the participants, Chiara Meichtry, assures the moderator and the public at home that they are looking for solutions in order to stop this 'exodus' towards other cantons or abroad. While doing so, she laughs and closes eyes for a duration longer than a blink.

(2) Meichtry: *Des solutions sont envisagées, voir on y* ***travaille****.*
(Solutions are foreseen, we are **working** on it.)

c) Eyes may be also used while thinking. When we are trying to remember something we can raise eyes up, when concentrating we may close eyes for a few seconds, isolating ourselves out of the surrounding space: this is the *cut off*, a type of eye-closure which can transmit information on the cognitive processes of the Sender (Morris 1977). These eye behaviours are not strictly communicative (Poggi 2007), in that they can be displayed exclusively to help the process of thought: they have the goal (either conscious or not) to help us concentrate and focus attention in order to reason better. Although by seeing us close our eyes our interlocutor can infer we are thinking, this doesn't imply that we intended to communicate this to him, so this eye closing is barely *informative*. But at times we may display our eye closing just to let the other know we are concentrating (and don't want to be disturbed or interrupted); in such a case, we can indeed speak of a *communicative* eye-closure.

7.2. Communicative eye closings

Having identified the items of gaze that in our view conveyed some meaning, we classified the meaningful items of eye closing as to their meaning. According to Poggi (2007), any communicative signal – words, prosody and intonation, gestures, gaze, facial expression, posture, body movement, therefore communicative eye closings too – can convey one of three basic kinds of information: about the World, the Sender's Identity, or the Sender's Mind. Information on the World concerns the concrete and abstract entities and events of the world outside the speaker (objects, persons, organisms, events, their place and time); Information on the Speaker's Identity concerns his/her age, sex, personality, cultural roots; while Information on the Speaker's Mind concerns the Speaker's mental states: his/her goals, beliefs and emotions. These kinds of information may be conveyed in verbal and body communication systems by means of specific signals called Mind Markers, more specifically, Belief Markers, Goal Markers and Emotion Markers.

7.3. Eye-closure and the Sender's Identity

Information about the Sender's Identity concerns the age, sex, personality or cultural roots of the person making the blink or eye-closure.

In the debate "Disability Insurance", Mr. Richoz, representing the blind people, counter-argues to his opponent's thesis, i.e. that the disabled should contribute to the decrease of the state's contribution to their support, by finding a job.

(3) Richoz : *A' la fin du processus on aurait fait des super chercheurs d'emplois certifiés, labélisés, à qui on aurait expliqué* ***comment chercher*** *un boulot, comment plaire à un employeur, comment dépasser l'handicap, mais au bout du compte, si on travaille pas sur le marché… c'est ça la réalité.*
(At the end of the process we would have transformed [the invalids] into super job searchers, to whom we would have explained **how to look for** a job, how to make a good impression to an employer, how to overcome their handicap, but in the end, if we don't work on the field… That's reality).

While reassuring the opponent (and the audience) about the actual invalids' efforts to obtain a qualification, search for a job, try to please the employer and to overcome their handicap, while uttering *comment chercher* (how to look for) **,** Richoz performs a *frown* and an *eye-closure*, which might be paraphrased as "I am concentrated in this effort", thus implying "we all are determined to do so". Richoz's *eye-closure* is somehow mimicking the invalids' determination in trying to do their best, thus conveying information on the invalids' identity. Taking into account that he himself makes part of the same category of people, and he himself attended training classes in order to obtain a qualification, we can say that his eye behaviour conveys information on his own identity.

7.4. Eye-closing and the Sender's Mind

Among the types of information on the Sender's Mind that can be conveyed by a communicative signal, Poggi (2007) distinguishes *Belief Markers, Goal Markers* and *Emotion Markers*. *Belief Markers* inform on the Sender's degree of certainty regarding the stated message, *Goal Markers* on one's goals while delivering the message and finally, *Emotion Markers* convey the emotions being felt during or regarding the situation described.

Belief Markers

Belief Markers inform about the degree of certainty we attribute to the beliefs we are speaking about. This information (to be distinguished from emphasis, that concerns Goal Markers and refers to the importance we attribute to the goal of communicating those beliefs) can be conveyed not only verbally, by verbal markers such as *absolutely*, *probably* or *possibly*, but also through gestural and eye behaviour. With an eye-closure, one can confirm either one's own or the interlocutor's utterances. The meaning conveyed by this kind of eye-closure is fairly equivalent to saying *'Yes'*, hence it counts as a confirmation.

In this example, the journalist Gabul expresses an opinion about the seriousness with which files are examined by the city council.

(4) Gabul: *L'impression que donne le vice-président à la municipalité, c'est qu'effectivement, les dossiers sont* ***mûris****, sont réfléchis, etc.*
(The impression given by the vice-mayor is that indeed, the files are carefully **examined**, reasoned, etc.)

While saying that the files are carefully examined (*mûris*), Gabul performs an eye-closure of confirmation which conveys his degree of certainty of his statement. It might then be paraphrased as "Absolutely, I am very certain of that".

In a previous paper, Poggi et al. (2010 b) proposed a classification of nods on the basis of the meanings they convey. In the light of these new findings on blinks, we can state that the eye-closure (especially if long in duration and with a higher tension on the lower eyelid) while nodding or while shaking head, conveys a higher degree of conviction with respect to nodding/head shaking alone. When accompanied by a nod or a head shake, eye-

closure can be seen therefore as an intensifier of the degree of conviction of the sender in what he is saying or hearing, like in the following examples.

In the first one, extracted from the debate on Disability Insurance, Mr. Rossini, a deputy of the Socialist Party, who is against the idea of reducing financial support to disabled persons, categorically rejects his opponent's opinion that he and his party promote a politics based on words and not on facts.

(5) Chevrier : *Vous avez simplement voulu faire de la politique politicienne…*
Rossini : *Non, on fait pas politique, non, on fait pas de politique politicienne.*
Chevrier: *…. à travers ce référendum, alors que sur le fond vous êtes convaincu que c'est une bonne révision.*
Rossini : ***Non***.
(Chevrier: You simply wanted to play party politics…
Rossini: No, we don't make politics, no, we don't play party politics.
Chevrier: ...by proposing this referendum, while deep down you are convinced that it's a good revision.
Rossini: **No**.)

While saying *'No'*, Rossini performs a *head shake* accompanied by an *eye-closure* which has the role of intensifying his being categorical when denying the accusations.

Emotion Markers

Another category of Mind Markers are *emotion markers*, i.e. signals bearing information on the Sender's emotions. Among the emotions that can be expressed by eye behaviour we mention *surprise*, either really felt or only acted, and *acted desperation*.

Surprise

A typical eye behaviour to signal surprise is *raising the eyebrows*; besides this, Ekman & Friesen (2007) mention *wide open eyes* as signals conveying surprise, adding that a high degree of intensity of this emotion may be also expressed by mouth opening (*jaw drop*). Such strong signals of surprise do not occur in political debates. Other signals are performed to convey surprise (real, pretended, or acted): *eyebrow raising* combined with *eyes wide open* and *repeated blinks*. We agree with Ekman & Friesen (2007) that surprise is expressed in general by *raised eyebrows* and *wide open eyes*, but our hypothesis is that surprise (only acted or actually felt at a certain moment in time and now re-expressed, therefore mimicked) can be conveyed by rapid repeated blinks. In this example, Mr. Feferler speaks about the surprise felt by other town hall workers and himself when a questionnaire came out in which the inhabitants of Valais were asked to assess the town hall's activity.

(6) Feferler: *Alors, écoutez, bon ben…Je dirais que quand ce questionnaire est sorti, à la veille des élections, ça nous a* ***un petit peu*** *surpris et puis je crois que cette surprise, elle pouvait s'expliquer parce qu'il y a avait les élections qui arrivaient.*
(Feferler : So, listen, well…I would say that when this questionnaire came out, a day before the elections, it surprised us **a bit** and I think that this surprise could be explained by the immediate arrival of elections.)

While pronouncing *un petit peu* ([it surprised us] a little), he makes a series of *rapid repeated blinks* accompanied by *raised eyebrows*, as if mimicking the surprise he felt in that particular moment when the questionnaire came out.

While this is a case of real surprise, actually felt at a particular moment in time, and now, in the moment of the story telling, recalled and iconically acted, here is an example in which surprise is not felt but only acted.

Acted surprise

Repeated blinks may occur in acted surprise, in this case being communicative: my (pretended) amazement in front of the speaker's statement or behaviour is so intense that I rapidly shake head and repeatedly blink, to show I want to convince myself I am not dreaming, like if I were rubbing my eyes for surprise or pinching myself to make sure I'm awake. While these behaviours are more likely performed when confronted with truly amazing situations, repeated blinks mimicking surprise are more often produced while listening to someone's discourse as a back-channel signal that conveys, in an indirect

manner, disagreement with the Speaker. In the debate *"Libre circulation"* (Free circulation) members of two different parties, Radicals and Christian Democrats, argue against each other. The former party sustains the free circulation of Polish workers, while the latter encourages the population to vote against it. In the fragment below Mr. Freysinger, member of the Christian Democrats speaking about the exodus of people from less economically developed countries towards Western countries, concludes:

(7) Freysinger: *Et c'est **pas ça** le modèle de la société équilibrée.*
(And **that** is **not** the model of a balanced society).

While saying *"c'est pas ça"* (that is not), Mr. Freysinger performs a series of rapid repeated blinks and makes a pause, gazing from the audience to the moderator and to his opponent, addressing, therefore, all of them. His rapid repeated blinks convey surprise and his eye behaviour seems to state 'I am very surprised that you don't realize in what an absurd society we are living in'. But since showing surprise means that what happens is completely unexpected, possibly awkward, acting surprised in this case is an indirect way to convey disagreement with the opponent.

Acted desperation
In another case, by a blink Mr. Freysinger enacts another emotion: desperation (Table 1).

(8) Moderator : *J'aimerais qu'on aborde la troisième partie de ce débat, à savoir si les garanties sont vraiment des garanties offertes par la **confédération**.*
(I would like to tackle the third part of the debate, more precisely the issue whether the warranties offered by the **confederation** are real warranties).

As an answer to the Moderator's question, Mr. Freysinger *shakes his head*, *raises eyebrows* and performs an *eye-closure* with *pressed eyelids*. His facial expression shows acted desperation, as if he were resigned in front of the Moderator's incapacity to understand the real situation. Also in this case, acted desperation, at the indirect level, conveys a deep disagreement.

Goal Markers
Goal Markers are all the signals that inform about the goal of the Sender's sentences (their performative) but also the structure of the sentences and discourses s/he is delivering, that is, how s/he intends to distribute information and connect sentences in a discourse. Thus, *meta-sentence* goal markers signal the beginning or the end of a sentence or phrase (syntactic goals, marked for example by intonation), or the comment (the new and more important information of the sentence, marked by emphasis); *meta-discursive* goal markers signal which parts, within the structure of his discourse, the Speaker considers important or less important, so much so to be possibly passed over.
Some items of both *blinks* and *eye-closures* in our corpus convey meta-sentence and meta-discursive information.

Syntactic eye-closure
Sometimes the eye-closure has a syntactic function: it signals the start of a sentence. In our corpus, this function is exploited in a case of misspelling and self-correction: one makes an error and signals one is restarting the sentence to correct oneself.

In the debate about the town hall's efficacy, the vice-mayor Mr. Feferler is talking about a decision made by the General Council: while quoting the numbers of votes, respectively, in favour, against and abstained, he makes a mistake, and then restarts to correct himself.

(9) Feferler : *Il faut savoir que le Conseil Général en 2003 a pris une décision par quarante-six 'oui', une abstention, **euh** quarante-six 'oui', un 'non' et six abstentions.*
(We must say that the General Council took a decision in 2003, with forty-six 'yes', one abstention, **euh**, forty-six 'yes', one 'no' and six abstentions).

As he realizes he has said "one abstention" instead of "one no", he performs a *rapid eye-closure* with *raised eyebrows* and a *violent nod*. The meaning of his body behaviour is 'I correct myself and I start all over again'. The *eye-closure* functions in this case as a

demarcation of where the Speaker stops and starts all over again. An alternative interpretation is that all three movements are triggered by the cognitive load of self correction.

Blinks too can work as demarcation signals. In the debate about "Héliski", Darbellay, a Green deputy, and Pouget, a helicopter pilot, discuss about whether taking people by helicopter to ski on the mountains should be banned since it represents a threat for the environment. Pouget, who claims this kind of sport is not at all harmful for nature, is interrupted by Darbellay, arguing against his thesis.

(10) Pouget : *Vous dites qu'on veut pas rentrer en matière. Non, pas sur une réduction parce que je pense…*
Darbellay : *Ah ah…*
Pouget : … ***on a on a*** *rien à gagner, à tous les niveaux […], on n'a rien à gagner d'une réduction du nombre de rotations en montagne, d'autant plus qu'elles sont quand même assez minimes…*
(Pouget : You say that we refuse to consider this issue. No, not the issue of a reduction [of flights], because I think…
Darbellay: Ah ah…
Pouget: … **we have we have** nothing to gain, at all levels, […] we have nothing to gain from a reduction of the flights number in the mountains, even more so since they are also rather rare…).

When Mr. Darbellay tries to intrude into Pouget's turn and take the floor, Pouget performs *two rapid blinks*, preceded by a *strict and irritated gaze* directed to his opponent. His eye behaviour might be rendered by the following sentence "I am irritated because you don't allow me to go on and therefore I start all over again". But at the same time the *double blink* marks the beginning of his repetition: *'on a on a'* (we have we have) and makes part of a strategy of floor keeping.

Emphasis blink

One of the Speaker's goals is to stress the main concepts of one's speech. Among the body communication strategies through which we emphasize the comment of our sentences, i.e. the new information, beat gestures and eyebrow raising are the most frequent, hence the most studied ones. But other signals convey emphasis too, such as a sudden *widening of eye aperture* or *repeated blinks*. *Rapid repeated blinks* can be used as a punctuation mark during speech: a Speaker performing a sequence of quick blinks while pronouncing an important concept may be signalling s/he has stated something important and attracting the interlocutor's attention on it.

This is what Mrs. Bressoud does in the debate "Mothers as educators". She is a frequent blinker, but moreover, while pronouncing key words for her argumentation, she performs a series of rapid repeated blinks to attract the listener's attention .

(11) Bressoud : *C'est pas dire qu'elles sont pas capables, la démarche est totalement* ***différente****, de pouvoir s'occuper des propres enfants et de pouvoir en deuxième temps de prendre en charge les enfants des autres.*
(It's not to say that they [mothers] are not capable, the approach is totally **different**, taking care of their own children and taking other people's children in charge).

Unimportance eye-closure

So far we have seen speakers whose blinks marked the key concepts of their discourse. In other cases, though, one may need to communicate that some topic can be left out since it is not essential for present discourse. Interestingly enough, this is not conveyed by a blink but by an *eye-closure*. We have seen cases of this in previous observation, but here is one from our present corpus.

While speaking about the total of flights made for Héliski, the helicopter pilot Pouget mentions that their number is not that important.

(12) Pouget : *On pourra parler plus tard du nombre des vols qui l'on fait en Héliski, qui n'est* ***pas si important que ça****, je pourrais vous donner des exemples en comparaison des transport qui l'on fait pour les cabanes de SAS, par exemple pour tout autre transport en montagne.*
(We could speak later about the number of flights we make for Héliski, it's **not that important as that**; I could give

you examples as compared to the number of transports we make for the SAS chalets, for instance, or for all other types of transport in the mountains).

While saying *n'est pas si important* (it is not that important), Pouget performs a *slow eye-closure*, that looks as a bodily synonym of what he says in words, meaning "I am skipping this part, as I don't consider it important for the present conversation".

8. Conclusions

The aim of this paper was to prove that eyes can communicate meanings not only while gazing, but even when not looking. Following our study, we can say that through blinks and eye-closure one can confirm the Speaker's speech, intensify or stress one's own discourse, mimic personal traits as determination or emotions such as surprise and desperation, delimit the beginning of a new sentence. Our approach in this paper was qualitative: first we distinguished between communicative and non communicative eye behaviours and then we tried to individuate the possible meanings conveyed by the communicative items of blinks and eye-closure. In our further work we will attempt a quantitative approach to investigate whether blinking is influenced by social context, culture and personality.

Acknowledgments. Research supported by the European Network of ExcellenceSSPNet (Social Signal Processing Network), VII Framework Program, G.A. N.231287.

References

Argyle, Michael & Cook, Mark (1976) *Gaze and mutual gaze*. Cambridge: Cambridge University Press.

Costa, Marco & Ricci Bitti, Pio, Enrico (2003) "Il chiasso delle sopracciglia". In *Psicologia Contemporanea*, 176: 38-47

Duncan, Starkey (1974) "Some signals and rules for taking speaking turns in conversations". In S. Weitz (Ed.) *Nonverbal communication*. Oxford: Oxford University Press.

Eibl-Eibesfeldt, Irenäus (1972). "Similarities and differences between cultures in expressive movements". In R. Hinde (Ed.) *Non verbal communication*. London: Cambridge University Press: 297-314.

Ekman, Paul (1979) "About brows: Emotional and conversational signals". In M. von Cranach, K. Foppa, W. Lepenies, & D. Ploog (Eds.), *Human ethology: Claims and limits of a new discipline: contributions to the Colloquium*. Cambridge University Press: 169-248.

Ekman, Paul; Friesen, Wallace & Hager, Joseph (2002) *Facial Action Coding System. The Manual.* Published by Research Nexus division of Network Information Research Corporation, Salt Lake City, USA.

Fornès Pallicer, A., Puig Rodrìguez-Escalona, M (in press). Comunicar con la Mirada en la Roma Antigua: el movimiento de pàrpados. Faventia.

Goodwin, Charles (1991) *Conversational organization. Interaction between speakers and hearers*. New York: Academic Press.

Hartmann, Björn; Mancini, Maurizio & Pelachaud, Catherine (2002) "Formational Parameters and Adaptive Prototype Instantiation for MPEG-4 Compliant Gesture Synthesis". In *Computer Animation* 2002: 111-119.

Heylen, Dirk (2005) "A closer look at gaze". Proceedings of the 4th International Joint conference on *Autonomous Agents and Multimodal Agent Systems 05.*

Kendon, Adam & Cook, Mark (1969). "The consistency of gaze pattern in social interaction". In *British Journal of Psychology*, 60: 48-94.

Kendon, Adam (1973) "A description of some human greetings". In R. Michael & J. Crook (Eds.) *Comparative Ethology and Behaviour of Primates*. New York: Academic Press: 591-668.

Leal, S.,Vrij, A. (2008) "Blinking during and after lying". In Journal of Nonverbal Behaviour,International Conference on *Interactive Virtual Agents*. Kos, Greece

Morris, Desmond (1977) *Manwatching*. London: Jonathan Cape.

Pelachaud, Catherine & Prevost, Scott (1994) "Sight and sound: Generating facial expressions and spoken intonation from context". In Proceedings of the 2nd ESCA/AAAI/IEEE Workshop on *Speech Synthesis*. New Paltz, New York: 216-219.

DePaulo, B. M., Kirkendol, S.E.: "The motivational impairment effect in the

communication of deception". In J.C. Yuille (Ed.) Credibility assessment. Dordrecht, The Netherlands: Kluwer, 51--70, (2003)

Poggi, Isabella (2002) "Mind markers". In M. Rector, I. Poggi, N.T., ed.: *Gestures. Meaning and use*. University Fernando Pessoa Press, Oporto, Portugal

Poggi, Isabella (2007) *Mind, Hands, Face and Body. A goal and belief view of multimodal communication*. Weidler Buchverlag

Poggi, Isabella, D'Errico, Francesca, Spagnolo, Alessia (2010 a) "The Embodied Morphemes of Gaze. In S. Kopp and I. Wachsmuth (Eds.): GW 2009, LNAI 5934, Springer-Verlag Berlin Heidelberg, pp. 34–46.

Poggi, Isabella, D'Errico, Francesca, Vincze, Laura (2010 b) "Types of Nods. The polysemy of a social signal". In Proceedings of the 7th International Conference on *Language Resources and Evaluation* (LREC), Malta, 19-21 May 2010.

Zuckerman, M., DePaulo, B.M, Rosenthal, R.: Verbal and nonverbal communication of deception. In L. Berkowitz (Ed.) Advances in experimental social psychology, vol. 14; New York: Academic Press, 1--57, (1981)

1. **Time**	**2.** **Speech**	**3.** **Action**	**4.** **Goal or meaning**	**5.** **Supergoal**	**6.** **Type**
1. 15.06 Moderator (Speaker)	*more precisely the issue is whether the warranties offered by the confederation are real warranties.*	**Head**: Head shake	No		
15.12 Freysinger (Listener)		**Gaze**: Eye closure Eyebrow raising	I am desperate → They really don't understand I am superior → Poor them, they really don't get it.		Communica-tive (Emotion)
2. 23.32 Gabul	*mu-municipalité*	**Gaze**: blink	Accompanies the effort of uttering the syllable		Non-communica-tive

Table 1 Annotation scheme

Towards an integrated view of gestures related to speech

Elisabeth Ahlsén
University of Gothenburg
Gothenburg, Sweden
eliza@ling.gu.se

Abstract

This study addresses the use of co-speech gestures in informal face-to-face interaction involving persons with and without aphasia (language disorder caused by acquired brain damage). A central question in aphasia research is whether gestures are better preserved when speech is impaired by aphasia and, if this is the case, can compensate for word finding problems in speech. This question is intimately related to the competing views between researchers who believe that gesture and speech are part of one system and generated in a totally interdependent way and researchers who believe that gestures and speech are generated by two different systems. In the first case, compensation would be impossible, whereas in the second case, compensation would be expected. A less categorical stance is suggested, based on an comparative empirical study of co-speech gestures in a database of 400 co-speech gesture produced by persons with and without aphasia.

1 Introduction

There are several reasons for studying co-speech gestures produced by persons with aphasia. (Gesture is here used in a wide sense for communicative body movements), one reason being the controversy concerning if and how the generation of gesture and the generation of speech are related. The idea of gestures possibly being more robust relates to the idea of gestures being evolutionary precursors of speech. There is a strong practical interest in finding out to what extent gestures can or cannot be used for compensation and how this can be used in communication therapy. It is also of great value to families and hospital staff to know more about if and how spontaneous gesturing can be used by persons with aphasia.

The theoretical controversy concerning the gesture-speech relation contains, on the one hand, (i) the view that speech and gesture are inextricably intertwined in development and generation (e.g. the growth point theory, which makes speech and gesture interdependent and simultaneous and which entails that if one is disturbed, so is the other (e.g. McNeill, 21992, 2000, 2007). On the other hand, (ii) the view that gesture and speech generation are two independent separate systems which means that gestures can replace or facilitate speech has been proposed, for example, by (Krauss et al., 2000, Hadar and Butterworth, 1997, Beattie and Shovelton, 2000, 2002, 2004). A less categorical view is that maybe gesture and speech generation are closely related but also to some extent independent. If gestures came earlier in evolution, they can be more robust and, thus, they can be candidates for compensatory use, either replacing words or adding information. Gestures can sometimes be more preserved in aphasia (e.g, Feyereisen et al., 1990, Ahlsén. 1985, 1991). There is, for example, the strong argument for stepwise evolution via less complex and more complex gestures to speech and language (from ”grasping an object” to ”Verb-Argument-structures”) presented by Arbib (2005), which draws on mirror neurons and the fact that Broca’s area developed on top of the mirror neuron (F4) area in the macaque

Related to this controversy, there is also the question whether gesture is mainly for the speaker or mainly for the hearer.

Some earlier findings in pursuing the questions above by studying mainly spontaneous gesture and speech production by persons with aphasia are the following. In persons with aphasia as a group an increase of

gestures in spontaneous speech can be found, compared to a reference group, i.e. a group with aphasia (although not all individuals) used significantly more gestures in spontaneous conversation than a matched group of persons without aphasia. Gestures were used spontaneously with compensatory function. Persons with severe apraxia (practically unable to produce actions, gestures, movements from instructions or to imitate them) still used spontaneous gesturing with compensatory function extensively. (Ahlsén, 1985, Macauley and Handley, 2005). Concerning the relation between action and communication, an Activity based Communication Analysis (c.f. Allwood, 2000, 2002) showed that a person with aphasia acquired a more favorable role and increased communicative ability in an activity which allowed action for communication, than in a pure verbal conversation activity. (Ahlsén, 2002)

Further support for a view that gesture use in aphasia can have a compensatory function was found in a case study of a person (HS) with an initially global aphasia which developed into a Wernicke's aphasia and further into a mainly anomic aphasia over a period of four years. HS was studied during three years of intensive treatment/courses (from 4 to 7 years post onset). Initially, he showed an extensive use of gestures – illustrating, pantomimic and others – together with a severe word finding problems. A decrease in gestures occurred, that paralleled an increased word finding ability (Ahlsén 1991), thus implying that the earlier use of gestures was not a general habit, but a compensatory use which disappeared when it was no longer needed.

The present study takes its point of departure in a perspective of embodied cognition and communication and is investigating gesturing behavior as a window to processing in finding and producing intended words or utterances. More specifically, it focuses on the correlation and complementarity of gestures and words.

The study relates to the following general questions: How much are gestures and words connected/intertwined in production? Are they disturbed in the same way when word finding problems occur? How much can gestures compensate for word finding problems?, and What can gestures tell us about the word finding process? The inclusion of data from persons with aphasia as well as the inclusion of "trouble spots" (see below) is intended to provide further information related to these questions.

2 Method

A corpus of 400 co-speech gestures associated with the production of content words and phrases in persons with and without aphasia was extracted from a corpus of video-recorded face-to-face interaction.

Two types of gesture contexts were included in two separate sub-corpora.
(i) The Verb-Noun (VN) Context: Gestures with representing/illustrating associated to the production of main Verbs and Nouns in fairly fluent speech
(ii) The Own Communication Management (OCM) Context_ Gestures associated with problems of word finding/word production involving other overt signs of own communication management (OCM) i.e. choice and change operations (cf. Allwood et al. 1990, Allwood et al. 2007).

The two subcorpora were identified with the purpose to study co-speech gestures both as related to verb and noun production (as examples of typical categorematic/content words) and to overtly manifested "trouble spots" in word finding/speech production.

For each of the two contexts a subcorpus of 100 gestures produced by persons with aphasia (The Aphasia corpus) and 100 gestures produced by persons without aphasia (The Reference corpus) was selected. Data was selected from 10 persons in each category. The corpus, thus contained four subcorpora with 100 co-speech gestures in each.

The corpus in total was coded according to the following coding schema.
- Share of noun versus verb context (for the VN context)
- Function (representational and/or OCM)
- Choice or change function (for the OCM context)
- Timing: stroke before, simultaneous with and/or after spoken "target word" (where a target word could be identified)
- Body part: 2 hands, 1 hand (left or right), head
- Gaze direction: towards the interlocutor or averted (specified for direction)

- Semantic features of content: shape, location, action, event
- Complexity features of hand movement: change of hand shape, complex hand-finger movements

Interrater reliability was ensured by originally coding an extended number of examples and subsequently chosing 100 examples for each of the four subcorpora where codings were in agreement between the two coders.

3 Results

A summarizing overview of the results is presented in Table 1 below.

Coded feature	Aph	Ref
Share of noun vs. verb context	49/51	44/56
Function: (repr in OCM context)	>30%	>30%
Choice vs. change function –	77/39	76/38
Gesture stroke before/simult. with word (VN context)	18/66	17/61
2 hands/1 hand		
VN context	32/64	46/54
OCM context	6/80	29/67
Head		
OCM context	35	12
Gaze at IL/averted		
VN context	70/30	96/4
OCM context	26/61	89/4
Semantic features:		
shape	36	26
location	23	13
action/event	60	72
Complex hand-finger movement	12	21

Aph = Aphasia database
Ref = Reference database

Table 1. Summary of results for selected features of gestures in relation to speech.

The share of noun versus verb contexts for gesturing turned out to be fairly similar for the aphasia and reference databases. However, the reference database contained more verb contexts than noun contexts, while this tendency was not as strong in the aphasia database. Furthermore, both the databases contained a number of action gestures for nouns as well as a number of shape gestures for verbs. This can mainly be explained by an action orientation of certain nouns, like ”keyboard”, which is illustrated by typing finger movements and an object/place orientation for certain verbs, like ”to bike”, which can be illustrated by a pointing gesture outlining a wheel.

In the verb-noun context, all the gestures contained illustrating/representational features. In the OCM context, however, the gestures tended to be self-activating, but a substantial share of these gestures (more than 30%), depending on the restrictions of the definition of illustrating feature as including some metaphoric gestures or not) also contained an illustrating/ representational feature, this tended to be somewhat more frequent in the aphasia database.

In the OCM context, the share of gestures with choice and change function, respectively, was the same in both the databases.

The timing of the gesture stroke in relation to the spoken ”target word” in the verb-noun context was the same in both the databases. Most often, the gesture stroke was simultaneous with the spoken word, but it can be noted that in 17-18% of the cases the gesture stroke preceded the spoken word.

As expected, the aphasia database contained more one-hand gestures, using the left hand, than the reference database. This applies to both of the contexts and, in general,, this can be seen as a consequence of an earlier or to some extent remaining right arm-hand hemiplegia. This, does, however, imply that the aphasia database contained less right hand and bimanual gestures. Especially bimanual gestures have been taken as a feature indicating an increased complexity of the gestures, compared to one-hand gestures.

There was also a considerable difference in the gaze direction during gesture production between the aphasia and reference databases, for both the contexts, although even more pronounced for the OCM context. The reference group in general upheld mutual eye contact with their interlocutors during gesturing, although somewhat less in the OCM

context than in the verb-noun context. The persons with aphasia, on the other hand, showed much more gaze aversion during gesturing in both the contexts, even in almost all the cases in the OCM context. Gaze aversion is generally taken as a sign of increased cognitive load and this is, then, an obvious feature related to word production and gesturing for persons with aphasia, even when no other overt signs of word finding problems are shown, as in the verb-noun database. It can also be noted that the gaze aversion can be further divided into subgroups like looking out into the air, looking down at the table, looking at one's own gesturing hands and seemingly looking at an imagined object or scene. The latter two of these subgroups did not occur in the reference database and can provide some cues related to the word-search/word-finding process. For example, looking at one's own gesturing hands has been interpreted as directing the gaze of the interlocutor to the gesture (cf. Gullberg and Kita, 2009) and can also relate to self-activation of information with the help of gesturing.

Some features of gesturing in both the databases were the occurrence of illustrating features in mainly self-activating gestures during word search and the occurrence of gesture strokes before the spoken word. There were, thus, certain possibilities of conveying compensating information via gesturing, when words are failing. Other common features were the action and object bias of a word sometimes overriding the related noun and verb word classes in the type of gesture, and the increased gaze aversion in cases of own communication management. Some differences were that the aphasia database, specifically, contained more one-hand gestures using the left hand (caused by hemiplegia), more gaze aversion, and more varied direction of gaze in cases of gaze aversion..

So what is the content of the gestures in the databases? When we turn to the semantic features of content, we find that the order of preference is the same in both the databases, with illustration of action/event being the most frequent feature, often accompanied by other functional features and some complexity on arm-hand movements, while illustration of action is less frequent and illustration of location, especially in relation to body, even less frequent. The two latter features co-occur quite often. In the aphasia database, however, the difference between action and object related gestures is smaller than in the reference database and there is, thus, more use of object and location features in the gestures produced by persons with aphasia.

4 Discussion

There are important similarities between the reference and aphasia databases in our study, which point to similar processing of both groups in generating gesture and speech production. It further points to a great deal of preserved gesture production in the persons with aphasia. Since the number of sampled gestures in the two databases was the same, no conclusions about the amount of gesturing can be drawn in this study, only about the features of gestures in relation to speech and they seem to be similar to a great extent. (There are, however, findings of an increased use of gestures by persons with aphasia in informal communication, cf. Ahlsén, 1985, Lott, 1999), although the frequency of gesture is likely to be subject to individual differences as well as other influences, such as the activity type. Although gestures and speech seem to mostly be generated in close relation, it is not immediately determinable from this overview analysis how much they are interdependent during the completion of the expression. The findings that the gesture stroke sometimes occurs before the spoken word and that some of the self-activating gestures related to own communication management in speech contain illustrating/representing semantic features indicate a possible discrepancy in timing as well as semantic content between gesture and word in the actual expression. The cases where a person with aphasia looks at his/her own gesturing hands or an imagined object or scene also point to a possible function of the gesture in evoking the spoken expression in the production of the speaker and/or the comprehension of the interlocutor. It seems likely that gestures can have a double function in both being of help for the producer and the recipient (or co-producer). See also studies by Rauscher, et. al. (1996), Kita (2000), Melinger and Kita, (2006), Ruiter (2006) and Morrel-Samuels and Krauss (1992).

There are, thus, features of co-speech gestures that make it possible for them to fill some compensatory functions. This does not, however, entail that the gestures are

necessarily intact, i.e. there is no evidence in the data that gesturing is not *at all* affected by the aphasia, even if gesturing is to a great extent functioning adequately in relation to speech. There are certain findings in our data that suggest that gesturing might be affected, in a primary and/or possibly secondary manner, in the persons with aphasia. These findings are that the semantic features of gestures are more related to objects, shapes and location in relation to the body and less related to action and complex functional movements in the aphasia database than in the reference database. The complexity of one-hand gestures produced by persons with traces of hemiplegia seems generally lower than that in the reference data, making this secondary influence hard to distinguish from a possibly more primary influence of a lower semantic complexity. From this overview data analysis, it can, thus, both be hypothesized that gestures in the aphasia group can be somewhat affected in relation to the aphasia (in a primary as well as a secondary way) and that gestures have the potential for compensatory use in cases of word finding problems. There is a possibility for some inter-dependence, as well as for a certain independence between gesture and speech.

There are a number of caveats related to overview results and necessitating a further, more detailed study of each co-speech gesture in its context. One such caveat is that while it is important to capture co-speech gestures in informal face-to-face interaction, this also involves a certain variation in the topics of discussion and there is some variation in topics between the two databases. Individual personalities and ways of expression of the subjects can also, to some extent, influence the selected databases, even though the selection was based on consecutive occurrences in 10 different persons for each of the two databases. Most of this possible influence was probably eliminated by the sampling procedure, but there might still be some differences and this will be the subject of further studies of the vocabulary co-occurring with the gestures.

Acknowledgements

We want to thank the Swedish Research Council for supporting this study (grant VR421 2006 1434)..

References

Ahlsén, E. (1985). *Discourse Patterns in Aphasia. Gothenburg Monographs in Linguistics* 5, University of Gothenburg, Department of Linguistics.

Ahlsén, E. (1991). "Body communication and speech in an Wernicke's aphasic - A longitudinal study", *Journal of Communication Disorders* 24, 1-12.

Ahlsén, E. 2002. Speech, vision and aphasic communication. In, Mc Kevitt, P., O'Nualláin, S. & Mulvihill, C. (eds) *Language, Vision and Music*. Amsterdam: John Benjamins, pp. 137-148.

Allwood, J. (2000). An Activity Based Approach to Pragmatics". In Bunt, H., & Black, B. (Eds.) *Abduction, Belief and Context in Dialogue: Studies in Computational Pragmatics*. Amsterdam, John Benjamins, pp. 47-80.

Allwood, J. (2001). Capturing Differences between Social Activities in Spoken Language. In Kenesei, I., & Harnish, R.M. (Eds.) *Perspectives on Semantics, Pragmatics and Discourse*. Amsterdam: John Benjamins, pp. 301-319.

Allwood, J., Ahlsén, E., Lund, J. & Sundqvist, J. (2007). Multimodality in own communication management. In J. Toivanen & P. Juel Henrichsen (Eds*.) Current Trends in Research on Spoken Language in the Nordic Countries*, Vol. II:, Oulu: Oulu University Press, pp. 10-19.

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

Beattie G.W., Shovelton H. K. (2004). Body Language. *Oxford companion to the mind*. Oxford University Press.

Beattie, G.W., & Shovelton, H. K. (2000). Iconic hand gestures and predictability of words in context in spontaneous speech. *British Journal of Psychology, 91*, 473-492.

Beattie, G.H., & Shovelton, H.K. (2002). An experimental investigation of some properties of individual iconic gestures that mediate their communicative power. *British Journal of Psychology, 93*, 179-192.

Feyereisen, P., Bouchat, M.-P., Dery, D., & Ruiz, M. (1990). The concomitance of speech and

manual gesture in aphasic participants. In E. Hammond (Ed.), *Cerebral control of speech and limb movements* (pp. 15–21). North Holland: Elsevier Science.

Gullberg, M. and Kita, S. (2009). Attention to speech-accompanying gestures: Eye movements and information uptake. *Journal of Nonverbal Behavior* 33(4): 251-257.

Hadar, U. & Butterworth, B. (1997. Iconic gesture, imagery and word retrieval in speech. *Semiotica*, 115, 147-172.

Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.), *Language and gesture: Window into thought and action*, pp.162–185. Cambridge, UK: Cambridge University Press.

Krauss, T. M., Chen, Y. & Gottesman, R. F. (2000). Lexical gestures and lexical access: a process model. In D. McNeill (Ed.), Language and gesture (pp. 261-283). New York: Cambridge University Press.

Lott, P. (1999). *Gesture and Aphasia*. Bern: Peter Lang.

Macauley, B.L. & Handley, C. (2005). Conversational gesture production by aphasic patients with ideomotor apraxia. *Contemporary Issues in Communication Sciences and Disorders 32*, 30-37.

McNeill, D. (1992). *Hand and Mind.* Chicago : The University of Chicago Press.

McNeill, D. (2000). *Language and gesture*. Cambridge : Cambridge University Press.

McNeill, David (2007). *Gesture and Thought*. Chicago: University of Chicago Press.

Morrell-Samuels, P. & Krauss, R. M. (1992) Word familiarity predicts temporal asynchrony of hand gestuers and speech. *Learning, Memory, and Cognition* 1992, Vol. 18, No. 3, 615-622.

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 (4), 226-231.

De Ruiter, J.P. (2006). Can gesticulation help aphasic people speak, or rather, communicate? *Advances in Speech-Language Pathology 8(2)* p.124-127.

Strategies of multimodality in communication following traumatic brain injury in adolescence

Åsa Fyrberg
University of Gothenburg
Göteborg, Sweden
asa.fyrberg@vgregion.se

Elisabeth Ahlsén
University of Gothenburg
Göteborg, Sweden
elisabeth.ahlsen@gu.se

Abstract

The purpose of this study was to explore the multimodal communicative ability of a young survivor of a moderate traumatic brain injury (TBI) in situations involving one or two other speakers.

A single subject design was applied, including a 17 year old adolescent with TBI. The study uses a triangulation of methods, evaluating both quantitative and qualitative data:

1) Analysis of Multimodal Communication Management (MCM) in video-recorded conversations.

2) Assessment of communicative skills in The Communicative Effectiveness Index - CETI (Lomas et al., 1989) by subject and parents.

3) Clinical neuropsychological and speech language assessments.

MCM differed with the number of interlocutors involved. In the two-partite dialogue (TWP), the tempo was lower compared to the three party conversations (TRP) and this facilitated language comprehension and turn-taking for the brain injured adolescent. Analyses in TWP showed frequent use of mutual gaze in collaboration with iconic hand gestures, particularly in moments of impaired word-finding. In TRP, the dominant role for the subject was as a listener since he rarely took turns in the dialogue.

The evaluation of daily communication in the CETI also identified trouble spots in high-speed communicative situations with several people involved. Formal tests verified reduced verbal abilities, corroborating impaired function in situations with high cognitive and communicative load.

1 Introduction

Communication problems following a traumatic brain injury (TBI) have been described as manifestations of general impairments to cognitive and executive systems (Ylvisaker and Feeney 2007) and cognitive-communication disorder the most prevalent form of communication disorders as a consequence of TBI (Sarno 1980). A definition is formulated in a position statement by the American Speech and Hearing Association (ASHA, 2005, p. 1):

"Cognitive-communication disorders (CDD's) encompass difficulty with any aspect of communication that is affected by disruption of cognition. Communication may be verbal or nonverbal and includes listening, speaking, gesturing, reading, and writing in all domains of language (phonologic, morphologic, syntactic, semantic, and pragmatic). Cognition includes cognitive processes and systems (e.g. attention, perception, memory, organization, executive function). Areas of function affected by cognitive impairments include behavioural self-regulation, social interaction, activities of daily living, learning and academic performance, and vocational performance."

The survival in victims of TBI has increased substantially in recent decades as a result of improved medical treatment methods. However, many survivors are left with lifelong cognitive and communicative impairments as a consequence of the trauma, severely affecting everyday communication skills (Wahlström Rodling et al., 2005).

For the ease of description, the concept "cognitive-communication" will henceforth be referred to as "communication", unless otherwise noted.

The impact of TBI traumas, especially in the moderate to severe cases, has the nature of a developing "invisible communicative disorder or handicap", corresponding to the fact that there are few immediately visible or audible external signs of a brain damage in many individuals (Chamberlain 2006). Subjects describe a lack of consistent empathic responses from others during recovery and some experience a difficulty from the environment to adjust and accept them (Roscigno et al., 2011).

The main goal for many adolescents suffering from communicative impairments after TBI is to recover their pre-injury level of functioning to fit in with the social environment they belong to. This may seem like a possible outcome after the conclusion of a period of hospital treatment and clinical assessment. However, it is not until demands are put on the young person to participate in everyday conversations, group dialogues or academic learning setups that the extent of the impediments becomes clear (Hux et al., 2010).

This study explores the use of multimodal communication patterns and how the analysis of such patterns can add to standard test proceedings in creating a more comprehensive description of the subject's communication and identify rehabilitation strategies.

1.1 The examination of communication after TBI

A traditional way to set goals for communicative rehabilitation after TBI is using formal assessment of speech and language to provide an outline for the intervention. When using standard aphasia tests where communication is usually not assessed, for instance in The Western Aphasia Battery - WAB (Kertesz, 1982) up to 30% or 40 % of the patients with TBI will show signs of impaired speech and language skills. These difficulties can consist of anomia expressed in impaired confrontation naming, word-finding, verbal association and comprehension (Ahlsén 2006). However, a conventional investigation of language competence based on phonological, syntactical and semantic skills fails to detect the problems in communication experienced by many individuals (McDonald 2000). Communication impairment after TBI is related to reduced language ability in some cases, like verb retrieval deficits in Broca's aphasia, but it seems that the majority of cases depend on more general cognitive difficulties. Researchers have found problems in the following areas: verbal learning and memory, discourse, meta-linguistic tasks, abstract and indirect language, complex lexical-semantic and morphosyntactic manipulation, theory of mind, social communication, and behavioural self-regulation (Ylvisaker and Feeney 2007).

The impact of the cognitive load in a home or school environment may expose difficulties that were just hinted in the clinical setup. A key limitation in clinical assessments is that tests of language functions tend to focus on the impairment perspective, failing to define the consequences of these deficits on functional communication skills (LaPointe et al., 2010). Standardized tests may be ''functional'', in the sense that they assess daily functioning, but because of the fact that the administration is standardized, the tests are always limited when it comes to describing the full potential of an individual's communication life (Fyrberg et al., 2007).

Other approaches can address these types of problems more adequately as has been more frequently discussed by researchers in the last two decades. A step away from traditional clinical assessments towards a description of the individual's communication in his/hers own environment may present the best context to understand and rehabilitate communication skills. Applying a social rather than a medical model requires a shift in perspective and in promoting social communication within natural contexts (Simmons-Mackie 2000). This "contextualized observation" is motivated by the fact that subjects with TBI often perform surprisingly better or worse in everyday contexts than can be predicted from standardized test performance (Ylvisaker et al., 2002).

Cognitive ethnography research combines traditional long-term participant observation with the micro-analysis of specific occurrences of events and practices in real life (Alač and Hutchins 2004). Conversation analysis focuses on microanalysis (Atkinson and Heritage 1984) and has been used by researchers to interconnect the data obtained in communication in social contexts with scores on formal language tests (Friedland and Miller 1998). To investigate the details of interaction in dialogues, such as "choice" or "change" functions in communicated messages, a protocol for Communication Management was developed by Allwood et al. (2007). The protocol looks at phenomena such as body gestures, hesitation and self-interruption and their role as "choice" or "change" mediators of an intended message.

The present study adopted the model of Communication Management to explore its relevance in the rehabilitation process of a young person with TBI**.**

1.2 Strategies of multimodality in communication after TBI

Face-to-face communication is multimodal which is important for the ability to participate in and to manage interaction after TBI. For example, intentional movements of arms, hands and head are used to convey a message; facial

expressions, eye gaze, sounds and body postures are other channels for a subject with a communication disorder to get a message through. Verbal statements can also be illustrated by role-playing. Multimodal communication can comprise prosodic features, pauses, sounds, silences and fragmentary responses in a dialogue and regulates interaction patterns such as turn-taking, feedback and communicative sequencing. Hence, different aspects of multimodality in communication are a focal point when it comes to creating content in a face-to-face interaction (Ahlsén 2003).

Three main components have been described that interact to convey a message in communicative situations: Firstly, factual information is mediated or co-constructed. Secondly, own communication and interaction is regulated and thirdly, emotions and attitudes are communicated (Goodwin 2006). This three-fold content is expressed with different degrees of conscious control and intentionality. On the one hand, the modality that is used to convey a message can require a rather high degree of control, such as in most word-production. On the other hand, a greater proportion of facial expressions, hand gestures and body movements are considered to be mobilized more automatically.

The type of information appears to influence the degree of control, in the sense that more of factual information seems to be produced with a greater degree of control and intentionality than most of the regulation of the speaker's own communication and emotions and attitudes (Ahlsén 2006). This implies that the cognitive effort is highly focused on conveying the linguistic part of the message and that the manner of speech, language, face expression and gestures are adapted to the main message on a more intuitive level in most informal face-to-face interactions.

The type of sign applied in information sharing will also demand a variation in controllability. Peirce's (1998) description of the triadic relations between the signs icon, index and symbol can further explain some of the multimodal communication patterns.

In a conversation, we typically "symbol-ically express" factual information while our hands "iconically illustrate" the same thing and our voice and face expressions "indexically" display our opinion of the topic we are speaking about or the person we are speaking to (Allwood 2002). This complex pattern puts high demands on a person with TBI since impaired cognitive functions will strongly influence the ability to make use of multimodality.

1.3 Communication management

In the model for Communication Management (CM), the planning of Own Communication Management (OCM) is considered a basic feature in face-to-face interaction. OCM represents a speaker's planning and implementation of an intended message in a dialogue. OCM has also been described in terms of hesitation, planning, disfluency, self-correction, editing and self-repair (Allwood et al., 1990). Another type of communicative mechanism is Interactive Communication Management (ICM), aiming at managing the interaction between interlocutors through systems for turn-taking, feedback and sequencing. To succeed in a dialogue, the speaker will need to plan what to say, as well as when to say it, and he or she will also need to continuously moderate the message depending on the response from other speakers. Consequently, OCM and ICM are closely tied, and in a continuous interactive process with the Main Message (MM). The overall purpose is to share main messages with other speakers and to make communication as smooth and fluent as possible (Figure 1).

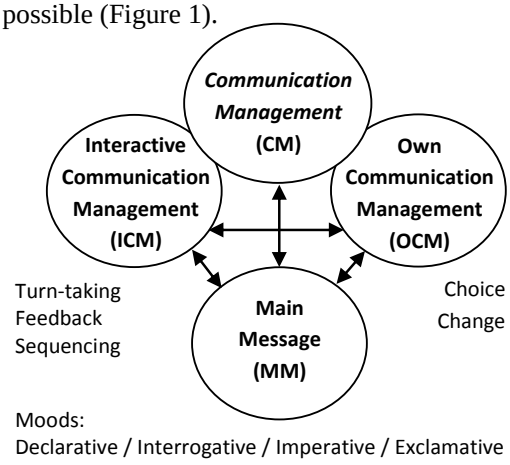

Figure 1. Main functions of Communication. (After Allwood et al., 2007)

Two main features are expressed in OCM - "choice" and "change". Firstly, "choice" phenomena will give the speaker enough time to administer the continuous planning of own content and expression in communication.

Choice can be expressed as tentative word-finding, memory retrieval, hesitation, planning a narrative and keeping the floor. Secondly, "change" features will allow the speaker to alter previously produced content and expressions on the basis of different feedback mechanisms, for

instance by auditive feedback from oneself or from the interlocutor. A change OCM can involve self-repetition and prosodic and/or gestural expressions. However, as Allwood et al. (2007) found in their study of 100 instances of speech based OCM's in informal conversations, OCM functions are often integrated with ICM and MM.

Analyses of gesticulation have been discussed as a method to explore multimodality functions in live communication of persons with aphasia (de Ruiter 2006). In the present study, analyses of OCM and ICM were chosen as methodology to describe multimodal communication in youth with TBI. The study of multimodality in the communication after TBI is a fairly new research area and, to the best of our knowledge, this tool has not been used with adolescents with TBI.

One of the aims of this study was therefore to examine if multimodal aspects supplement formal assessments to create a more comprehensive description of the subject's communication and contribute to identify strategies and goals in the rehabilitation process.

2 Method

2.1 Subject

The participant was a 16-year old male (PJ) who was found unconscious after a downhill skiing accident. In the medical reports, Loss of Consciousness (LOC) was estimated to approximately 30 minutes, and the Glasgow Coma Score was 9. Post-traumatic amnesia (PTA) prevailed for 2 days. Magnetic Resonance Imaging (MRI) findings of the brain revealed scattered subcortical contusions as well as haemorrhages in the frontal, midbrain and temporal left cerebral regions. There was also evidence of grade 1 DAI injuries (Diffuse Axonal Injury) in the left frontal lobe, indicating a degeneration of white matter in this area. Subsequently PJ was diagnosed with a moderate TBI.

During initial hospitalization PJ regained many of his previous abilities. He appeared to have a fairly relevant self-awareness. Gross and fine motor functions were assessed as intact, apart from a pain and stiffness of the neck. Neuropsychological findings showed normal functions in isolated tasks carried out in quiet surroundings, except for verbal memory capacity.

The speech language report identified adequate language abilities, when performed without time pressure and at a limited level of abstraction. However, PJ had explicit difficulties to focus his attention to spoken messages and consequently had problems storing the heard information. This resulted in a limited language comprehension in communicative situations, despite age adequate results in single tests.

After discharge from the hospital and acute-care settings he was sent home. Four months later, PJ was again referred to a clinic, this time a rehabilitation centre, after failing to cope with his home and school environment. The medical referral indicates that he exhibited extensive symptoms of anxiety but had declined counselling. Major obstacles when it came to functioning in his previous academic setting concerned initiating, structuring and planning activities and a tiredness that prevented him from participating in class-room activities as before.

During the subsequent 10 month rehabilitation period, data concerning daily communication functioning in the home and school environment was obtained. The report revealed impaired naming, word-finding, verbal memory and a delay in constructing meaningful messages in a conversation.

2.2 Procedure

Multimodal Communication Management measures: Two live conversations were recorded on videotape at the conclusion of the treatment period. Both recordings involved an unstructured dialogue between the subject and one or two interlocutors (Figure 2 and 3). None of the participants had met previously.

The instruction for the bi-partite conversation (figure 2) was to talk freely for 10 minutes in a "first acquaintance conversation".

In the tri-partite talk (figure 3), the participants agreed on a common topic of conversation, "travelling", for an informal talk. Subsequently, an investigation of communication functions in the two videotapes was made according to multimodal communication analyses (Allwood et al., 2007). The choice of analysed aspects was made according to two main functions: Own Communication Management (OCM) and Interactive Communication management (ICM).

In these two contexts, hand gestures, gaze and head movements as well as smiles and non-verbal sounds perceived as communicative were registered. Articulated words or sentences in conjunction with the gesture were also accounted for. The relations between the vocal-verbal and the gestural production were explored.

Findings in patterns for turn-taking and questioning were linked to PJ's self-confrontation and evaluation of the video-recordings.

Figure 2. The bi-partite conversation. Faces are blurred to secure anonymity.

Figure 3. The tri-partite conversation. Faces are blurred to secure anonymity.

Self-confrontation of live conversations in video transcripts was applied to clarify the interplay between the vocal and the gestural modalities. **The CETI:** The Communicative Effectiveness Index (Lomas et al. 1989) was originally developed for persons with stroke. It is a 16-item questionnaire for estimation of functional communication based on daily communicative functions. Examples of described functions are: "Getting somebody's attention", "Having a one-to-one conversation" and "Being part of a conversation when it is fast and there are a number of people".

For the purpose of this study it was translated into Swedish and used for evaluation of communication by PJ as well as his parents.

Formal tests: Traditional neuropsychological assessments as well as a speech language evaluation were made in clinical surroundings at the beginning of the treatment period.

2.3 Ethical considerations

The study was approved by the Regional Ethical Review Board.

3 Results

3.1 Multimodal Communication Management Results

The outcome of the two live conversations was very different concerning PJ's vocal-verbal participation. In the first conversation with one interlocutor, he contributed substantially more to the conversation than in the second talk with two speakers.

His role talking to two people was more as a listener than an interlocutor. Tables 1 and 2 show the occurrences of interrupted turn-taking, completed turn-taking and instances of asked questions for PJ and the interlocutors in the bi-partite and the tri-partite conversations.

The attempts to initiate turn-taking in the tri-partite conversation were trouble spots since they were delayed and consequently ignored by the other speakers who had already moved on to a new topic. The overall impression was that the other participants interacted partly as interviewers and that PJ was excluded from the turn-taking as the tempo was perceived higher and he had difficulties keeping up with the turns.

Interrupted turn-taking		**Completed turn-taking**		**Asking questions**	
Other speaker	*PJ*	*Other speaker*	*PJ*	*Other speaker*	*PJ*
3	15	15	10	34	9

Table 1. Frequency of turn-taking and questioning in the bi-partite conversation N=2

Interrupted turn-taking		**Completed turn-taking**		**Asking questions**	
Other speaker	*PJ*	*Other speaker*	*PJ*	*Other speaker*	*PJ*
–	3	17	–	19	1

Table 2. Frequency of turn-taking and questioning in the tri-partite conversation N=3

However, PJ had better chances of taking initiative when talking to one person, due to a slower speech rate in the conversation and less competition for the turn.

His turns were longer and more elaborated compared to in the three party conversations where his contributions consisted of mainly one sentence utterances. The phrases were essentially answers to asked question from one of the other participants in the three party talks and not results of PJ's own turn-taking initiative. Hand gestures were frequently used as OCM in the bi-partite conversation (Example 1 and Table 3).

Example 1. The interaction of OCM, hand gesture and gaze in an utterance (// signifies a prolonged silent pause).

Speaker PJ: *// Silverringen // äum.. de..e.. // en lägenhet där //*

(// The Silver Ring // ehum.. it.. is..// an apartment there //)

Speech	//	The Silver Ring //	ehum	it	is //	apartment
Type	silence	noun	OCM word	pronoun	adjective	noun
Gesture	Palm down, fingers spread circular, illustrating a ring. Gaze at interlocutor (IL).	Palm still down, fingers spread circular, illustrating a ring. Gaze at IL.	Fingers collected, index finger pointing down. Gaze at IL.	Continued hand movement with index finger making a circular movement Gaze to side.	Hand and fingers collected. Gaze still to side.	Hand closed. Gaze at IL.
Duration	2 secs	2 secs	3 secs			3 secs

Table 3. The interaction of OCM, hand gestures and gaze in an utterance.

In the above example, PJ answers the question "Where do you live?" and the hand gesture is accompanied by gaze direction at interlocutor. The gestures occur before the elicited content-bearing word and appear to serve the main purpose to trigger word-finding.

The duration of the interval by which the stroke of the gesture preceded the target word corresponded to the duration of the gestures. The gestures continued after the onset of articulation of the lexical affiliate.

This touches on the findings in a previous study by Morrel-Samuels and Krauss (1992) that showed how gestures help speakers access and retrieve lexical items from their mental lexicon. The researchers found that the less familiarity can be assumed in the lexical affiliate, the greater the interval by which the gesture precedes it.

Furthermore, the familiarity of the lexical affiliate was also related to the gesture's duration: the less familiar, the longer the duration of the associated gesture.

In the case of PJ' performance, one might argue that his impaired naming and word-finding as well as a reduction of verbal processing speed and verbal memory creates a similar condition, where verbal functions appear elusive and unfamiliar and require prolonged time to emerge in live conversations. In the research area of expressive gesture abilities in individuals with aphasia, persons with Broca's aphasia were found to be slow to initiate movement, have long pauses but also to have frequent use of iconic gestures and beats (Duffy et al., 1984). The speech related to Broca's aphasia is characterized by a slow and effortful articulation with no significant disturbance in language function. The condition resembles the expressive language difficulties experienced by PJ, as well as the site of the lesion in his left frontal lobe which is similar to the neurological basis for Broca's aphasia.

PJ used gestures to manage the communication. However, the number and the nature of the gestures varied with the number of completed turn-takes. When talking to two persons, PJ used no gestures of the hand to manage the conversation.

Instead he closed his eyes and smiled while struggling with word-production in the one case of gestural OCM (Table 5). During a major part of the conversation, he acted as a listener to the other two speakers and held his hands clasped in front of him, at the sides of the body or the arms held behind his back. However, smiling and using other ICM strategies to demonstrate

participation were frequent and adequate, despite a partial absence of own verbal contributions (JP's comparison and analyses of the dialogues are reported in the end of this section). Tables 4 and 5 contain PJ's distribution of different gesture types in OCM and ICM in the two conversations.

Gesture	OCM	ICM
Hand gesture	**8**	–
Gaze down	**4**	**1**
Head shake	**1**	–
Gaze up	**2**	**1**
Gaze to side	**8**	**3**
Head nod	–	**1**
Smile	**3**	**24**
Non-verbal sounds	**7**	**30**

Table 4. **The bi-partite conversation:** PJ's production of gestures in OCM and ICM

Gesture	OCM	ICM
Closed eyes	**1**	–
Smile	**1**	**15**
Non-verbal sounds	–	**32**

Table 5. **The tri-partite conversation:** PJ's production of gestures in OCM and ICM

In the bi-partite conversation, gestures for word-finding describing spatial location and action were used in eight cases of completed turn-taking. This was clearly expressions of OCM performed at a lower pace when PJ talked only to one person. In this situation, he had enough time to use the gesture during silent pauses to trigger a delayed word-finding during his turn (Example 2 and Table 6).

Example 2. The interaction of OCM, hand gesture and gaze in an utterance (// signifies a prolonged silent pause).

Speaker PJ:
eller jag börjar om... um um // två veckor

(that is, I start in… ehum.ehum // two weeks)

PJ's qualitative description of the video-recordings confirmed a clear discrepancy in the experience of communication management depending on the number of speakers involved. Speech rate in one of the interlocutors in the tri-partite conversation was perceived as high by PJ which further limited his overall language comprehension of the dialogue.

Speech	ehum..ehum	//	two weeks
Type	OCM word	silence	noun phrase
Gesture	Gaze to side.	Lifted collected hand, index finger making two circular movements. Gaze to side.	Gaze at IL.
Duration	3 secs		

Table 6. The interaction of OCM, hand gestures and gaze in an utterance.

Comprehension was also reduced by unknown topics in the talk. Turn-taking and initiative was managed more easily in the bi-partite dialogue as the speech rate of the interlocutor was lower here. PJ did not want to laugh so much during the conversations and had wanted to use his hands more for gesturing. When unable to understand, he did not ask for a clarification. His overall feeling was that new people do not regard him as being serious and avoid his eye-contact. This statement was, however, not confirmed in the analyses of the video-recordings.

3.2 The CETI results

The ratings on the CETI made by the parents and PJ occurred at the beginning of the rehabilitation period, six months post trauma. Repeat test scores were recorded 10 months later, at the closure of the period. In both the initial test score as well as in the repeat score, PJ evaluated his own communicative ability *"as able as before the brain injury" (*score = 100) in a total of 9 communicative situations. Four of these 100 % items were rated before onset of the treatment period, and a further 5 items were registered at follow-up. Apparently PJ experienced 4 communicative functions as being completely unaffected by the trauma and additional 5 functions as being recovered to present status at the end of the treatment period.

The parents, however, did not on any given occasion perceive their sons communication as *"as able as before the brain injury",* a fact that was mirrored in their estimations. Their highest points of registration were between 75 % and 98 % (12 items) with six of these ratings occurring before the treatment period and six items after.

However, in these ratings, there are two items, 11 and 13, "Starting a conversation with people who are not close family" and "Understanding writing" that indicate a major change over time, of a 50 (51) % improved capacity. For the other ratings, there is no major change of performance registered.

3.3 Formal test results

The subject performed seemingly well on all tests in the WAIS-III. Full Scale IQ-results of 101 indicate an average cognitive level of functioning. However, a discrepancy of 25 IQ points between the verbal and visual domains, with limitations in verbal functions, was apparent.

4 Discussion

In the videotaped interactions of Multimodal Communication Management, PJ was the more passive vocal-verbal interlocutor in both dialogues. However, he managed to interact using multimodal expressions, a great proportion of all instances of communication management was judged to be expressions of ICM, thereby upholding the interaction non-verbally. In the case of OCM, hand gestures and gaze down were the most frequent gestures. This is consistent with the findings by Allwood et al. (2007) in their study of Communication Management in conversations between healthy subjects.

Furthermore, gestures preceded the affiliated word in most cases and the delay between gesture and target word was 2-4 seconds. Morrell-Samuels and Krauss (1992) found that the onset of gestures usually precedes the target word. The researchers also found that the less familiar a word is, the larger is the time interval by which the gesture that precedes the speech. This might explain the interval between PJ's gesturing and naming in the conversations, since delayed and tangential word-finding as well as verbal memory limitations were trouble spots after the trauma.

During the rehabilitation period PJ elaborated the use of multimodal cues to participate in conversations, despite persisting problems with verbal comprehension and expressions. From a communication treatment perspective, the cognitive functioning of the adolescent allowed a development of insights in the possibilities and obstacles in communicative situations. By using gaze, smiles and postural techniques, he was able to participate as a teammate even in the instances of reduced language comprehension. To appreciate the role as a listener and the importance of this stance in the joint creation of meaning-making proved an important technique to uphold a conversation and, above all, to save face in moments of comprehension difficulties.

The results on the CETI are consistent with previous findings in investigations of Health Related Quality of Life (HRQL) after TBI (Stancin et al., 2002). Specifically, the researchers found that parents rated their children's HRQL less favourably than the young person did themselves. This implies that adolescents might be inclined to underestimate the impact of their own health and functioning and hence report higher HRQL compared to their parents.

Conventional MR images are poor predictors of functional outcome in patients with TBI. However, as in the case with the adolescent in this study, neurological findings helped explain some of the core deficits underlying the difficulties experienced after a brain injury. The DAI-lesions in the left frontal lobe of PJ reflected a slower rate of processing speed and initiative. Damage to the left temporal lobe corresponded to the word-finding problems, the reduced processing of auditory input and to the verbal memory limitations. In functional communication, this may have affected the impaired language comprehension ability, as was documented in the video-taped conversations.

5 Conclusions

The results in this study support the notion that a triangulation of methods is a fruitful approach to investigate and treat consequences of communication impairment after TBI. Future research should include trials in more persons with TBI, and an extension of the method to compare more recorded situations of multimodal communication management during the rehabilitation period.

References

Ahlsén, E. 2003. Communicative contributions and multimodality – Reflections on body communication, communication aids for persons with language disorders and dialogue systems. In *Proceedings of the 1st Nordic Symposium on Multimodal Communication*.

Ahlsén, M. 2006. *Introduction to Neurolinguistics*. Amsterdam: John Benjamins.

Alač, M., and E. Hutchins. 2004. I See What You Are Saying: Action as Cognition in fMRI Brain Mapping Practice. *Journal of Cognition and Culture* 4:629-661.

Allwood, J. 2002. Bodily communication dimensions of expression and content. In *Multimodality in language and speech systems*. Boston: Kluwer Academic.

Allwood, J., E. Ahlsén, J. Lund, and J. Sundqvist. 2007. Multimodality in own communication management. In *Current Trends in Research on Spoken Language in the Nordic Countries*. Oulu: Oulu University Press.

Allwood, J., Nivre, J. and Ahlsén, E. 1990. Speech Management—on the Non-written Life of Speech. *Nordic Journal of Linguistics* 13 (1):3-48.

ASHA - American Speech-Language-Hearing-Association. 2005. Roles of Speech-Language Pathologists in the Identification, Diagnosis, and Treatment of Individuals With Cognitive-Communication Disorders. [Position Statement].

Atkinson, J. Maxwell, and John Heritage. 1984. *Structures of social action : studies in conversation analysis*, *Studies in emotion and social interaction*,. Cambridge: Cambridge Univ. Press.

Chamberlain, D. J. 2006. The experience of surviving traumatic brain injury. *J Adv Nurs* 54 (4):407-17.

de Ruiter, Jan Peter. 2006. Can gesticulation help aphasic people speak, or rather, communicate? *International Journal of Speech-Language Pathology* 8 (2):124-127.

Duffy, Robert J., Joseph R. Duffy, and Patricia A. Mercaitis. 1984. Comparison of the performances of a fluent and a nonfluent aphasic on a pantomimic referential task. *Brain and Language* 21 (2):260-273.

Friedland, D., and N. Miller. 1998. Conversation analysis of communication breakdown after closed head injury. *Brain Inj* 12 (1):1-14.

Fyrberg, A., M. Marchioni, and I. Emanuelson. 2007. Severe acquired brain injury: rehabilitation of communicative skills in children and adolescents. *Int J Rehabil Res* 30 (2):153-7.

Goodwin, C. 2006. Human sociality as mutual orientation in a rich interactive environment: multimodal utterances and pointing in aphasia. In *Roots of human sociality: culture, cognition and interaction*. Oxford: Berg.

Hux, Karen, Erin Bush, Samantha Zickefoose, Michelle Holmberg, Ambyr Henderson, and Gina Simanek. 2010. Exploring the study skills and accommodations used by college student survivors of traumatic brain injury. *Brain Injury* 24 (1):13-26.

Kertesz, A. 1982. *Western Aphasia Battery test manual*: Grune & Stratton.

LaPointe, Leonard L., B. E. Murdoch, and Julie A. G. Stierwalt. 2010. *Brain-based communication disorders*. San Diego: Plural Pub.

Lomas, J., L. Pickard, S. Bester, H. Elbard, A. Finlayson, and C. Zoghaib. 1989. The communicative effectiveness index: development and psychometric evaluation of a functional communication measure for adult aphasia. *J Speech Hear Disord* 54 (1):113-24.

McDonald, Skye. 2000. Editorial Putting communication disorders in context after traumatic brain injury. *Aphasiology* 14 (4):339-347.

Morrel-Samuels, Palmer, and Robert M. Krauss. 1992. Word Familiarity Predicts Temporal Asynchrony of Hand Gestures and Speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 18 (3):615-615-622.

Peirce, Charles S. 1998. What Is a Sign? In *The essential Peirce : selected philosophical writings. Vol. 2, 1893-1913*, edited by C. J. W. Kloesel and N. Houser. Bloomington: Indiana University Press.

Roscigno, C. I., K. M. Swanson, M. S. Vavilala, and J. Solchany. 2011. Children's longing for everydayness: Life following traumatic brain injury in the USA. *Brain Inj* 25 (9):882-94.

Sarno, M. T. 1980. The nature of verbal impairment after closed head injury. *J Nerv Ment Dis* 168 (11):685-92.

Simmons-Mackie, N. 2000. Social approaches to the management of aphasia. In *Neurogenic communication disorders : a functional approach*, edited by L. E. Worrall and C. M. Frattali. New York ; Stuttgart: Thieme.

Stancin, T., D. Drotar, H. G. Taylor, K. O. Yeates, S. L. Wade, and N. M. Minich. 2002. Health-related quality of life of children and adolescents after traumatic brain injury. *Pediatrics* 109 (2):E34.

Wahlström Rodling, M., M. Olivecrona, L-O. D. Koskinen, B. Rydenhag, and S. Naredi. 2005. Severe traumatic brain injury in pediatric patients: treatment and outcome using an intracranial pressure targeted therapy- the Lund concept *Intensive Care Med* 31:832-839.

Ylvisaker, M., and T. Feeney. 2007. Pediatric brain injury: social, behavioral, and communication disability. *Phys Med Rehabil Clin N Am* 18 (1):133-44, vii.

Ylvisaker, M., R. Hanks, and D. Johnson-Greene. 2002. Perspectives on rehabilitation of individuals with cognitive impairment after brain injury: rationale for reconsideration of theoretical paradigms. *J Head Trauma Rehabil* 17 (3):191-209.

Author Index