NEALT Proceedings

Northern European Association for Language Technology

Editors:

Jens Allwood
Elisabeth Ahlsén
Patrizia Paggio
Costanza Navarretta
Kristiina Jokinen

mmc 4th Nordic Symposium on Multimodal Communication 15th - 16th November 2012

SKI SKI

CHALMERS | UNIVERSITY OF GOTHENBURG

Proceedings of the
4th Nordic Symposium on Multimodal Communication

NMMC 2012

November 15-16, 2012 • Gothenburg, Sweden

Linköping Electronic Conference Proceedings

NEALT PROCEEDINGS SERIES

Vol. 21

Proceedings of the 4th Nordic Symposium on Multimodal Communication

November 15-16, 2012
University of Gothenburg
Sweden

Editors

Jens Allwood
Elisabeth Ahlsén
Patrizia Paggio
Costanza Navarretta
Kristiina Jokinen

NORTHERN EUROPEAN ASSOCIATION
FOR LANGUAGE TECHNOLOGY

Linköping Electronic Conference Proceedings, No. 93
Linköping University Electronic Press
Linköping, Sweden, 2013
ISSN: 1650-3686 (print)
ISSN: 1650-3740 (online)
ISBN: 978-91-7519-461-5
URL: www.ep.liu.se/ecp_home/index.en.aspx?issue=093

NEALT Proceedings Series, Vol. 21
ISSN 1736-6305 (Online)
ISSN 1736-8197 (Print)

Program Committee

Elisabeth Ahlsén, University of Gothenburg, Sweden

Jens Allwood, University of Gothenburg, Sweden

Patrizia Paggio, Copenhagen University, Denmark and University of Malta

Costanza Navarretta, Copenhagen University, Denmark

Kristiina Jokinen, University of Helsinki, Finland

Organizing Committee

Elisabeth Ahlsén

Jens Allwood

Stefano Lanzini

Pavel Rodin

Reviewers

Nick Campbell, Ireland

Jens Edlund, Sweden

Joakim Gustafsson, Sweden

Pentti Haddington, Finland

Bart Jongejan, Denmark

Arne Jönsson, Sweden

Matthias Rehm, Denmark

Alexandra Welenman, Sweden

Contents

Preface 1

Multimodal turn management in Danish dyadic first encounters
Costanza Navarretta and Patrizia Paggio 5

What's in a gesture?
Elisabeth Ahlsén and Jens Allwood 13

Overlaps in Maltese: a comparison between map task dialogues and multimodal conversational data
Alexandra Vella and Patrizia Paggio 21

Bipedics: Towards a next category of kinesics. An empirical investigation of attitude and emotion through simple leg and foot gesture
Peter O'Reilly 31

Multimodal communication in intercultural health care interactions
Elisabeth Ahlsén and Nataliya Berbyuk Lindström 39

Predicting the attitude flow in dialogue based on multi-modal speech cues
Peter Juel Henrichsen and Jens Allwood 47

Multimodal feedback expressions in Danish and Polish spontaneous conversations
Costanza Navarretta and Magdalena Lis 55

Showing interest during first acquaintance
Gülüzar Tuna, Jens Allwood and Elisabeth Ahlsén 63

Prosodic expressions of emotions and attitudes in communicative feedback
Gustaf Lindblad and Jens Allwood 71

Vagueness and Gesture
Jens Allwood and Evelyn Vilkman 79

Annotating attitudes in the Danish NOMCO corpus of first encounters
Anette Luff Studsgård and Costanza Navarretta 85

Attitudinal emotions and head movements in Danish first acquaintance conversations
Bjørn Nicola Wessel-Tolvig and Patrizia Paggio 91

Preface

The articles collected in this volume are a selection of papers presented at the 4th Nordic Symposium on Multimodal Communication that was held at the University of Gothenburg on 15-16 November 2012. The symposium was organized by the SCCIIL Interdisciplinary Research Center and the Division of Communication and Cognition at the Department of Applied IT at the University of Gothenburg and the NOMCO (Multimodal Corpora for the Nordic Languages) NORDCORP project. The symposium was supported by the Swedish Research Council (VR) and FORTE.

The symposium continues a tradition, established by the Swedish Symposium on Multimodal Communication, held from 1997 to 2000, and then continued by the two Nordic Symposia on Multimodal Communication held in 2003 and 2005, the workshop held at NODALIDA in Odense in 2009 and the Third Symposium on Multimodal Communication in Helsinki 2011.

Several studies based on and related to the NOMCO project were presented at the symposium and appear in this volume, most of them based on the Nordic corpora of First acquaintance interactions. Other studies on multimodal communication in the volume deal with other data and other aspects of multimodal communication. Studies on a number of topics and languages are represented in the volume. Two main topics of the papers are multimodal communication in relation to interaction regulation, e.g. turn management and communicative feedback, and multimodal communication related to emotions and attitudes.

The symposium received about 50 submissions from a number of countries in Europe, Asia, Africa and the US , 23 of which were accepted for oral or poster presentation in Gothenburg, after each being reviewed by two members of the Panel of reviewers and the Program committee. Four invited keynote speakers, Karl Grammer, Dirk Heylen, Daniel Västfjäll and Michael Kipp, were also included in the symposium program. The 12 papers included in these Postproceedings were submitted after the symposium and each paper was was revised after having been reviewed by the Program Committee and two members of the Panel of reviewers..

The first paper, by Navarretta and Paggio, studies turn management in Danish first encounters. It also includes measuring the length of contributions and one of the findings is that both male and female participants speak more when the conversation partner is female.

Ahlsén and Allwood report a perception experiment where the subjects are asked to identify the word most likely coproduced with a unimodally presented mimicked gesture. Some of the results were that gestures that were originally produced by persons with aphasia were as easy to interpret as those produced by persons without aphasia and that subjects with a different cultural-linguistic background from that of the person originally producing the gesture interpreted the gestures almost as easily as subjects with a similar background

The focus in the paper by Vella and Paggio is on overlaps in Maltese. The study includes map task dialogues and studies of face-to-face conversations and shows differences in the frequency and function of overlaps between the two conditions, where overlaps are used to achieve optimal information exchange in the Map Task dialogs, while they are a sign of ease and familiarity in free conversations.

O'Reilly introduces a study based on a corpus and interviews, as well as on an experimental empirical study of what he calls "Bipedics", i.e. communication through leg and foot gesture. This is one of several contributions on multimodal communication of emotion and attitudes. The results of both O'Reilly's studies support a link between certain bipedic gestures and the expression of attitudes and emotions.

The study by Ahlsén and Berbyuk Lindström takes the different phases of a particular social activity – the intercultural health care interaction – as its point of departure and describes the different types and functions of multimodal communication in these phases. Functions of multimodality, for example in enhancing comprehension and establishing rapport, are exemplified and discussed.

Henrichsen and Allwood have developed an approach to automatic prediction of attitude flow in dialog from multimodal speech cues and report on the results of using this approach in a first experiment. Their results include a recommended set of analytical annotation labels and a recommended setup for extracting linguistically meaningful data even from noisy audio and video signals.

A comparative study of Danish and Polish multimodal feedback in interaction is provided by Navarretta and Lis. Even though the same types of head movements and vocal expressions are used in the two languages, the Polish data contain more multimodal feedback in general and more repeated multimodal feedback. A relation between familiarity and repetitive feedback is also found.

Tuna, Allwood and Ahlsén attempt to capture the multimodal cues showing the attitude of interest in first acquaintance dialogs. Multimodal expressions connected with showing interest mainly include types of five body movements/gestures; gaze, head movements, holistic face, hand movements and body postures. The expression

of interest only and interest in combination with other affective-epistemic states is also analyzed.

Another study of emotions and attitudes, by Lindblad and Allwood, focuses on measuring their expression through prosodic features. The study uses studio recorded feedback words read with the intention to express different affective-epistemic states (AES) for a perception experiment as well as an acoustic analysis of these features. The results varies, i.e. there is more agreement among the subjects on some AES than others, which might reflect different degrees of dependence on prosodic cues in relation to other expressive features. The method shows promise for further studies.

The paper by Allwood and Vilkman focuses on how vagueness, unspecificity, approximation, uncertainty and hesitation (VUAUH phenomena) are reflected in word and gestures in a corpus of political debates. Some VUAUH types seem to be connected to certain gestures, e.g. approximation to head waggle.

Capturing multimodal expressions of emotions and attitudes in videorecorded interactions requires coding and/or automatic analysis and different methods have been used in the studies reported in this volume. Luff-Studsgård and Anderson contribute a methodology paper on annotating attitudes, where they suggest annotation in the PAD (pleasure, arousal, dominance) space in combination with annotation labels, in order to reach a higher intercoder agreement.

The last paper, by Wessel-Tolvig and Paggio, also deals with attitudinal emotions, focusing on head movements in Danish first encounters and their relation to reported attitudinal emotions in post-experiment questionnaires. The findings suggest a slight positive correlation between number of head nods and a more positive attitude.

On behalf of the organizing committee

Jens Allwood, Elisabeth Ahlsén

Multimodal Turn Management in Danish Dyadic First Encounters

Costanza Navarretta
University of Copenhagen
Copenhagen, Denmark
costanza@hum.ku.dk

Patrizia Paggio
University of Copenhagen and
University of Malta
Valletta,Malta
paggio@hum.ku.dk

Abstract

This paper presents studies of multimodal turn management in a Danish corpus of video recorded conversations between two young people who meet for the first time. More specifically, we investigate multimodal behaviours by which the conversation participants indicate whether they wish to give, take or keep the turn. In this study we present quantitative analyses of such cues, as well as an investigation of the length of the individual participants' speech contributions. The quantitative studies comprise body behaviours which have not been previously investigated with respect to turn management, so that it not only confirms preceding studies on turn management in English but also provides new insight on how speech and body behaviours are used synchronously in communication. The investigation of the participants' vocal contributions shows gender differences in that male participants talked more when interacting with a female than they did when their interlocutor was a male, while female participants talked more when interacting with a female than when they interacted with a male.

Keywords: multimodal communication, multimodal corpora, turn management, annotation

1 Introduction

Human-human communication is multimodal by nature because people communicate through both speech and non-verbal behaviours. This article is about speech and non-verbal[1] cues with a turn management function in the Danish NOMCO first encounters corpus (Paggio and Navarretta, 2011).

Turn management is the process by which conversation participants regulate the interaction flow (Allwood et al., 2007). This is done by both verbal and body behavioural cues (Kendon, 1967; Yngve, 1970; Ford and Thompson, 1996; Duncan, 1972; Allwood et al., 2007; Hadar et al., 1984).

Sacks et al. (1974) propose pre-defined *turn-taking* rules to model the way in which the participants regulate their turn flow smoothly, in other words avoiding too long pauses and speech overlaps. Schegloff (2000) adds to the turn-taking system an overlap-management system in order to account for the many overlaps which have been found in real life conversations, see inter alia (O'Connell et al., 1990; Cowley, 1998). Overlap management is only needed, according to Schegloff, when overlap is problematic, that is when it occurs for longer time.

Several researchers disagree with the view of a pre-defined turn-taking system because turn-taking depends on many factors such as the conversation setting, the cultural environment, the degree of familiarity and the number of the participants (O'Connell et al., 1990; Cowley, 1998; Du-Babcock, 2003; Tanaka, 2008). Furthermore, silence and overlaps, in both speech and bodily behaviours, should be seen as a natural part of conversation, signalling that people communicate in synchrony (Campbell, 2009; Esposito et al., 2010).

In this paper, we present two studies of turn management cues expressed by various body behaviours comprising head movements, gaze, hand movements, facial expressions and body postures in a corpus of dyadic Danish first encounters, and we relate these findings to the literature. Then, we focus on how long the participants's turns are depending on their different interlocutors. The rest of the paper is organised as follows. In section 2 we discuss relevant background literature while in section 3 we describe our corpus and the relevant

[1] In this paper, we use non-verbal in the sense of pertaining to body behaviour rather than speech.

annotations. In section 4 we present our analyses of the data, section 5 contains the discussion, and in section 6 we conclude and present future work.

2 Background Studies

Different studies underline the central role of specific vocal and body behaviours in turn management. For example, Kendon (1967) and Argyle and Cook (1976) investigate the role of gaze direction and of mutual gaze, respectively, while Duncan (1972) annotates speech and body behaviours in dyadic English conversations and identifies verbal and non-verbal cues by which speakers signal that they want to give the turn to the interlocutor (*turn-yielding* in Duncan's terminology). The verbal cues which Duncan defines comprise the following phenomena: a) intonation, b) the use of hedges, such as *you know* and *I guess*, and c) syntax. Only one non-verbal cue indicating that the speaker is terminating the turn is identified: the completion of on-going hand gestures. In Duncan's conversations at least two cues co-occurred when speakers showed that they wanted to release the turn.

Hadar et al. (1984) analyse the occurrences of head movements in conversations between four participants. In these data, they find that linear movements of the head ("postural shifts") often occur after "grammatical" pauses both between clauses and sentences. Furthermore, postural shifts are also identified towards the initiation of speech, both between speaking turns and between syntactic boundaries inside speaking turns. Their conclusion is that head movements are involved in regulating turn taking and marking syntactic boundaries inside speaking turns. They also find that smaller and quicker movements tend to occur after dysfluencies inside grammatical boundaries, especially after short pauses.

Differing from the other researchers, Duncan and Fiske (1977) focus on the behaviours of the listeners. They propose to distinguish backchannelling signals which do not provide new semantic information from regular turns..

In the rest of the paper, we discuss to what extent speech cues and body behaviours are also present in our data in connection to turn management.

3 The Data

Our data is the Danish NOMCO corpus of first encounters[2]. The corpus was collected and annotated under the NOrdic Multimodal COrpora (NOMCO) project. Comparable conversations of first encounters were also collected and annotated in Finnish and Swedish (Paggio et al., 2010; Navarretta et al., 2011; Navarretta et al., 2012).

The Danish corpus consists of 12 dyadic conversations with the length of approximately 5 minutes each. The participants are all young people, university students or university educated, aged from 21 to 36. The participant population comprised 6 females and 6 males who had a common acquaintance, but did not know their interlocutors in advance. Each subject participated in two conversations, one with a female participant and one with a male participant, and the two conversations were recorded on two different days (Paggio and Navarretta, 2011). The participants were instructed to talk in order to get acquainted, as if they met at a party, and they were only told that they participated in a project on Danish.

The interactions were recorded by three cameras at the University of Copenhagen. Frontal views of each subject and a panorama view of the two participants standing in front of each other are available. The three camera views are shown in 1 and 2.

Figure 1: Snapshot from a conversation: frontal camera views

3.1 The annotations

The corpus was orthographically transcribed in PRAAT (Boersma and Weenink, 2009) with time stamps at the word level. Stress and phrasal information are also available. The transcriptions

[2]This section is almost identical to the description of the corpus provided in earlier papers.

Figure 2: Snapshot from a conversation: side camera view

were imported in the multimodal annotation tool ANVIL (Kipp, 2004) and the body behaviours were annotated according to the MUMIN annotation scheme (Allwood et al., 2007). The scheme provides pre-defined features describing the shape and function of gestures and their semiotic type (Peirce, 1931). Since body behaviours are multifunctional, they can be assigned more functions at the same time. Body behaviours which are judged to be semantically related to speech segments produced by the gesturer or the interlocutor can be linked explicitly to these in the annotation (Allwood et al., 2007).

The body behaviours were annotated by a coder and then corrected by a second coder. Disagreement cases were resolved by a senior coder. The analyses discussed in this paper are based on the final concerted version (Paggio and Navarretta, 2011; P.Paggio and Navarretta, 2012).

Navarretta et al. (2011) give an account of intercoder agreement tests on the annotations which, depending on the categories, resulted in kappa scores (Cohen, 1960) between 0.60-0.90.

In this study, we use the annotations of head movements, facial expressions and body postures related to turn management. The features describing the shape of these behaviours are presented in Table 1. Body postures are annotated with information on direction, whether the body is facing the interlocutor, and what the movement of the shoulders is. Facial expressions are described with a general face attribute and an eyebrows attribute. Finally, head movements are described by the form of the movement and an attribute indicating whether the movement is performed one or more times.

The MUMIN scheme distinguishes the following six turn management behaviours:

- TurnTake: signals that the speaker wants to take a turn that wasn't offered, possibly by interrupting
- TurnHold: signals that the speaker wishes to keep the turn
- TurnAccept: signals that the speaker is accepting a turn that is being offered
- TurnYield: signals that the speaker is releasing the turn under pressure
- TurnElicit: signals that the speakers is offering the turn to the interlocutor
- TurnComplete: signals that the speaker has completed the turn.

Shape attribute	**Shape values**
BodyDirection	BodyForward, BodyBackward, BodyUp, BodyDown, BodySide, BodyTurn, BodyDirectionOther
BodyInterlocutor	BodyToInterlocutor, BodyAwayFromInterlocutor
Shoulders	Shrug, ShouldersOther
HeadMovement	Nod, Jerk, HeadForward, HeadBackward,Tilt, SideTurn, Shake, Waggle, HeadOther
HeadRepetition	Single, Repeated
General face	Smile, Laugh, Scowl, FaceOther
Eyebrows	Frown, Raise, BrowsOther

Table 1: Shape Features of Head Movements, Facial Expressions and Body Postures

4 Turn Management in the Corpus

4.1 Turn management and gesture types

The corpus contains 18000 speech tokens comprising filled pauses. Table 2 shows the body behaviours annotated in the corpus, the body behaviours with a turn management function, and their percentage.

Table 3 shows how the turn management associated with body behaviours is distributed across the three different behavioural types.

In table 4, the body behaviours which are most frequently related to a turn management function in this corpus are shown.

Behaviour	Total	Turn	%
Head	3117	738	24
Face	1448	247	17
Body	982	223	23
Total	5547	1208	100

Table 2: Turn management body behaviours

Behaviour	%
Head movements	61
Facial expressions	20.5
Body postures	18.5

Table 3: Turn management distribution across body behaviours

The tables show that all three body parts are related to turn management, and not only head movements, hand gestures and gaze on which preceding studies mainly have focused. The second table also shows that more types of head movement than those indicated in the literature are relevant to turn management in this corpus. Interesting are especially the occurrences of forward and backward movements of the head which have not been related earlier to turn management.

Some of these head movements may be accompanied by movements of the torso and the body. However, we have not considered whole body behaviour here.

The most frequently assigned turn management categories are TurnHold, TurnAccept and TurnElicit, while TurnYield, which in our coding scheme is used to code turn/releasing under pressure, is extremely rare in the corpus. This reflects the type of social activity and the communicative situation: people who meet for the first time are

Turn M. Behaviour	No.
SideTurn	217
HeadForward	127
EyebrowsRaise	126
Tilt	104
Smile Shake	98
HeadBackward	76
Nod	72
BodyTurn	63
	52

Table 4: Most frequent turn management behaviours

Behaviour	Turn M.Type	No.
Head	TurnHold	217
	TurnAccept	202
	TurnElicit	196
	TurnTake	112
	TurnYield	10
Face	TurnElicit	105
	TurnAccept	96
	TurnTake	28
	TurnHold	15
	TurnYield	3
Body	TurnAccept	88
	TurnElicit	60
	TurnHold	37
	TurnTake	34
	TurnYield	4

Table 5: Turn management related types and body behaviours

both friendly and polite and, in Denmark, this also implies not interrupting the interlocutor. Table 4.1 shows the turn management categories which are assigned more frequently to head movement, facial expressions and body postures.

The table indicates that each body behaviour type is mostly related to two or three specific turn management functions: for head movements these are TurnHold, TurnAccept and TurnElicit; for facial expressions they are TurnElicit and TurnAccept; and finally for body postures, they are TurnAccept and TurnElicit.

4.2 Multimodal turn shift cues

In the above analysis we have looked at each body part independently, however in the data several body behaviours often co-occur. A more truly multimodal approach is presented below in an analysis of part of the corpus. More specifically, we investigate in two first encounter conversations whether the turn offering cues observed in English conversations by Duncan (1972) and described earlier, also hold in our corpus. As already noted, Duncan finds that at least one of six cues connected to intonation, speech content and hand gestures occurred in turn eliciting situations. When more than a cue is present, they occur simultaneously or in tight sequences.

In this study, we only include turn alternations without overlapping speech since we want to verify on our data Duncan's results (1972) work. We

also exclude answers to direct questions because the question is an explicit turn eliciting cue. The turn shifts relevant to our study are 35% of the turn shifts in the two conversations. In 95% of these, the speaker concludes a syntactic phrase. The pitch level, on the other hand, only goes down at the end of the phrase in 10% of the cases. The speaker's head and body freeze in all cases, and similarly, the speaker and interlocutor always have eye contact. The participants in the two conversations analysed do not move their hands very frequently. Thus, the speaker finishes talking and moving their hands nearly simultaneously only in three of the turn shifts considered here. The syntactic completion of a phrase and a pitch different from an intermediate pitch level are two of the cues described by Duncan (Duncan, 1972), but we only found occurrences of the former in the two conversations. It can be argued that the freezing of the head and body are parallel to the termination of hand gestures by the speaker which Duncan mentions as the body cue associated with turn-yielding. Furthermore, also in our data the speaker ends the production of speech and co-occurring hand movement nearly simultaneously. As for the use of gaze, the fact that speakers look at their interlocutors to indicate that they want to release the turn confirms a tendency also found in other studies (Kendon, 1967; Argyle and Cook, 1976; Jokinen, 2011).

Duncan (1972) also finds that speakers signal that they are offering the turn to the interlocutor by using hedges such as *I think* and *I guess*.

We have only found few vowel lengthenings on the stressed syllable of a terminal clause in the two conversations, and there are no occurrences in the corpus of hedges in turn eliciting situations. In other words, the presence of hedges does not play a significant role in our corpus.

4.3 Turn length

The length of speakers' turns can be studied from different perspectives. In the third study reported here, we investigate whether there are differences in the amount of speech produced by the participants depending on the gender of the interlocutor. In particular, we look at the number of words uttered by each participant and the length of their turns. The pattern which results from this study is the following. Female participants utter more words and keep the floor for longer time when they talk with another female than they do when they interact with a male. Male participants, on the contrary, talk for longer time and utter more words when their interlocutor is a female than they do when the interlocutor is a male. This is true in all conversations with one exception in which the female participant speaks more than her male interlocutor. However, she also speaks more than her female interlocutor and is in fact the participant who talks more than all the other participants. In sum, she can be considered an outlier in the sample.

5 Discussion

Since more than 20% of the communicative head movements, facial expressions and body postures annotated in the corpus have a turn management function, our study shows that all these behaviours often have a turn management function. This confirms the fact that humans use all their body when they synchronise their contributions.

We also see that different body parts are often related to specific turn management functions, an aspect that ought to be investigated further.

The less frequently occurring turn management functions in the corpus are TurnYield and TurnComplete (the latter only occurs once). The fact that only few occurrences of turn release under pressure are found, reflects the type of social interaction as well as the culture. The participants meet for the first time, thus they want to be kind and friendly, and avoid interrupting each other as a consequence. As for TurnComplete, the phenomenon is more relevant in different communication situations, e.g. interviews, in which the speaker is asked for something and stops talking when the answer to the question is complete.

Many types of head movement are involved in turn management in addition to nods and shakes, which many earlier studies have treated as typical turn shift signals. This finding is similar to the conclusion drawn in another analysis of the same corpus, in which we found that all types of head movements are also used by the participants to give or elicit feedback (Paggio and Navarretta, 2011).

Whilst the role of body cues for turn shift was analysed quantitatively across the whole corpus, two of the conversations were studied with special regard to how various syntactic and prosodic features contribute to turn management. In particular,

the analysis of turn eliciting cues in the two files confirms Duncan's observation that a speaker's completion of a syntactic phrase signals that the speaker wants to relinquish the turn. The presence of hedges, vowel lengthening and pitch change, on the other hand, are either seldom or never found in the data. Our analysis also shows that the speaker can signal that they are offering the turn to the interlocutor not only by finishing off on-going hand gestures, but also by bringing the head and body to a standstill while keeping their gaze on the interlocutor.

The investigation of how long the participants talk shows that five out of six men talk more when interacting with a person of the opposite gender than with one of the same gender. Conversely, female participants tend to be more talkative with an interlocutor of the same gender. Thus, the data indicate that there is a gender difference when males and females participate in first encounters. Whether this is a general tendency or an idiosyncratic characteristic of our corpus should be investigated further.

6 Conclusions and Future Work

In this paper we have presented three studies of turn management behaviours in the Danish NOMCO first encounters corpus. The first study, where we look at the type of body behaviours involved in turn management, confirms preceding studies on turn management, but it also provides new insights. In fact, our data shows that all kinds of head movement, facial expression and body posture are used as turn management cues. We also found a relation between the turn management types which occur most frequently in this corpus and the type of social activity in which the participants are involved. Furthermore, the data show that each body behaviour is often related to specific turn management types.

The second study, which investigates vocal and non-verbal cues of turn shift in part of the corpus, shows both similarities and dissimilarities with the cues found by Duncan in English dyadic conversations. More specifically, the syntactic cues hold also in our data, while the intonation patterns seem not to be the same in the two languages. We found also that the head and the body (Duncan focuses on the hands) complete ongoing movements to signal that the speaker is finishing the turn and is prepared to give the floor to the interlocutor. This analysis of co-occurring behaviours will in future be extended to the entire corpus.

In the third study, where we looked at turn length and number of words in a turn, and how these varied for each participant depending on the conversation, we found an interesting gender difference, which to our knowledge has not been reported earlier. In our corpus, thus, males talk more when they interact with females, while females talk more when their interlocutor is another female. However, we would like to test whether this tendency can also be seen in other datasets and in different types of conversations before making any general claim about gender differences.

Several issues of relevance to the phenomenon of turn management were not touched upon here and could be studied either based on this corpus or adding datasets for different languages. For example, we did not consider mirroring or synchronised behaviours of the participants, see e.g. (Campbell, 2009). Nor did we investigate whether the degree of familiarity of the interlocutors affects turn management, or to what extent turn management cues vary from language to language. Especially the last issue, which is in line with preceding comparative studies of body behaviours and other communicative functions in different cultures and communicative situations (Lu et al., Under publication; Maynard, 1987; Navarretta et al., 2012; Navarretta and Paggio, 2012), can be studied by comparing the findings described here with similar turn management data from parallel multimodal corpora from the NOMCO collection.

Acknowledgments

This research was carried out for the Nordic NOMCO project, which is funded by the NORDCORP program under the Nordic Research Councils for the Humanities and the Social Sciences (NOS-HS). We want to thank the NOMCO project partners: Elisabeth Ahlsén and Jens Allwood from University of Gothemburg and Kristiina Jokinen from University of Helsinki as well as the Danish annotators, Sara Andersen, Josephine B. Arrild, Anette Studsgård and Bjørn N. Wessel-Tolvig.

References

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C. & Paggio, P. (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing. *Multimodal Corpora for Modelling*

Human Multimodal Behaviour. Special Issue of the International Journal of Language Resources and Evaluation, 41(3–4), 273–287.

Argyle, M. & Cook, M. (1976). *Gaze and mutual gaze*. Cambridge University Press, Cambridge, UK.

Boersma, P. & Weenink, D. (2009). *Praat: doing phonetics by computer*. Retrieved May 1, 2009, from http://www.praat.org/.

Campbell. N. (2009). An audio-visual approach to mea- suring discourse synchrony in m ultimodal conversa- tion data. In *Proceedings of Interspeech 2009*, pp. 12–14.

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

Cowley, S. J. (1998). Of Timing, Turn-Taking, and Conversations. *Journal of Psycholinguistic Research*, 27(5), 541–571.

Du-Babcock, B. (2003). A comparative analysis of individual communication processes in small group behavior between homogeneous and heterogeneous groups. In *Proceedings of the 68th Association of Business Communication Convention*, pages 1–16, Albuquerque, New Mexico, USA.

Duncan, S. Jr. and D.W. Fiske, D. W. (1977). *Face-to-face in- teraction*. Erlbaum, Hillsdale, NJ.

Duncan, S. Jr. (1972). Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology*, 23(2), 283–292.

Esposito, A., Campbell, N., Vogel, C., Hussain, A. & Nijholt, A. (eds). (2010). *Development of Multimodal Interfaces: Active Listening and Synchrony*, volume 5967 of *LNCS*. Springer Verlag.

Ford, C. E. & Thompson, S. A. (1996). Interactional Units in Conversation: Syntactic, Intonational, and Pragmatic Resources for the Management of Turns. In E. Ochs, E.A. Schegloff, and S.A. Thompson, editors, *Interaction and Grammar*, pp. 134–184. Cambridge University Press, Cambridge.

Hadar, U. Steiner, T. J. & Clifford Rose, F. (1984). The timing of shifts of head postures during conversation. *Human Movement Science*, 3(3), 237–245.

Jokinen, K. (2011). Turn taking, utterance density, and gaze patterns as cues to conversational activity. In *Proceedings of ICMI-MMI*, Alicante, Spain, November.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26, 22–63.

Kipp, M. (2004). *Gesture Generation by Imitation - From Human Behavior to Computer Character Animation*. Ph.D. thesis, Saarland University, Saarbruecken, Germany, Boca Raton, Florida, dissertation.com.

Lu, J., Allwood, J.& Ahlsén, E. Under publication. A study on cultural variations of smile based on empirical recordings of Chinese and Swedish first encounters. In D. Heylen, M. Kipp, and P. Paggio, editors, *Proceedings of the workshop on Multimodal Corpora at ICMI-MLMI 2011*, Alicante, Spain, Nov.

Maynard, S. (1987). Interactional functions of a nonver- bal sign: Head movement in Japanese dyadic casual conversation. *Journal of Pragmatics*, 11:589–606.

Navarretta, C. & Paggio, P. (2012). Verbal and nonverbal feedback in different types of interactions. In *Proceedings of LREC 2012*, pp. 2338–2342, Istanbul Turkey, May.

Navarretta, C., Ahlsén, E., Allwood, J., Jokinen, K. &, Paggio, P. (2011). Creating Comparable Multimodal Corpora for Nordic Languages. In *Proceedings of the 18th Conference Nordic Conference of Compu- tational Linguistics*, pages 153–160, Riga, Latvia, May 11-13.

Navarretta, C., Ahlsén, E., Allwood, J., Jokinen, K. & Paggio, P. (2012) Feedback in Nordic first-encounters: a comparative study. In *Proceedings of LREC 2012*, pp. 2494–2499, Istanbul Turkey, May.

O'Connell, D. C., Kowal, S. & Kaltenbacher, E. (1990). Turn-Taking: A Critical Analysis of the Research Tradition. *Journal of Psycholinguistic Research*, 19(6):345–373.

Paggio, P. & Navarretta, C. (2011). Head movements, facial expressions and feedback in Danish first en- counters 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*, pp. 583–590, Orlando, FL, USA, July. Springer.

Paggio, P., Ahlsén, E., Allwood, J., Jokinen, K. & Navarretta, C. (2010). The NOMCO multimodal Nordic resource - goals and characteristics. In *Proceedings of LREC 2010*, pp. 2968–2973, Malta, May 17-23.

Peirce., C. S. (1931). *Collected Papers of Charles Sanders Peirce, 1931-1958, 8 vols*. Harvard University Press, Cambridge, MA.

Paggio, P. & Navarretta, C. (2012). Head movements, facial expressions and feedback in conversations - empirical evidence from danish multimodal data. *Journal on Multimodal User Interfaces - Special Issue on Multimodal Corpora*.

Sacks, H., Schegloff, E. & Jefferson, G.. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4), 696–735.

Schegloff, E. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, 29:1–63.

Tanaka, H. (2008). Communication strategies and cultural assumptions: An analysis of French-Japanese business meetings. In S. Tietze, editor, *International Management and Language*, pp 154–170. Routledge, New York, NY.

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

What's in a gesture
On verbs, nouns, actions and objects as reflected in gestures of persons with and without aphasia

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

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

Abstract

This study treats the semantic interpretation of co-speech gestures produced with nouns and verbs. One set of 30 gestures was originally produced in conversation by speakers with aphasia, whereas another set of 30 gestures was produced by speakers without aphasia. Each gesture was mimicked by the experiment leader to a panel of judges. The interpreted meaning was written down by the panel of 13 subjects, 7 with the same linguistic and cultural background as the original producers and 6 with other linguistic-cultural backgrounds. The purpose was to study the possible influence on the interpretations of (i) aphasia – no aphasia, in the originally producing group,(ii) cultural background in the panel, (iii) verb vs. noun (or action vs. object orientation) of the originally co-produced word, and (iv) the level of abstraction of a gesture-word-combination. The results showed no influence from aphasia in the producer or cultural background in the interpreting panel. Action gestures tended to be more frequent for both persons with and without aphasia than object gestures and were used also with some nouns. The level of abstractness was captured in the interpretation of about 75% of the items and in the remaining 25%, the interpretations tended to be more abstract than the originally co-produced word.

Keywords: gesture, aphasia, action gesture, object gesture, abstractness

Introduction

This study investigates the types of co-speech gestures that occur with verbs and nouns in the speech production of a number of persons with aphasia and in a reference database of gestures produced by persons without aphasia. A pilot study was carried out with the purpose of finding out what semantic-semiotic features could be identified in a number of gestures produced by (i) persons with aphasia and (ii) persons without aphasia. The range of semantic features occurring in the interpretations of a certain gesture can be interpreted as the meaning potential of that gesture, i.e. what elements of meaning/-s can be associated with it (Allwood 2003). The panel of judges were 13 academics, 7 of which had the same cultural-linguistic background as the persons originally producing the gestures (Swedish) and 6 of which had other cultural-linguistic background.

Background

The meaning of bodily expressions is less clear than that of vocal verbal expressions, since it is less conventionalized and to a greater extent dependent on indexical and iconic information (cf. Peirce 1932). The relation of gestures and words have been described from two quite different standpoints and also from some intermediate positions. The first standpoint is that iconic gestures and speech are "inextricably intertwined" in development and generation and thereby both interdependent and simultaneous. This is the view of McNeill (1992, 2000), who assumes a common "growth point" for words and iconic gestures. From this type of standpoint, if words are impaired, so are gestures. The opposing standpoint is that words and gestures are independent and that gestures have a mainly self-activating role in facilitating speech production (Hadar et al 1996, Hadar & Butterworth 1997, Beattie & Shovelton 2000, 2002, 2004, 2005). An intermediate view is that gestures are closely related to words, but to some extent independent and more robust. The two latter standpoints harmonize with the view that gesture came earlier in evolution, can be

more robust and can therefore be used for compensation when word finding/production is impaired. In this view, iconic gestures can be replacing words or adding information and they can sometimes be more preserved in aphasia (e.g. Feyereisen & Havard 1999, Ahlsén 1991, Lott 1999). Thus, gestures can be closely related to speech and possibly disturbed when there is a language disorder like aphasia, but, at the same time, they can still be more robust and fill a compensatory function.

There are some hypotheses about what gestures can do for speech, i.e. facilitate word finding and structuring of the spoken contribution. De Ruiter (2006) reports studies by Rimé, Schiaratura, and Ghysselinckx (1984), Graham and Argyle (1975), and Graham and Heywood (1975), where the relation between speech and gesture concerning the activation of *spatial features* has been interpreted in different ways with respect to activation of speech. De Ruiter proposes own problem detection and reallocation of communicative content between modalities as a more communica-tive explanation. Kita suggested that analytic thinking organizes information by hierarchically structuring decontextua-lized conceptual templates (analytic templates) (Kita 2000, Melinger and Kita 2006). Raucher, Krauss & Chen (1996), among others, point out that *language production (i.e. the speech flow) is affected if the use of gesture is inhibited*. This also points to an activating or structuring function of gestures.

A number of proposed theories placing an increased focus on the functions of alignment, embodiment and mirroring functions in communication point to the importance of analyzing the role of gesture in relation to speech (cf. Gallese & Lakoff, 2005, Simmons & Barsalou 2003). Arbib (2005) strongly argues for stepwise evolution via first less complex and then more complex gestures to structured speech and language (an evolution from grasping an object to producing Verb-Argument-structures), drawing on mirror neurons and the fact that Broca's area developed on top of the mirror neuron (F4) area in the macaque.

How are gestures and speech related in terms of the brain areas and brain mechanisms involved? Suggestions, based on, among other things, brain activity data in perception experiments using fMRI (functional Magnetic Resonance Imaging) and ERP (Evoked Response Potentials) have been put forward. The experiments have used paradigms like varying mismatch in words, gestures or both in relation to a previous utterance (e.g. Özürek et al. 2007) or varying the conditions of speech reception between the conditions of no picture, picture but no gesture, gesture matching speech and gesture mismatching speech (e.g. Wu & Coulson 2007). Several areas have been found to be active, but most hypotheses concern Broca's area and adjacent areas (BA45/47) and the premotor area (BA6). Broca's area, which has been found to be active in speech perception, especially perception of verbs of movement and tool based action and sentence processing, has also been found to be active in the perception of co-speech gestures. Özürek et al 2007) propose that Broca's area is a center for integration of speech and gesture perception. BA6 is an area which responds automatically to gestures and other actions. This activation can be modulated by preceding speech. It, thus, appears that brain areas of speech and gesture perception overlap and therefore it is hypothesized that the same type of perception is used for both in an integrated processing. A topic of discussion is how much this processing takes place mainly by activation of sensory-motor "perception maps" or by higher-level co-activation of networks and whether and whether these higher-level networks can be considered as "amodal" or just hierarchically coordinated sensory-motor (cf Simmons & Barsalou 2003). This touches on the relation between meanings and words as well as on the relation between nouns and verbs (Crepaldi et al. 2011). Two main questions are highly relevant for the present study of how gestures accompanying words can be interpreted. The first is (i) whether sensory and motor processing are linked, and the second is (ii) whether nouns and verbs use overlapping areas and are related to similar-related or basically dissimilar concepts.

Purpose and Research Questions

Specific questions in this study are:

1) Is there a detectable difference in how well subject can identify the meaning (here operationalized as the word originally accompanying the gesture) of reproduced gestures from the database produced by persons with aphasia and

from the database produced by reference persons without aphasia?
2) Is there a difference in the ability to identify the meaning and elements of meaning in gestures between persons with the same cultural and linguistic background as the persons producing the gestures and persons with a different cultural and linguistic background?
3) To what extent can a gesture be identified as having a main relation to a noun or a verb – or to an object or action related word?
4) To what extent can the level of abstractness of a gesture be identified? This question concerns whether a particular gesture illustrates a concrete word, such as "head" or a more abstract word, such as "conception".

Method

Material

The analysis was based on gestures extracted from a database of 100 gestures produced together with a spoken noun or verb by 10 persons with aphasia and 100 gestures produced with a spoken noun or verb by 10 reference persons without aphasia. The data had been coded with respect to target word, gesture form, semantic features and a number of other variables. From each of the two data sets, 60 clearly identifiable hand gestures, associated with the spoken production of either a noun or a verb, were randomly selected and mimicked from the video – recording by the experiment leader, giving a total of 120 hand gesture stimuli.

Subjects and Procedure

The 120 hand gestures mimicked by the experiment leader were shown one by one in isolation (i.e. without context or accompanying speech) to a group of academics, 7 Swedish and 6 non-Swedish, in a group experiment, where they were asked to write down the meaning of the gesture or the closest related word (or phrase with target word underlined). The experiment leader was sitting in front of the group and showing each gesture with one repetition, then pausing until all the subjects had written down their response, before mimicking the next gestures. (This procedure was chosen for two reasons: (i) to respect the anonymity of the persons with aphasia in the study, (i) to only produce the type of hand gesture without any context of other accompanying speech, facial expressions, context factors etc. Although there is an element of loss of authenticity in the procedure, the hand movements were quite distinct and easy to mimic.

Analysis

The written responses were analysed with respect to:

- the number of response words for each of the two data sets (aphasia and reference data set) corresponding to the words originally produced by the person making the gesture (i.e. the same word or a near synonym)
- the number of gestures originally produced with a verb, where the response was instead a noun and the number of gestures originally produced with a noun, that were responded to by a verb by the participants in the experiment
- the number of gestures where the word given by the participants in the experiment differed in degree of abstractness from the word produced by the person originally producing the gesture.

Results

The number of words in the responses corresponding to the words originally accompanying the gestures is presented in table 1.

The proportion of words that corresponded to the original word produced with the gesture can be estimated to somewhere between 16-20 percent. This means that the "target word" could be identified from the gesture alone, with no context, in about 1/5 or a little less of the cases.

The question of whether "target words" would be harder to identify in relation to hand gestures originally produced by persons with aphasia than to hand gestures originally produced by reference persons (question 1) was negatively answered. The gestures produced by persons

with aphasia were as easy to interpret as (or possibly easier than) the gestures produced by the reference group.

Table 1. Number of words corresponding to the word originally accompanying the gesture

	Reference database	Aphasia database
Subject:		
A. Swedish	12	16
B. Swedish	11	12
C. Swedish	7	13
D. Swedish	12	16
E. Swedish	10	9
F. Swedish	10	12
G. Swedish/ Eng/Norw	7	7
Total Swedish	69	85
Mean Swedish	9.86	12.14
H. Italian	11	8
I. Arabic	7	9
J. Pakistani	9	11
K. Turkish	11	11
L. Chinese	10	12
M. Chinese	8	11
Total non-Swedish	56	62
Mean non-Swedish	9.33	10.33
Total	56	61
Mean	9.33	10.10

Considering the similarity in cultural background between the subjects in the experiment and the persons originally producing the gesture (question 2), there was no statistically significant difference in ability to identify the "target word", although the persons with non-Swedish background had slightly lower numbers. It can, thus, not be hypothesized from this pilot experiment that cultural-linguistic background plays a substantial role in the ability to interpret spontaneously produced gestures occurring with verbs and nouns.

A tendency to produce action-oriented gestures with not only verbs, but also nouns, has been noticed in the original database (Ahlsén & Schwarz, 2012). This can occur because the noun itself is action oriented, (e.g. a cut, a throw) or because the gesture is perhaps more holistically related to a whole phrase or clause or related to an adjacent verb.

Table 2. Number of verb responses to gestures originally produced with nouns.

	Reference Database	Aphasia database
Subjects		
A. Swedish	3	1
B. Swedish	2	1
C. Swedish	2	1
D. Swedish	2	1
E. Swedish	2	1
F. Swedish	3	1
G. Swedish /Eng/Norw	1	1
Total Swedish	15	7
H. Italian	4	0
I. Arabic	3	0
J. Pakistani	2	0
K. Turkish	2	1
L. Chinese	3	0
M. Chinese	1	0
Total non-Swedish	15	1
Total	30	8

There were considerable differences in how many verbs are given as responses for hand gestures originally produced with nouns by persons with and without aphasia (question 3). The frequency is fairly low in both groups, but clearly higher for the gestures produced by persons without aphasia (30 vs. 8).

The different items varied with respect to how much the level of abstractness of the response words matched that of the words originally produced with the corresponding gestures (question 4). All in all 25% or the responses were clearly at a different level of abstractness than the original words, the vast majority of these being more abstract interpretations of the gestures.

Example 1-3 below illustrates the responses from our 13 subjects (A-M) for two gestures originally co-produced with the verbs for “been placed” and “shrinks” and with the noun “tree”.

Example 1)

Originally co-produced with: “Been placed” (Sw. placerats)

Gesture: Both hands in front of chest, about 20 cm apart with palms towards each other, movement up and down to the right.

Responses.

A. Moving object
B. Putting in one place
C. Division, first, then (abstr)
D. Throw away (thrash)
E. Put there
F. End of discussion (abstr)
G. Roof top (A->O)
H. Put it on a side, forget (abstr)
I. Indicating fed up (abstr)
J. Put it somewhere
K. Throw something down
L. Put sth down to the right side
M. Put it somewhere/there

Most of the responses are related to placement ((9 out of 13) and one of them is adding also an abstract interpretation ”put it on a side, forget”. Other, more abstract responses are: ”division, first, then” and ”indicating fed up”, There are two noun responses ”roof top” (concrete, focusing on the form of the gesture) and ”end of discussion” (more abstract).

Example 2)

Originally co-produced with: “Shrinks” (Sw. krymper)

Gesture: both hands in front of stomach, palms toward each other, fingers bent, about 50 cm apart, movement of hands coming together with some tension, slowly, leaving a smaller distance between them (about cm).

Responses:

A. Sphere (A->O)
B. Together (abstr) (A->adv)
C. Press together -> focus on something (abstr)
D. Focus (V), condense (abstr)
E. Make it smaller
F. Put together
G. Coming together
H. Put all together/assemble it
I. Almost done
J. -
K. Smaller, something is getting smaller
L. Squeeze the object
M. Put them altogether/make it smaller

Only three of the responses contain the focused meaning of shrinking: “make it smaller”, “smaller, something is becoming smaller” “making it smaller”. The meaning of coming or putting together seems more immediate, as in “press together”, “put them together”, “coming together, put all together/assemble it” and “put them all together. This meaning can also be more abstract, as in ”focus on something” and “focus, condense”. Another abstract interpretation is “almost done” (possibly taking the gesture as time or a task shrinking). Finally, there is one noun interpretation “sphere”, focusing on the form of the gesture.

Examples 3)

Originally produced with: “Tree” (Sw. träd)

Gesture: both hands lifted high in front, palms towards each other, first close, then lowered coming further apart, then coming toghether again to about 10 cm apart, then lowered in parallel about 30 cm.

Responses:

A. Showing a shape – possibly woman
B. A man or a person
C. Round at the top getting thinner – showing form
D. Tree
E. Showing the shape of something
F. Show the form
G. Female earth mother (showing hip rounding)
H. ”This shape”
I. -
J. –
K. Symbolizing a woman/female body”
L. A tree
M. Narrow it down

All of the subjects (except two, who gave no response), gave an interpretation of the form of the gesture.

Discussion and conclusions

This study was an exploratory pilot study involving only a small group of participants. it can, thus, be the basis of hypotheses to be tested further on a larger population, rather than more definite conclusions.

Regarding question 1, the finding that the meaning of gestures produced by persons with aphasia was not harder to interpret (in terms of identifying the originally co-produced word) than the meaning of gestures produced by other persons is indicative of similarities in the use of gestures and no general loss of the ability to make comprehensible gestures caused by the aphasia, when there are no noticeable word finding episodes. Both words and gestures were produced by the subjects in the sample used for this study. It does, thus, not directly address the question of gestures occurring in word search episodes where there are noticeable word search/word retrieval problems, which remain to be studied. Such studies will add information about the possible activating and/or compensatory role of gesturing with respect to word production.

The study does not point to any major differences in the ability to identify the meaning of a gesture depending on cultural and linguistic background (question 2), i.e. same or different culture and language as the language producer. The group of subjects is, however, small and represents only five different languages, although they are from widely different parts of the word. The results, however, point in the direction of more general principles for interpreting gestures that apply across cultures and languages, which would make at least some gestures more robust than spoken words in intercultural communication.

The finding that there is a clear difference in how many gestures originally accompanying nouns that elicited verbs as responses between the gestures produced by persons with and without aphasia (30 vs. 8 instances) is interesting (question 3). There can be many reasons for this result. One reason is that the non-aphasic reference database contains a higher frequency of iconic action gestures accompanying nouns than the aphasia database (Ahlsén & Schwarz forthcoming). The reason for this, in turn, could be that the non-aphasic speakers have a higher speech rate. It could then, possibly, be quite easy, especially for a slow and/or complex gesture, to “spill over” in temporal co-occurrence from a verb to a noun. Gestures could also be more holistically planned in relation to chunks or phrases of speech. Many people with aphasia tend to speak slower and perhaps focus more on each word. Since persons with aphasia have greater difficulties in general in mobilizing nouns, they might also therefore have a tendency to produce more gestures related to nouns.

Iconic gestures are produced with abstract as well as concrete nouns and verbs (question 4). What makes a person interpret a gesture as more abstract is probably the accompanying word and other context. In this experiment, such context is missing. The subjects, however, do interpret gestures as related to quite abstract words in almost 25% of the cases, as we have seen.

In ordinary face-to-face interaction, there are a several converging sources of information, three important sources being the *meaning potentials*, i.e. the possible meanings of a gesture and of a word in a specific context.

How does the *meaning potential* of a particular gesture restrict the interpretation of the accompanying word and vice versa? For example, if a person raises his/her hand, this could mean several things, if he/she says the word big this could also mean several things (c.f. a big grape or a big house), but if the raised hand and the word *big* are coproduced, the merged meaning potentials restrict the meaning of both expressions. In this study, only the meaning potential of gestures in isolation was studied. In a future study, we hope to study how the meaning potential of gestures and speech are integrated and, thus, contribute to an investigation of how in face-to-face interaction the multimodal integration of meaning potentials facilitates under-standing.

Acknowledgements

This research has been supported by the European Community´s seventh Framework Programme (FP7/2007-2013) under grant agreement no.231287(SSPNet) and by the project: "Multimodal Corpora in the Nordic Countries" under the NORDCORP program, the Nordic Research Council for the Humanities and the Social Sciences (NOS-HS).

References

Ahlsén, E. (1991). Body communication and speech in a Wernicke's aphasic – A longitudinal study. *Journal of Communication Disorders, 24*, 1–12.

Ahlsén, E. (2011). Towards an integrated view of gestures related to speech, *Proceedings of the 3rd Nordic Symposium on Multimodal Communication. Editors: Patrizia Paggio, Elisabeth Ahlsén, Jens Allwood, Kristiina Jokinen, Costanza Navarretta. NEALT Proceedings Series*. 15 (2111) s. 72-77.

Ahlsén, E. & Schwarz, A. (2013). Features of aphasic gesturing. Clinical Linguistics and Phonetics). (Early online doi:10.3109/ 02699206.2013.813077).

Allwood, J. (2003). Meaning Potential and Context. Some Consequences for the Analysis of Variation in Meaning. In Cuyckens, Hubert, Dirven, René & Taylor, John R. (eds). *Cognitive Approaches to Lexical Semantics*. Moulton de Gruyter, pp. 29-65.

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

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

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. W., & 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.

Beattie, G. W., & Shovelton, H. K. (2004). Body language. In *Oxford companion to the mind.* Oxford: Oxford University Press.

Beattie, G. W., & Shovelton, H. K. (2005). Why the spontaneous images created by the hands during talk can help make TV advertisements more effective. *British Journal of Psychology, 96,* 21–37.

Crepaldi, D., Berlinger, M., Paulesu, E., & Luzzatti, C. (2011). A place for nouns and a place for verbs? A critical review of grammatical-class effects. *Brain and Language, 116,* 33–49.

De Ruiter, J. P. (2006). Can gesticulation help aphasic people speak, or rather, communicate? *Advances in Speech-Language Pathology, 8*, 124–127.

Feyereisen, P., & Havard, I. (1999). Mental imagery and production of hand gestures while speaking in younger and older adults. *Journal of Nonverbal Behavior, 23,* 153–171.

Gallese, V., & Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in reason and language. *Cognitive Neuropsychology, 22*, 455–479.

Graham, J. A., & Argyle, M. (1975). A cross-cultural study of the communication of extra-verbal meaning by gestures. *International Journal of Psychology, 10,* 57–67.

Graham, J. A., & Heywood, S. (1975). The effects of elimination of hand gestures and of verbal codability on speech performance. *European Journal of Social Psychology, 5,* 189–195.

Hadar, U., & Butterworth, B. (1997. Iconic gesture, imagery and word retrieval in speech. *Semiotica, 115,* 147–172.

Hadar, U., Wenkert-Olenik, D., & Soroker, N. (1996). Gesture and the processing of speech in aphasia. *Brain and Language, 55,* 180–182.

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.

Lott, P. (1999). *Gesture and aphasia.* Bern: Peter Lang.

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

McNeill, D. (1992). *Hand and mind.* Chicago: The University of Chicago Press.

McNeill, D. (2000). *Language and gesture.* Cambridge, UK: Cambridge University Press.

McNeill, D. (2007). *Gesture and thought.* Chicago: University of Chicago Press.

Melinger, A., & Kita, S. (2006). Conceptual load triggers gesture production. *Language and Cognitive Processes, 22*, 473–500.

Özyürek, A., Willems, R. M., Kita, S., & Hagoort, P. (2007). On-line integration of

semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of Cognitive Neuroscience, 19,* 605–616.
Peirce, C. S. (1932). *Collectd Papers of Charles Sanders Peirce.* Harvard University Press.
Rauscher, F. H., Krauss, R. M., & Chen, Y. (1996). Gesture, speech and lexical access: The role of lexical movements in speech production. *Psychological Science, 7,* 226–231.
Rimé, B., Schiaratura, L., & Ghysselinckx, A. (1984). Effects of relative immobilization on the speaker's nonverbal behavior and on the dialogue imagery level. *Motivation and Emotion, 8,* 311–325.
Simmons, K., & Barsalou, L. W. (2003). The similarity-in-topography principle: Reconciling theories of conceptual deficits. *Cognitive Neuropsychology, 20,* 451–486.
Wu, Y. C., & Coulson, S. (2007). How iconic gestures enhance communication: An ERP study. *Brain and Language, 101,* 234–245.

Overlaps In Maltese: A Comparison Between Map Task Dialogues And Multimodal Conversational Data

Alexandra Vella
University of Malta
alexandra.vella@um.edu.mt

Patrizia Paggio
University of Malta
University of Copenhagen
patrizia.paggio@um.edu.mt

Abstract

This paper deals with overlaps in spoken Maltese. Overlaps are studied in samples from two different corpora, one consisting of Map Task dialogues, and the other of free face-to-face conversations. The results show that the number and function of the overlaps vary with the presence or absence of pre-defined roles, the nature of the dialogue and the subjects' familiarity with the situation. Overlaps are used to achieve optimal information exchange in the Map Task dialogues, and are a sign of ease and familiarity in the free conversations.

Keywords: overlaps, MapTask dialogues, face-to-face conversations, Maltese

1 Background

The fact that spontaneous speech does not consist of neatly arranged and separated turns, and that speakers, on the contrary, speak over each other and interrupt each other, has been observed by many. In Schegloff (2000) it is recognised that overlaps play a role in what the author calls *talk-in-interaction*, in spite of the general view held by conversational analysts that overlaps are minimised in the turn-taking mechanism (Gardner et al, 2009). More recently, Campbell et al. (2010) have observed that in a free group conversation the number of short, often overlapping utterances, is much larger than the number of longer distinct ones. An interesting way to study overlaps is to examine their nature and function in different types of interaction. Examples are Cetin and Shriberg (2006), who analyse overlaps in a number of different corpora, and Adda-Decker M. et al. (2008), who introduce a framework to measure overlaps in political speech.

A general insight arising from the Cetin and Shriberg study which is directly relevant to the present paper, is the fact that whether or not the participants in a conversation have clearly defined roles plays a significant function in the amount of overlap one may observe. In particular in chaired meetings, in which the general interaction is controlled by the chair, there is little overlap. The effect of medium (whether the interaction happens face-to-face or takes place over the phone) is less important. Conversely, from the data analysed in the Campbell et al. paper, familiarity seems important, such that the more familiar people are with each other, the more overlap they produce when they talk.

This study presents a preliminary comparison between conversations taken from two different corpora of spoken Maltese with specific reference to the issue of overlap, thereby testing the hypothesis that overlaps are used to different degrees and for different purposes in different communicative situations.

The aims of the study are to see (i) how frequent overlaps are in the two corpora, (ii) what types of overlap occur, (iii) how overlaps are distributed between the speakers. In general, we are interested in investigating whether there are systematic differences in the two corpora due to different features such as the presence or absence of pre-defined roles, the different degrees of familiarity between the speakers, or the nature of the conversation.

2 Overlaps: definition and types

An overlap is a stretch of time of variable duration where two speakers talk over each other, and which may or may not result in a change of speaker (Fig. 1 and 2).

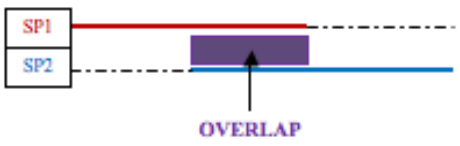

Figure 1: Overlap with speaker change

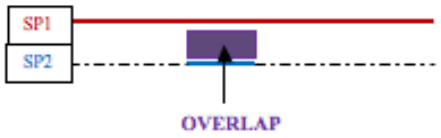

Figure 2: Overlap without speaker change

Different types of overlap may also be distinguished based on different functional categories which include:

1. Overlap in the context of *feedback* (also called Acknowledgement in Carletta et al., 1997): there is no competition for the floor and no change of speaker. This can be lexical (e.g. *orrajt*/*owkey* 'all right, okay', *sewwa*/*tajjeb* 'good') or quasi-lexical (e.g. *mhm*/*eħe*).
2. Overlap in the context of *questions* which require a yes or no answer (*Query-YN* in Carletta et al., 1997): the current speaker relinquishes the floor and a change of speaker is expected. (Preliminary scrutiny of the data suggests that overlap is less likely, though not impossible in the case of *wh*-questions – *Query-W* in Carletta et al., 1997).
3. Overlap involving *interruption*: the two speakers are competing for the floor. The current speaker can retain or relinquish the floor.

We will attempt to establish to what extent overlapping in our data can be characterised using these three functional types.

3 The corpora

3.1 The Maltese Map Task Dialogues

Figure 3: Example of Leader's map used in the Map Task dialogues for Maltese

The first corpus consists of eight Maltese Map Task dialogues which form part of the MalToBI corpus (Vella and Farrugia, 2006). The corpus was designed to be representative of spoken Standard Maltese, participants being carefully selected with a view to balance in terms of age, sex and educational background. The Maltese Map Task design is similar to that used for the HCRC Map Task corpus (Anderson et al., 1991). Two participants engage in a communication gap activity. The aim is for the participant in the Leader role to describe the route on the Leader Map to the participant in the Follower role, whose task is to draw the route in accordance with the information provided by the Leader. The Maps are not identical, thus necessitating an element of negotiation. The Maltese Map Task dialogues involve 16 speakers (8 females and 8 males): half the females fulfil the Leader role and the other half the Follower role, and similarly in the case of the male speakers.

3.2 The Multimodal Corpus of Maltese

The second collection is the multimodal corpus of Maltese MAMCO, which consists of twelve video-recorded first encounter conversations between pairs of Maltese speakers.

Figure 4: Screenshots from the MAMCO corpus. Total side view and split semi-frontal view.

Twelve speakers participated (6 females and 6 males), each taking part in two different short conversations that took place in a recording

studio. The setting and general organisation of the collection replicate those used in the Nordic NOMCO corpus (Paggio et al., 2010) so that it will be possible in future to use the corpora for inter-cultural comparisons. Contrary to other similar collections, in the Maltese Map Task corpus all participants could see each other. As a result, the Maltese Map Task data are directly comparable to the MAMCO data in that non-verbal as well as verbal means of communication were available to speakers for use (only audio recordings of the Maltese Map Task data are available, however).

3.3 The two corpora at a glance

In both corpora the speech has been or is being (in the case of MAMCO) transcribed using Praat (Boersma and Weenink, 2009) and following the guidelines described in Vella et al. (2010). An annotation of head movements is also planned for the multimodal corpus.

Table 1 below provides a comparison of the two corpora along a number of different parameters.

Map Task	MAMCO
Dialogues	Dialogues
Subjects sitting facing each other with two tables between them	Subjects standing at comfortable speaking distance
Unidirectional microphones	Lapel microphones
No cameras	Cameras
Can see each other (face and torso)	Can see each other (entire body)
Have to solve a task	Talk freely
Different roles	No predetermined role
Familiarity not an issue	Do not know each other

Table 1: Features of the two corpora

The most significant features from the point of view of the quantity and types of overlap to be expected from the subjects are the last three, which we shall discuss briefly.

First of all, the Map Task dialogues are task-oriented, while the MAMCO conversations are free face-to-face interchanges. The subjects are only instructed to try to get to know each other, but they are free to choose their own topics of conversation. We consider the MAMCO dialogues examples of *natural* conversation although they take place in a studio, and are *provoked* by the experimenter. So, how natural are they really? In order to investigate this aspect, subjects were presented with a post-experiment questionnaire in which they were required to assess each interaction with scores from 1 (lowest) to 5 (highest) along various parameters having to do with how comfortable they had felt during the conversations. Fig. 10 shows the average scores obtained for each parameter during the first and second recording (the two experiments each participant took place in were scheduled on separate days). For most of the parameters, the self-rated scores fall between 3.5 and 4.5, indicating that the interactions were judged by the participants themselves as reasonably natural. There is a significant increase in the ratings given on the second day as the subjects were more used to the situation and the setting (two-tailed paired t-test, $p=0.0019$).

As far as the role division is concerned, there is a clear distinction in the Map Task dialogues between the Leader, whose task it is to describe the route, and the Follower, whose task involves implementing the directions given. In MAMCO, on the contrary, the participants all have equal status from the point of view of the interaction.

Finally, although the participants did not in fact know each other, familiarity, or rather lack of such, is not really an issue in the Map Task corpus. By contrast, it is a pre-requisite in MAMCO, since the corpus is intended to represent first encounter situations.

3.4 Corpus features and overlaps

In both types of data, overlaps are defined as temporal segments in which the conversation participants speak at the same time. However, the degree and function of overlap are presumably quite different because of the different features of the corpora.

Based on the findings by Cetin and Shriberg (op. cit.), we would expect a greater degree of overlap in the MAMCO conversations because neither of the speakers has a predetermined leading role. In other words, both have to negotiate the floor. On the other hand, the relative discomfort of having to speak to a stranger standing in an artificial space, while being recorded, may inhibit the speakers from producing overlaps. Therefore, we would also expect overlaps to increase as the dialogue proceeds, as speakers get more comfortable with the situation and also more familiar with each other.

As for the functions of overlap in the Map Task dialogues, these include that of assuring the Leader, who gives the instructions, that an instruction has been understood (or the opposite) as well as of maintaining continuity with a view to task completion. Since one of the speakers has a leading role in the dialogue, we expect this person mostly to keep the turn at the end of an overlap. In the first encounter conversations, by contrast, the two participants' main objective is to break the ice and keep up the dialogue. There is therefore, at least based on the nature of the interaction, no reason to expect that one of the speakers should overlap more than the other. If there are differences between the speakers with respect to overlap, this may be due to different factors, e.g. personality traits.

These expectations were verified by extracting the overlaps in selected interactions and carrying out a (limited) quantitative and qualitative comparison across the two corpora.

4 Quantitative analysis

Only two videos have been analysed so far (one for each corpus), therefore the results reported here are tentative and require validation on the basis of an analysis of the rest of the corpus data. Note also that, since the two corpus samples are so small, it made no sense to carry out significance tests at this stage.

The first dimension along which we want to compare the two corpora is the degree of overlap. We looked at this in two different ways by measuring (i) the overlap time over the total talking time, (ii) the proportion of overlap time to approximately half way through the dialogue, and (iii) the proportion of overlap time in the rest of it. The three sets of measures are shown in Fig. 5. For each measurement, the bar on the left represents the Map Task, and the one on the right MAMCO.

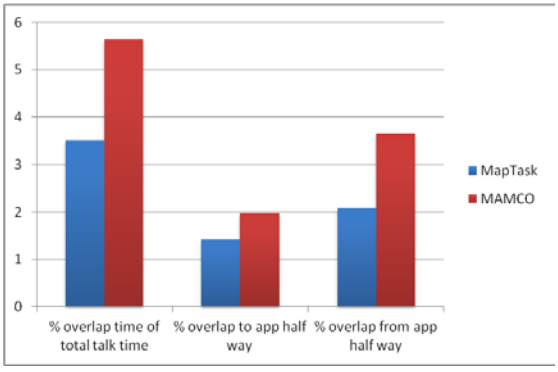

Figure 5: Proportion of overlaps in the two samples

The total length of the two samples is 223.97s for the Map Task file and 207.91s for the MAMCO one. The average overlap length is 0.36s with considerable variation (from 0.04s to 0.96s). As expected, the three measures show that there is substantially more overlap in the MAMCO sample, and also that the proportion of overlap time increases in the second part of the interaction in both samples and especially in MAMCO.

If we look at how the overlaps are distributed between the two speakers (Fig. 6 and 7), we can again observe differences between the two samples. Whereas in MAMCO there are no noticeable differences between the two speakers, in the Map Task sample the Follower (upper bar region) has more overlap time (Fig. 6), whilst the Leader (lower bar region) has a large number of (shorter) overlaps (Fig. 7) .

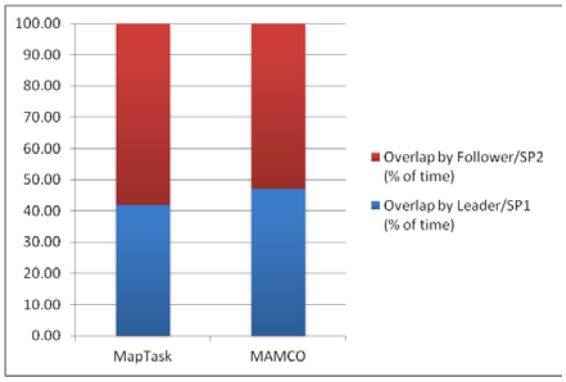

Figure 6: Distribution of overlaps between the two speakers (overlap time)

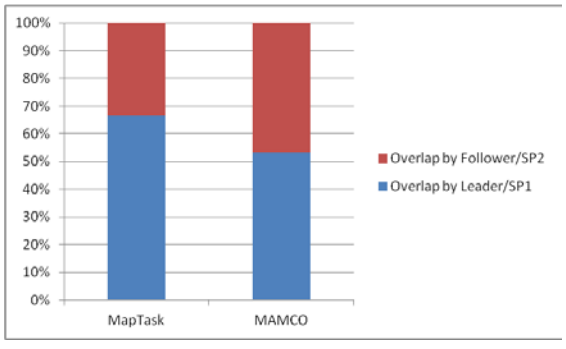

Figure 7: Distribution of overlaps between the two speakers (numbers of overlap)

The difference can be easily understood in terms of the different roles. When the Follower overlaps, the purpose is that of asking for ex-

planations and sometimes commenting on apparently incorrect instructions (type 2 or 3 in our list of functional types). By contrast, the Leader's overlaps are mostly of the feedback giving type (type 1) to answer questions and confirm expectations to then carry on with the instructions.

Let us now look at how overlap relates to speaker change and turn taking.

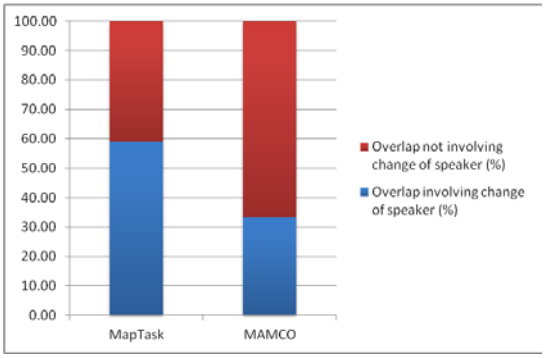

Figure 8: Overlaps and change of speaker

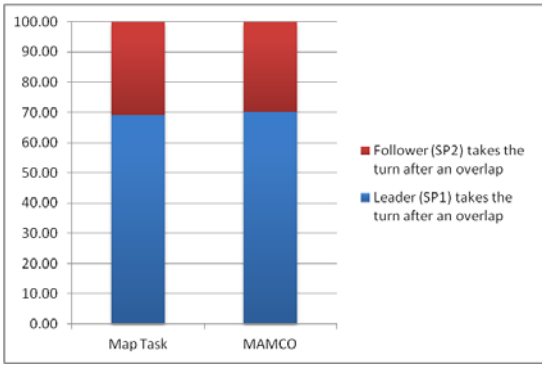

Figure 9: Overlaps and turn taking

Fig. 8 shows what proportion of the overlaps in the two samples result in a change of speaker (lower bar region), while Fig. 9 illustrates which of the speakers takes the turn after the overlap if there is a change.

In the Map Task sample, 60% of the overlaps result in a change of speaker, while the proportion drops to only 30% in MAMCO. If there is a change, one of the speakers takes the turn more often than the other in both samples. In the Map Task, it is the Leader (lower bar region in Fig.9), exactly as we were expecting. The typical situation in which this change takes place is one in which the Follower asks a question to make sure they are doing the right thing, and the Leader answers (by overlapping) and then takes over. The reason why one of the speakers in the MAMCO sample mostly takes the turn after having overlapped, on the other hand, is not caused by any intrinsic characteristic of the dialogue. Rather, it is probably due to the personality and engagement of the specific subject. It could be said that this subject is taking a leading role.

To sum up, the data we have from these small samples tentatively confirm our expectations of the fact that overlapping would be different in quantity and nature in the two corpora. In the following section, we look more closely at specific examples.

5 Qualitative analysis

Examples of the different types of overlap identified in section 2 are presented below.

The first is the type occuring in a context of *feedback*. This type of overlap can involve quasi-lexical as well as lexical elements. An example from the MAMCO corpus involving the use of quasi-lexical elements is the following:

SP1: għandi z-zijiet minn hemmhekk.
I have aunts from there.
SP2: [Mhm.
Mhm.
SP1: In-nanna+] (.) minn Bormla.
My grandmother is from Bormla.

In this and the examples that follow, square brackets are used to indicate the parts of the speakers' turns which overlap. Pauses internal to a turn are shown using (.). For this example, a printscreen of the View & Edit Praat object is also shown in Fig. 11. In the figure the overlap segment is clearly marked across the various annotation tiers. In the exchange, acknowledgement of the fact that the transfer of information has been successful is provided by the use of 'Mhm'. The current speaker continues speaking while the interlocuter gives this feedback, hence the overlap. There is, however, no competition for the floor and no change of speaker. Similar exchanges are also common in the Map Task corpus.

Feedback-related overlaps involving lexical feedback also occur in these data. An example from the Map Task is the following:

SP1: jew Dar Millenia
either Millenia House
SP2: **Dar Millenia** [sewwa
Millenia House, right
SP1: jew] Vjal il-Mara
or Lady Alley

SP1 is providing information on alternative possible routes. SP2 acknowledges receipt and understanding of the information given by SP1 by repeating the location and then adding the lexical element 'sewwa' *right*, to show that he had understood. Again, similar examples can be found in the MAMCO corpus. Note that competition for the floor is not in evidence in these cases, the overlap serving rather to acknowledge successful transfer of information. An interesting feature is the use of repetition of some element from the interlocuter's prior turn (indicated in bold above) in some part of the turn involved in the overlap.

Instances of the second type of overlap, that occuring in the context of *questions* which require a yes or no answer, also occur in both the Map Task and the MAMCO data.

In these examples the current speaker relinquishes the floor by virtue of the very fact of asking a question which requires an answer. A change of speaker is therefore expected. The overlap occurs as a result of a slightly earlier "entry" by the speaker taking the floor, and again not for reasons to do with competition, but rather in a show of cooperative behaviour.

For instance in an example from the Map Task corpus, SP1 provides the answer 'Ija' *yes* to the question 'Minn Triq Mannarino' *Through Mannarino Street?* whilst SP2 is still completing his question. In a similar example, SP1 anticipates the end of the question, in this case a tag question 'Imma s-sitt waħda teżi, hux veru?' *But the sixth one (=year) is a thesis, right?*, with her answer 'Eżatt.' *Exactly*.The third type of overlap identified involves *interruption* of some sort resulting from the two speakers competing for the floor. The current speaker can retain or relinquish floor. Relevant instances are found in both corpora. An example from the Map Task corpus is the following:

SP2: [hemm naqra bogħod
it's rather far
SP1: Trid issib] (.)
You need to find
SP2: biex ngħaddi
to go
SP1: Eħe.
Yes.
SP2: minnha.
that way.

Here, SP1 makes an attempt at giving a new instruction, overlapping, in so doing, with SP2, who is commenting on the difficulty of carrying out an earlier instruction. After a brief pause, SP2 continues with his turn, however, managing to retain the floor to the extent that SP1 not only relinquishes the floor, but proceeds immediately to provide SP2 with feedback ('Eħe' *yes*) on the content he had been trying to transfer at the point when she attempted (and failed) to take the floor.

By contrast, the current speaker (SP1) in the following example from MAMCO relinquishes the floor:

SP1: Mela mill-Università [forsi ġieli rajt wiċċek.
So it's from University that I may have seen your face
SP2: Imma+ ee] (.)
But ee
għandi z-zijiet hemmhekk. In-nanna+
I have aunts from there. My grandmother

Here there is clear competition, each speaker continuing to develop their own separate thread, competing for the floor in the process. It is noteworthy that SP2 enhances his attempt at taking the floor by (i) lengthening the final syllable of 'imma' *but*, (ii) further holding on to his turn through the use of the filled pause 'ee', and (iii) pausing briefly before continuing to speak. These strategies achieve the desired effect: SP1 relinquishes the floor.

A final example will serve to illustrate the use of overlap for a purpose other than acknowledging that transfer of information has been successful, willingly relinquishing one's turn in order to get information required, or negotiating the floor (the three functional categories illustrated above). The following exchange is involved:

1SP1: [Dort ma' Triq l-Ewwel
I went around the Street of the 1st
2 SP2: Nibqgħu sejrin] (.)
We continue on
3 SP1: ta' [Mejju
of May
4 SP2: għal] Triq l-Ewwel ta' Mejju
1st May Street
5 SP1: u (.) għaddejt issa
and I now passed
6 SP2: U għaddejna minn ee (.)
And we've gone through FP
Misraħ il-Lejl [issa
Night Square now
7 SP1: Owkey.]
Okay.

There are 3 overlaps in the above excerpt. The first of these, between 1 and 2, involves complete overlap. After a brief pause there is a second overlap involving SP1 completing transfer of information on the street name in question 'l-Ewwel **ta' Mejju**'; SP2, the Leader, completes the instruction he had been in the process of giving. At this point, the two speakers converge, with SP1 saying she had got to the location in question ('għaddejt issa'), and SP2 restating the current position ('u għaddejna minn'). The last overlap involves feedback on the part of SP1, who is now eager to give reassurance to SP2 that, following the earlier breakdown in communication, realignment has taken place.

6 Discussion and conclusion

Our expectations that overlaps would not be used in the same way in the two corpora have been confirmed, although the small size of the samples used in the analysis renders the results tentative.

As predicted, the lack of predetermined roles in MAMCO as opposed to the clear role division in the Map Task corpus, gives rise to more overlaps in the former. We also see that in both samples, the amount of overlap increases as the dialogue proceeds, showing that the frequency of overlap is dependent on subjects' familiarity with each other and with the situation. The importance of role assignment is also reflected in the fact that in the Map Task dialogues, the Leader mostly has the turn after an overlap involving a change of speaker. Interestingly, participants in free conversations can take on a leader role and show similar turn-taking behaviour.

In spite of the differences, however, there are also similarities in the two data sets, as shown by the qualitative analysis of a number of chosen examples. In particular, the view of overlaps that emerges from the analysis of both corpora is not one in which overlaps are used as an aggressive feature. Rather, overlaps can be seen as a means to achieve optimal information exchange in the map-oriented dialogues, or as a sign of familiarity and ease in free face-to-face conversations. In other words, the view that "optimal" conversation should manifest itself in smooth turn taking without overlap, and that overlaps are detrimental to an optimal exchange, does not capture what happens in either task-related or free dialogues.

In future, we intend to provide a more solid empirical foundation for our results by analysing the full range of recordings in the two corpora.

Acknowledgments

We would like to acknowledge the work of Sarah Agius, Marija Debono and Luke Galea, who transcribed the MAMCO conversations. This work was possible through funding from the University of Malta's Research Grant Fund project LINRP06-02.

References

Adda-Decker M., Barras, C., Adda, G., Paroubek, P., Boula de Mareüil P. & Habert, B. (2008). Annotation and analysis of overlapping speech in political interviews, in *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08).*

Anderson, A. H., Bader, M., Bard, E.G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. & Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech* 34, 351-366.

Boersma, P. & Weenink, D. (2009). Praat: doing phonetics by computer (Version 5.1.05) [Computer program].

Campbell, N. & Scherer, S. (2010). Comparing Measures of Synchrony and Alignment in Dialogue Speech Timing with Respect to Turn-Taking Activity. In *Proceedings of Interspeech*, pp. 2546-2549.

Carletta, J., A., Isard, S., Isard, J., Kowtko, J., Doherty-Sneddon, G. & Anderson, A. (1997). The reliability of a dialogue structure coding scheme, *Computational Linguistics* 23 (1), 13-32.

Cetin O. & Shriberg. E.E. (2006). Overlap in Meetings: ASR Effects and Analysis by Dialog Factors, Speakers, and Collection Site. MLMI06 (3rd Joint Workshop on Multimodal and Related Machine Learning Algorithms), Washington DC.

Gardner et al. 2009. The underlying orderliness in turn-taking - Examples from Australian talk, *Australian Journal of Communication*, 36(3).

Paggio, P., Allwood, J., Ahlsén, E., Jokinen, K. & NavarrettaC. (2010). The NOMCO Multimodal Nordic Resource - Goals and Characteristics, in Calzolari et al. (eds.) *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC '10)*, pp. 2968–2974, Valletta, Malta.

Schegloff, E. A. (2000). Overlapping Talk and the Organization of Turn-Taking for Conversation. *Language in Society*, 29, 1, 1-63.

Vella, A. & Farrugia, P.-J. (2006). MalToBI – building an annotated corpus of spoken Maltese. *Speech Prosody 2006*, Dresden.

Vella, A., Chetcuti, F., Grech, S. & Spagnol. M. (2010). Integrating annotated spoken Maltese data into corpora of written Maltese, in *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC '10)*, Workshop on Language Resources and Human Language Technologies for Semitic Languages, pp. 83-90, Valletta, Malta.

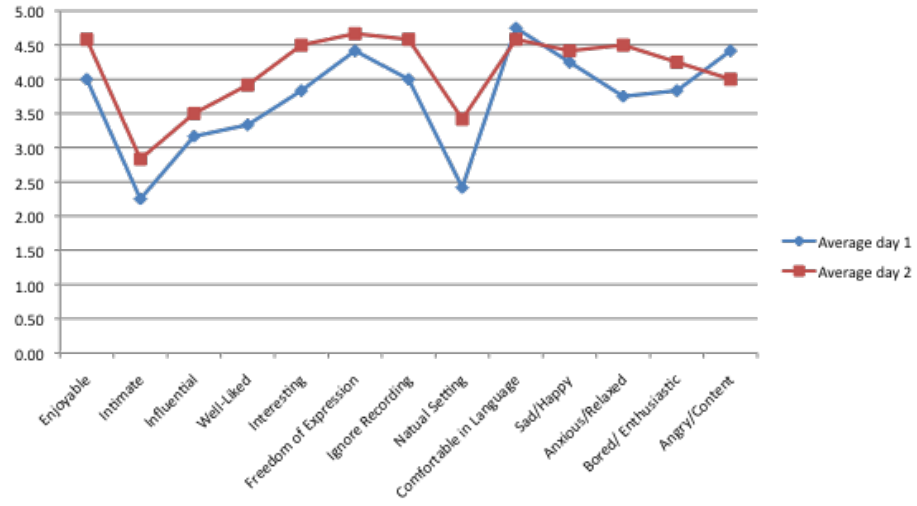

Figure 10: Questionnaire average scores, first and second conversation.

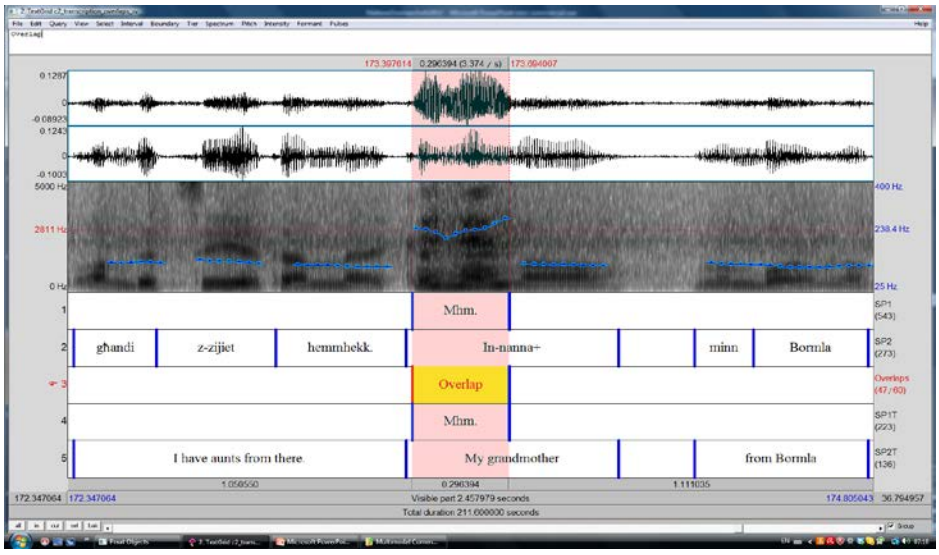

Figure 11: MAMCO example of overlap involving use of the quasi-lexical element *mhm*. The five annotation tiers from top to bottom are used for: transcription of SP1's speech (1), transcription of SP2's speech (2), overlap annotation (3), translation of SP1's speech (4), translation of SP2's speech (5).

Bipedics: Towards a new category of kinesics. An empirical investigation of the expression of attitude, and emotion, through simple leg and foot gesture

Peter O'Reilly
University of Gothenburg
Gothenburg, Sweden
peter.oreilly@ait.gu.se

Abstract

Previous research in the field of nonverbal behaviour and communication has neglected a possible link between simple leg and foot posture and movement (or bipedic gesture) and the expression of attitudes and emotions. The present investigation explored this link in two studies; Study 1 employed analysis of a corpus that consisted of video recordings of first encounter dyadic interaction alongside interactants' self-reported measures of liking for their conversation partner. Study 2 employed a quasi-experimental design whereby participants were asked to interpret liking between interactants portrayed by mannequin dolls. The results from both studies support a link between certain bipedic gestures and the expression of attitudes and emotions. It is hoped that these findings stimulate further research on this neglected part of the human body and its communicative affordances.

Keywords: attitude display, bipedic gesture, kinesics, leg and foot positioning

1. Introduction

In 2004 the Heinz Nixdorf museum began the employment of an anthropomorphic artificial agent to interact with its visitors (Wachsmuth, 2008). Based at what is reputed to be the world's largest computer museum and known simply as 'Max' his design team at Bielefeld University noted communicative competencies such as 'small talk', demonstrable personality traits, and the expression of a variety of emotions (Pfeiffer et al., 2011).

Being a three-dimensional, full-bodied artificial agent Max communicates with visitors and colleagues using a variety of sense modalities. For instance; the hearing sense modality being activated via sound from Max's speech and the visual sense modality being activated via Max's facial expression, bodily gesture and posture. All of these conveying meaning and managing interaction (e.g. via turn management) through a synchronized interplay of speech and bodily communication.

However, in respect to the latter of these when observing Max in action he nearly always appears standing behind a desk, or projected from only the waist up. This, it would appear, is based on the assumption that the lower half of the body plays no part in nonverbal gestural or postural communication.

This assumption may arise from an absence in the scientific research literature investigating the role that the lower half of the human body – the legs and the feet – might play in nonverbal communication.

A corollary of this follows like so – does Max lose any aspect of communication through the loss of one half of his body (the lower half)? And, beyond human-machine interaction, what role might legs and feet play in human to human multimodal communication?

Research that investigates these questions not only contributes to a body of research seeking an understanding of nonverbal, multimodal communication, but also incorporates practical applications. These include; (i) informing the development of communicatively convincing full-bodied artificial agents of the future, with (ii) further commercial application to service encounter situations using *Artificial Embodied Agents* (Salomonson et al., 2013). As well as (iii) contributing to social skills training (Argyle, 1988), which in turn (iv) can be of utility to professionals that rely on interview interaction for the gathering of information (e.g. medicine, therapy and law enforcement).

2. Background

This investigation's empirical journey starts with Ekman & Friesen's (1969b) well known and influential taxonomy which provides for five categories of nonverbal behaviour; *Emblems*, *Illustrators*, *Affect Displays*, *Regulators* and *Adaptors* (Ekman & Friesen, 1969b). These categories, or modes of communication, rely exclusively upon the visual sense modality and for the most part require no sound to transmit meaning (the snapping of fingers and clapping of hands being obvious exceptions). Within this taxonomy

nonverbal communication primarily transmits information pertaining to emotion, its intensity and nature, in four ways; (i) body acts, (ii) body positions, (iii) facial expression, and (iv) head orientation (Ekman and Friesen, 1967). However, within their taxonomy and 'four ways' of encoding emotion there is no room for the role that interactants' legs or feet might play. Indeed, within the specific context of deception cue leakage, the authors state that, "The feet and legs are almost in all respects the worst nonverbal senders" (Ekman & Friesen, 1969a, p.94).

This theoretical and empirical position has both represented and influenced the body of research such that, as already noted, previous research that examines the role that simple leg/foot gesture plays in multimodal communication is limited.

James (1932) was one of the first to explore the relationship between bodily posture, movement and the expression of emotion by analysing 1,200 observations of 347 different postures. However, in spite of the extensive analyses and use of a full-body mannequin / human actor the study failed to take any account of the lower half of the human body (James, 1932).

Smith-Hanen's (1977) study of nonverbal communication in therapeutic settings found that different leg positions were significantly related to perceptions of warmth and empathy, but also noted that, "…the effects of the various leg positions were more complex than the arm positions" (Smith-Hanen, 1977, p.87). Harrigan and colleagues (1985) reported in their study of physicians' use of nonverbal communication that certain symmetric and asymmetric leg positions were significantly related to participant ratings of rapport (Harrigan et al., 1985). In the intervening period between James' (1932) study and those of Smith-Hanen (1977) and Harrigan et al. (1985) other studies have referred in passing to this part of the human body. For instance; linked with quasi-courtship behaviours in therapeutic sessions (Dittmann et al., 1965; Scheflen, 1964, 1965) interviews (Ekman, 1969a) and social encounters (Dittmann & Llewellyn, 1969; Mehrabian, 1968, 1969, 1972/2007). As Harrigan sums up, when compared to research upon the head, face and hands there has been a neglect and a "lack of comprehensiveness" in respect to arms and legs in nonverbal communication (Harrigan, 2008, p.178).

More recently Dael and colleagues (2011, 2012) have developed their Body Action Posture (BAP) coding system in an attempt to code the expression of emotion by all parts of the human body. Using the GEMEP (Geneva Multimodal Emotion Portrayals) corpus their studies however have failed to account for simple leg/foot movement, and position, citing visibility and technical difficulties (Dael et al., 2012).

The present investigation did find one source of material relating to this part of the human body; popular literature (Navarro, 2008; Pease, 1991). Caution was taken in the handling of this material due to a commonly cited lack of scientific method (Harrigan, 2008) and reported "grossly exaggerated claims" (Lecci et al. 2008. p.70). However, careful analysis of these works provided for a basic taxonomous model of simple leg/foot movement and positioning, referred to as *bipedic gestures* in the present investigation, which were then aligned to different emotions and attitudes for the purpose of experimental testing (Fig. 1 below).

Positive Emotion & Attitude	Positive Foot Pointing (standing)	**Basic Emotions**: joy and/or surprise
	Positive Leg Crossing (standing)	**Circumplex (Dimensional) Model of Emotion**: positive valence, high or low arousal
	Positive Leg Crossing (seated)	**Attitudes**: friendliness, submissiveness, either stable or temporal
Negative Emotion & Attitude	Negative Foot Pointing (standing)	**Basic Emotions**: fear, disgust, anger or distress
	Negative Leg Crossing (standing)	**Circumplex (Dimensional) Model of Emotion**: negative valence, high or low arousal
	Negative Leg crossing (seated)	**Attitude**: hostility, dominance, either stable or temporal

Fig. 1: The bipedic gesture model; gestures aligned to different models' concepts of emotions and attitudes.

In respect to the concepts of attitude and emotion there is a significant absence of any one theoretical model which has gained consensus within the scientific literature. In their study Dael and colleagues (2012) list three types of emotion theory; Basic Emotion Models, Dimensional Models, and Componential Models. Reviewing research on attitude Bohner & Dickel (2011) define attitude as, "an evaluation of an object of thought" and highlight two important features; attitudes as being either stable cognitive constructs within memory or something more temporary (Bohner & Dickel, 2011, p.392). Argyle (1988) reports that factorial analysis studies of attitude research reveals two dimensions; Dominance-Submissiveness and Friendliness-Hostility. Furthermore, Argyle argues that emotions and attitudes can be viewed as broadly similar behavioural phenomena based on (i) frequently similar nonverbal display characteristics, and (ii) similar speeds of display onset/cessation (Argyle, 1988, p.86).

Taking the lead from Argyle's position emotion and attitude are subsequently reduced and conceptualized in the present investigation to positive or negative, and aligned to appropriate bipedic gestures from the source literature (Navarro,

2008; Pease, 1991) as summarized above (see Fig.1 and Fig. 2).

Fig. 2: The bipedic gestures portrayed by actors. The arrows indicate positive attitude / emotion towards an object of interest. In negative attitude / emotion orientation the leading foot orientates away and leg forms a barrier to the object of disinterest.

Drawing these threads together; the purpose of the present investigation is to empirically explore and seek validation of bipedic gestures. In doing so ascertaining whether simple leg and foot gestures are associated with the expression of emotion and attitude.

3. Method

Two studies were conducted to explore the link between bipedic gesture and the expression of emotions and attitudes.

Study 1: Corpus Analysis

Design & Materials: Drawing upon the methodological approaches adopted in recent related studies (Dael et al., 2011, 2012) an ex post facto experimental design (Coolican, 1994) was employed using video material from the SSKKII/SCCIIL interdisciplinary centre at the University of Gothenburg. This material forming part of the wider NOMCO corpora (Paggio et al., 2010). This video corpus possessed advantages over others used in previous research (such as GEMEP) as it recorded full-body interaction (see Fig.3), it recorded 'real encounters' with potential for enhanced ecological validity, and it was accompanied with self-report attitudinal data whereby interactants had rated liking for their conversational partner post-conversation.

Participants: The video corpus consisted of 40 recordings of 37 adults (15 male / 22 female) aged approximately between their early twenties and mid to late thirties. Conversations were conducted in Swedish and all but one of the interactants were native Swedish speakers. After extraction and synthesis of data 20 video clips of dyadic first encounter interactions were coded and analysed involving 10 females and 8 male interactants (some interactants were involved in more than one conversation, but never with the same interactant).

Fig. 3: Screen captures from corpus video material used in Study 1.

Procedure: Self-report attitude questionnaires from the video corpus were coded using scale items deemed relevant to liking. These questionnaires were then ranked and arranged into the ten highest and ten lowest scores. A high score denoting a high level of liking and positive attitude/emotional orientation towards the conversational partner and a low score denoting the opposite. Content analysis of the corresponding 20 video clips was conducted using coding units from the bipedic gesture model (see Fig. 1 and 2). Due to coding difficulties caused by observed leg/foot posture and movement not adhering precisely with those within the bipedic gesture model subsequent analysis was adapted by focusing on (i) Frequency of Change in bipedic gesture (number of changes during the encounter), (ii) Leg Crossed behaviour (total cumulative seconds spent in a leg crossed position) and (iii) Negative Leg Pointing (total cumulative seconds spent with leading leg/foot orientated away from partner). Analysis was conducted using an independent samples *t*-test to determine statistical significance in differences between groups arranged according to high/low levels of liking vis-à-vis high/low levels of positive attitude/emotion orientation.

Study 2: The Mannequin Experiment

Design & Materials: Using a quasi-experimental design (Shadish et al., 2002) with a nonrandom, convenience sample Study 2 utilized props to represent human interaction, consistent with previous research (James, 1932; Little, 1968). The props consisted of three artists' mannequin dolls (named A, B, and C) and were used in different scenarios designed to simulate various social interactions. Consistent with this study's experimental design the independent variable (IV) utilized to manipulate participant responses was the positioning of the mannequins' legs and feet according to the bipedic gesture model (Fig.1 and Fig. 2). The subsequent dependent variable (DV) was participants' choice of mannequin according to which they felt was the most or least liked in each scenario. A total of 15 scenarios distributed across 19 web pages was administered online to participants using a proprietary survey tool (SurveyGizmo$_{TM}$).

Experimental controls included; (i) use of a blind control whereby participants were not informed of the full reasoning behind the scenarios until the end of the experiment. (ii) Each scenario being displayed to each participant in the same order without use of counterbalancing to control for order effects. (iii) The selection of a design of mannequin that was minimalistic, gender neutral, and lacking in dress or adornments that might be construed as indicative of status or culture. And (iv) the avoidance of anthropomorphizing the dolls by using human names (e.g. 'Bill', or 'Ingegerd') but instead opting for the use of ambiguous referents 'A', 'B' and 'C'. The latter two controls being employed to counter bias potential in participants' responses.

Participants: A total of 91 participants attempted the Mannequin Experiment of whom 61 (35 females / 26 males) completed all 15 scenarios. In terms of age 39.3% of participants were aged between 25 and 34 years, 41% aged between 35 and 54 years, and the remainder outside of these ranges. Participants originated from eleven different countries with the majority from the UK (47.5%), Sweden (29.5%), and the US, Canada and Australia (combined: 11%). Other nationalities included Cuba, Germany, Hungary, Iraq, Malaysia and Russia (combined: 12%).

Procedure: Three 15cm tall artists' mannequin dolls were arranged into 15 different scenes; four seated and eleven standing, depicting different social interaction scenarios. Each scenario was staged so that bipedic gestures displayed liking and a positive attitude/emotion orientation between the three mannequins.

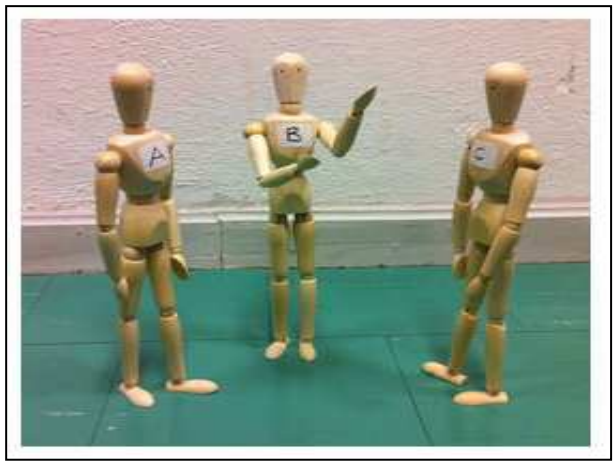

Fig. 4: Scenario 1 from the Mannequin Experiment depicting standing bipedic gestures in simulated social interaction.

Questions were constructed using a forced choice response format ('Our Mannequins are having a chat – from this image who is Mannequin A more interested in and liking more? Is it…[a] Mannequin B, or [b] Mannequin C' – see Fig. 4 above). These were uploaded onto a proprietary online survey tool which was used to administer the experiment and to collate responses.

Analysis was conducted using the χ^2 'goodness of fit' test to determine statistical significance. This approach was taken as 14 of the 15 scenarios provided participants with only a limited choice of responses; one of which followed the experimental prediction. As participants had a 50/50 chance of selecting a 'correct' response (the DV) the χ^2 'goodness of fit' test enabled a determination of whether responses were occurring by chance or as a result of an underlying variable (the IV).

Fig. 5: Scenario 14 from the Mannequin Experiment depicting seated bipedic gestures in simulated social interaction.

Ethical Considerations: Participants were advised that they could withdraw at any time, that

participation was anonymous and confidential, and full explanations were provided to achieve informed consent. The final page of the online survey tool also invited further general feedback and one such request was received.

4. Results

The results from the present investigation are arrange around its two studies.

Study 1: Corpus Analysis

Grouped Data	n = number participants	$\bar{x}$ Attitude Scores	Bipedic Gesture Change (frequency)	Negative Foot Pointing ($\bar{x}$ seconds)	Negative Leg Crossing ($\bar{x}$ seconds)
Highest Attitude Score	10	30.8	17.6	123.6	54.5
Lowest Attitude Scores	10	14.2	19.8	170.6	89.9
Female Participants	10	21.6	16.2	144.6	61.8
Male Participants	10	23.4	21.2	149.1	82.6
Highest Scoring Females	5	29.8	15.6	120.2	23.2
Highest Scoring Males	5	31.8	19.6	127.0	85.8
Lowest Scoring Females	5	13.4	16.8	169.0	100.4
Lowest Scoring Males	5	15.0	22.8	171.2	79.4

Table 1: A summary of findings obtained by using the adapted units of analysis. Columns denoting seconds reflect total cumulative time spent in that bipedic gesture/posture. Frequency represents a count of changes in bipedic position.

Table 1 (above) summarises findings from the present investigation's first study. The 'Attitude Score' reflects the average score obtained from participants' self-report questionnaire in terms of gauging liking and degree of positive attitude/emotion orientation to their respective conversation partner. A high 'attitude' score reflects positive orientation, a low 'attitude' score reflects a negative attitude/emotion orientation.

The frequency and duration of certain observed bipedic gestures proved to be statistically significant upon analysis.

The ten participants with the highest attitude scores ($\bar{x}$=30.8) displayed less Negative Foot Pointing ($\bar{x}$=123.6 seconds) compared with participants with the lowest attitude scores ($\bar{x}$=14.2) who displayed more Negative Foot Pointing ($\bar{x}$=170.6 seconds). Analysis of this difference revealed a statistically significant effect (t = −0.784, df = 11.67, $p < 0.05$).

The ten participants with the highest attitude scores ($\bar{x}$=30.8) also displayed less Negative Leg Crossing ($\bar{x}$=54.5 seconds) compared with participants with the lowest attitude scores ($\bar{x}$=14.2) who displayed more Negative Leg Crossing ($\bar{x}$=89.9 seconds). Analysis of this difference revealed a statistically significant effect (t = −1.011, df = 13.693, $p < 0.05$).

A possible gender effect was observed whereby the five female participants with the lowest attitude scores ($\bar{x}$=13.4) displayed more Negative Leg Crossing ($\bar{x}$ =100.4 seconds) than the females with the highest attitude scale scores ($\bar{x}$=29.8) who displayed 4.29 times less Negative Leg Crossing ($\bar{x}$=23.2 seconds).

It was also observed that male participants changed their bipedic gestures more frequently than female participants. Data related to highest attitude scores ($\bar{x}_{male}$ = 19.6, $\bar{x}_{female}$ = 15.6), lowest attitude score ($\bar{x}_{male}$ = 22.8, $\bar{x}_{female}$ = 16.8) and gender generally ($\bar{x}_{male}$ = 21.2, $\bar{x}_{female}$ = 16.2) all reflected this pattern (Table 1).

However, analysis of the observed gender patterns failed to confirm these as statistically significant effects.

Frequency of Bipedic Gesture Change within comparison groups (e.g. within male/female participants groupings) appeared to exhibit less difference and so appeared more stable.

Study 1: Corpus Analysis

Summary data (Fig. 6, Table 2) and statistical analysis (Table 2) revealed that participants' responses followed experimental predictions contained within of the bipedic gesture model and were highly statistically significant.

The experiment's first depiction of a bipedic gesture, in Scenario 1, elicited 79.1% of responses in line with prediction. Analysis of this data revealed the result to be statistically significant (χ^2 = 30.87, df = 1, $p < 0.001$).

Overall, 13 out of 15 of the scenarios produced a statistically significant result of which 12 followed prediction and one followed in the opposite direction to predicted response.

Scenario	n	% of Responses following Prediction	χ^2 value	*p* value
1	91	79.1	30.87	0.0001
2	88	84.1	40.91	0.0001
3	86	73.3	18.60	0.0001
4	84	64.2	6.86	0.0088
5	80	67.5	9.80	0.0017
6	75	76.0	20.28	0.0001
7	72	18.1	29.39	0.0001
8	71	62.0	4.07	0.0437
9	70	70.0	42.33	0.0001
10	70	91.4	48.06	0.0001
11	70	74.3	16.51	0.0001
12	70	81.4	27.66	0.0001
13	61	52.2	0.13	0.7184
14	61	85.5	34.79	0.0001
15	61	43.5	1.17	0.2794

Table 2: Responses for each scenario showing number of participant responses (n), % of responses in line with prediction, and corresponding χ^2 and *p* values.

Table 2 reveals a response fatigue effect where participant mortality is seen to steadily increase throughout the 15 scenarios with 32.9% of those who started failing to complete. It can also be observed from Fig. 6 that the profile of bars on the bar chart is not incrementally increasing or decreasing indicating that the data is free from an order effect.

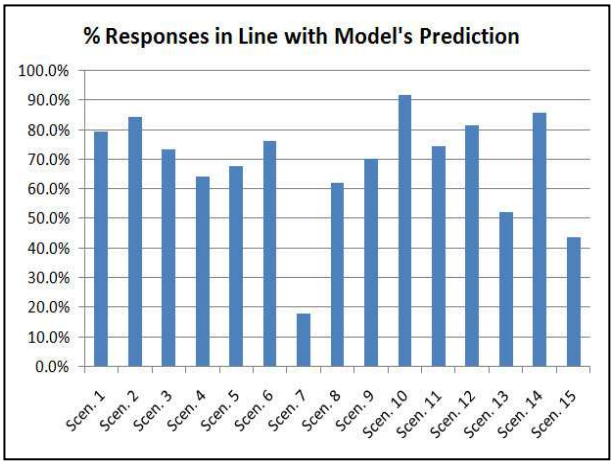

Fig. 6: Chart showing each scenario (1-15) where the blue bars represent the proportion of participant responses in line with bipedic gesture prediction.

5. Discussion

The purpose of the present investigation was to empirically explore and attempt to validate whether simple leg and foot movement and positioning, or bipedic gestures, could express attitude and emotion.

Two studies were conducted where the data and results obtained from both pieces of research supported this contention. Additionally, some insight was provided into methodological problems associated with researching this part of the human body.

Results from the Corpus Analysis study indicate that when an individual meets someone for the first time (in a standing position) their legs and feet will behave in certain ways depending on their attitudinal and emotional orientation to that other person.

Participants from the first study crossed their legs less, and pointed their leading foot away from their conversation partner less, when interacting with a person they later reported liking towards. This lending support to the bipedic gestures of Negative Leg Crossing and Negative Foot Pointing.

Two possible gender effects were also observed where firstly, female participants who liked their conversation partner least displayed nearly four and a half times more Negative Leg Crossing than those female participants who liked their conversation partner most. And secondly, male participants were observed changing their bipedic gesture more often than female participants regardless of positive or negative attitudinal orientation. These results, whilst marked, failed to meet the appropriate thresholds of statistical significance. However, they are suggestive of a possible future path of inquiry alongside other individual differences such as age, personality and culture.

As mentioned methodological problems were encountered and these were caused, in the main, by the surprising complexity of this part of the human body. A pair of legs and feet combined have 120 bones (with more muscles) with which to position and shape a wide range of positions, shapes and postures. These were observed to be performed rapidly and with high frequency. Categorizing and making these fit into a fixed model, or taxonomy, of gestures proved difficult to the extent that from the six basic bipedic gestures only two could be practically measured. Though it should be noted that two of the bipedic gestures were automatically excluded in Study 1 because the corpus used recorded only standing interaction.

These methodological issues are consistent with the experience of previous studies (Dael et al., 2011, 2012; Smith-Hanen, 1977). As well as being consistent with Ekman and Friesen's (1967) commentary and dismissal of this part of the human body as being difficult and not worthy of enquiry. It is perhaps these factors that have resulted in the

present absence of research related to bipedic gestures.

In spite of methodological issues experienced in Study 1 the experimental data and results obtained from Study 2 provided empirical support for all six bipedic gestures.

In Study 2's fifteen scenarios bipedic gestures were simulated alongside the manipulation of other nonverbal cues such as gaze, bodily posture, head position, and arm and hand movement (see Fig. 4 and Fig. 5). In twelve of these scenarios where the mannequin's bipedic gesture had been manipulated to display liking and a positive attitudinal/emotion orientation to a specific other mannequin, participants accurately decoded this without prompting and over (or alongside) other nonverbal cues.

Although encouraging various limitations can be identified with Study 2's experimental design. First, the forced choice format only made use of two response options in the scenarios which perhaps had a channeling effect on responses. A more complex design might make use of more response options for participants to use, or alternatively, make use of a likert scale where participants estimate how much mannequin A likes mannequin B or C. An additional improvement includes varying the order of the scenarios for each participant so they respond to each scenario in a randomized and different sequence. Although this counterbalancing control is usually used in treatment of order effects – of which none were observed in the present investigation – their employment might nonetheless have made these very encouraging results more robust.

In respect to the theoretical implications of these findings, and where a taxonomous bipedic gesture model might fit, a return to the start of this investigation's empirical journey is appropriate. As mentioned, Ekman and Friesen's (1969b) pervasive and influential scheme arranges nonverbal behaviour into five distinct categories. A possible theoretical question that arises is whether bipedic gestures are a new, sixth kinesic category to be positioned alongside *Emblems*, *Illustrators*, *Affect Displays*, *Regulators* and *Adaptors*, or whether they are one of these. If the latter, then which of these categories would provide the best fit?

At this stage it is perhaps premature to begin fashioning theoretical implications until more work has been conducted. A perhaps interesting question remains though in respect of whether bipedic gestures would constitute a new and sixth category of kinesics.

Other questions that were encountered in the course of the present investigation include, (i) what are the implications of left or right footedness? Does a left footed person display Negative Foot Pointing with their left foot? (ii) The present investigation's first study used corpus material of first encounter dyadic interaction between approximately similar individuals. What might be the effects of status, culture, gender, age, personality and context to display rules? All of these remain unanswered and perhaps represent interesting questions that can be taken forward within future research.

6. Conclusion & Future Research

The results from the present investigation supports a link between certain leg and foot movements and positions, or bipedic gestures, and the expression of attitude and emotion. However, the findings here only represent a beginning to investigating a part of the human body neglected in the literature concerned with nonverbal, multimodal communication. Replication and further investigation with the inclusion of individual differences, culture, context and status may all prove interesting avenues. Ultimately, it is hoped that the present investigation encourages more empirical research which in turn will add to a neglected body of enquiry.

7. Acknowledgements

The research reported here was submitted as a thesis paper under the MSc. Programme in Communication at the University of Gothenburg in July 2012. Thanks goes to staff and fellow students at the SSKKII/SCCIIL Interdisciplinary Centre as well as at the Department of Applied Information Technology, for all their support, inspiration and guidance.

8. References

Argyle, M. (1988). *Bodily Communication* (2nd edit). London: Routledge.

Bohner, G. & Dickel, N. (2011). Attitudes and Attitude Change, *Annual Review of Psychology*, *62*, 391-417.

Coolican, H. (1994). *Research Methods and Statistics in Psychology* (2nd ed.). London: Hodder & Stoughton.

Dael, N., Mortillaro, M. & Scherer, K. R., (2012). The Body Action and Posture Coding System (BAP): Development and Reliability. *Journal of Nonverbal Behavior*, *36*, 97-121.

Dael, N. & Scherer, K. R. (2011). Emotion expression in body action and posture. *Emotion*, *12*(3), 1085-1101.

Dittmann, A. T., Parloff, M. B. & Boomer, D. S. (1965). Facial and Bodily Expression: A study of Receptivity of Emotional Cues. *Psychiatry*, *28*, 239-244.

Dittmann, A. T. and Llewellyn, L. G., (1969). Body Movement and Speech Rhythm in Social Conversation. *Journal of Personality and Social Psychology*, *11*(2), 98-106.

Ekman, P. & Friesen, W. V. (1967). Head and Body Cues in the Judgment of Emotion: A Reformulation. *Perceptual and Motor Skills*, *24*, 711-724.

Ekman, P. & Friesen, W. (1969a). Nonverbal Leakage and Clues to Deception. *Psychiatry*, *32*(1), 88-105.

Ekman, P. & Friesen, W. (1969b). The Repertoire of Nonverbal Behavior: Categories, Origins, Usage and Coding, *Semiotica*, *1*, 49-98.

Harrigan, J. A. (2008). Proxemics, kinesics, and gaze. In: Harrigan, J. A., Rosenthal, R. & Scherer, K. R. (Eds.) (2008). *The new handbook of Methods in Nonverbal Behavior Research.* Oxford: Oxford University Press, pp. 137-198.

Harrigan, J. A., Oxman, T. E., & Rosenthal, R. (1985). Rapport expressed through nonverbal Behaviour, *Journal of Nonverbal Behaviour*, *9*(2), 95-110.

James, T. (1932). A Study of the Expression of Bodily Posture, *Journal of General Psychology*, *7*, 405-437.

Lecci, L., Snowden, J. & Morris, D. (2008). Using Social Science to Inform and Evaluate the Contributions of Trial Consultants in the Voir Dire. *Journal of Forensic Psychology Practice*, *4*(2), 67-78.

Little, K. B. (1968). Cultural Variations in Social Schemata, *Journal of Personality and Social Psychology*. *10*(1), 1-7.

Mehrabian, A. (1972/2007). *Nonverbal Communication*. London: Aldine Transaction.

Mehrabian, A. (1969). Significance of Posture and Position in the Communication of Attitude and Status Relationships. *Psychological Bulletin*, *71*(5), 359-372.

Mehrabian, A., (1968). Relationship of Attitude to Seated Posture, Orientation and Distance. *Journal of Personality and Social Psychology*, *10*(1), 26-30.

Navarro, J. (2008). *What Every Body is Saying – An Ex-FBI Agent's Guide to Speed-Reading People*. New York: Harper.

Paggio, P., Allwood, J., Ahlsén, E., Jokinen., K. & Navaretta, C. (2010). The NOMCO Multimodal Nordic Resource – Goals and Characteristics. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M. & Tapias, D. (Eds.). *Proceedings of the Seventh Conference on International Language Resources and Evaluation* (LREC'10) Valletta, Malta. May 19-21. European Language Resources Association (ELRA).

Pease, A. (1991). *Body Language – How to read others' thoughts by their gestures*. London: Sheldon Press.

Pfeiffer, T., Liguda, C., Wachsmuth, I. & Stein, S. (2011). Living with a Virtual Agent: Seven Years with an Embodied Conversational Agent at the Heinz Nixdorf Museums Forum. In: Barbieri, S., Scott, K. & Ciolfi, L. (Eds.), *Proceedings of the Re-Thinking Technology in Museums 2011 - Emerging Experiences*. Limerick: Think Creative and The University of Limerick, pp. 121.131.

Salomonson, N., Allwood, J., Lind, M. & Alm, H. (2013). Comparing Human-to-Human and Human-to-AEA Communication in Service Encounters. Journal of Business Communication, *50*(1), 87-116.

Scheflen, A. E. (1964). The Significance of Posture in Communication Systems. *Psychiatry*, *27*, 316-331.

Scheflen, A. E. (1965). Quasi-courtship Behavior in Psychotherapy. *Psychiatry*, *27*, 316-331.

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). Experimental & Quasi-Experimental Designs for Generalized Causal Inference. Melbourne: Wadsworth Cengage Learning.

Smith-Hanen, S. S. (1977). Effects of Nonverbal Behaviors on Judged Levels of Counselor Warmth and Empathy, *Journal of Counseling Psychology*, *24*(2), 87-91.

Wachsmuth, I. (2008). 'I, Max' – Communicating with an artificial agent. In: Wachsmuth, I. and Knoblich, G. (Eds.), *Modeling Communication with Robots and Virtual Humans*. Berlin: Springer, pp.279-295.

Multimodal communication in intercultural health care interactions

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

Nataliya Berbyuk Lindström
University of Gothenburg
Gothenburg, Sweden
berlinds@chalmers.se

Abstract

This paper investigates and discusses the role of multimodal communication, especially gesture and facial expressions, in intercultural doctor-patient interactions. Three main phases of typical doctor-patient interactions were studied in a database of recorded interactions between foreign physicians in Sweden and their Swedish patients. The phases were analyzed from the point of view of (i) communicative challenges and (ii) the functions of multimodal communication in each of the phases. Types of content, levels of intentionality and different types of signs are discussed and the discussed functions are exemplified with extracts from the database.

Keywords: multimodal, gesture, health care interaction

Multimodal communication – Introduction and framework of analysis

All face-to-face communication is multimodal, i.e. we use the auditory and visual modes of perception and we produce speech, other sounds and gestures (communicative body movements of head, face, arms, hands, fingers and body posture). There is a natural and most often quite unconscious interplay between different means of expression, where, the distribution of the burden of communication can be dynamically altered, as an adaptation to the conditions of communication. In addition to speech and gestures, objects such as written documents, tools and instruments can be used in medical interaction and in the physical examination, the tactile modality is also to some extent used.

Types of content and types of signs

Many types of content con be conveyed by gestures, the three most important ones for face-to-face interaction being *factual information or main message, interaction regulation* and *expression of emotions and attitudes*.

We can describe three types of signs, according to Peirce (1931-39): *icons* , which represent something by means of a similarity relation (e.g. outlining the form of an ear to refer to an ear), *indeces*, which represent something by means of contiguity (i.e. pointing to an ear to refer to an ear), and *symbols*, which represent something through an arbitrary relation (i.e. using the word "ear" to refer to an ear). Icons and indeces are called motivated signs and they are motivated by a link of analogy which is established by contiguity of the body movement and its reference and then subject to abstraction (cf. Calbris 2011).

Degrees of intentionality and control

How aware and consciously in control are we in our use of motivated and arbitrary signs? The different types of signs bear a certain relation to different levels of intentionality and awareness in communication. There is a continuum of degrees of intentionality in communication and the three levels presented here are trying to capture part of this: 1) indicating is unconscious communication (such as becoming pale from fear), 2) displaying is intentionally showing something (i.e.intentionally coughing to show that you have a cold), and 3) signaling is consciously showing that you show something (e.g. by saying "I have a cold", cf. Allwood 2002).

Gestures, sign types, degrees of intentionality, and cultural/linguistic conventions

There is a tendency for gestures to 1) more often be motivated signs (icons or indeces) and 2) more often be produced at a lower level of intentionality and control than speech. This means that when we study gestures, we are interested in motivated signs and a relatively low level of intentionality.

At the same time, the meaning of gestures is to a great extent determined by their context. This context can consist of: preceding gestures, preceding and simultaneous speech and the activity context where the gesture occurs. An important prerequisite for a gesture to function communicatively is that the participants in the interaction have the same convention for what feature is significant in the gesture, in its context, to render a specific meaning.

What is a communicative body movement and how is it distinguished from action in general? Actions can become gestures when they are interpreted as communicative. Since gestures are not always very consciously produced, it can sometimes be difficult to determine whether a certain action should be counted as a gesture or not. We want to explore all possible means of communication, including speech, gesture, other actions and use of props (text, pictures, objects etc.).

Consequences for gestures in intercultural health communication

Gestures have to be studied in video-recorded interaction, since the participants are often not conscious enough about them to be able to answer questions about what gestures they use and why.

Gestures are most often motivated (i.e. they have a relation to the content they represent that is based on similarity or contiguity/causality rather than arbitrary, as is the case for symbols (e.g. spoken words). This means that most gestures have a fundamentally different relation to what they refer to than most spoken words. They are, thus, not necessarily subject to the linguistic restrictions of symbolic-arbitrary communication of an intercultural interaction between persons with different linguistic backgrounds. Believing that gestures cannot be used for communication if speech cannot be used is therefore not warranted in any simple way. Such a position could be the main prediction related to theories of gesture and speech as "intertwined" and generated together (cf. McNeill 2000). On the contrary, gestures can fill essential communicative functions as well as activating word finding and structuring of one's own speech (cf. Calbris, 2011).

Gestures are, nevertheless, not completely universal, but do show some cultural and conventional variation. This means that not only should we believe that gestures can be very useful for intercultural communication, we should also be conscious that there is a certain risk for misunderstanding of gestures that are based on different cultural conventions.

The role of gestures in relation to speech in intercultural health communication

Doctor-patient communication is only one of the many types of intercultural communication in health care settings. But it is an important activity, which comprises many of the most critical features of intercultural communication and has an interesting structure with respect to the use of gestures and action.

The most typical structure of a doctor-patient interaction is 1) opening-greetings, case history and current health problem, 2) physical examination, and 3) prescriptions, advice and closing of the interaction. The three phases have different characteristics. The first phase is a spoken interaction, where the patient provides most of the spoken information, although the case file often provides unspoken information to the doctor. The doctor is mainly asking questions and eliciting speech, but also interpreting the answers. The second phase is a more physical and action oriented phase, where speech is mostly only instrumental in managing the physical examination and takes the form of requests and questions from the doctor and responses, sometimes also clarification questions from the patient. There are sometimes also short mentioning of results. The third phase, again, is a mainly spoken verbal phase, this time with the doctor as the main speaker and the patient as the recipient. We will now take a closer look at the three phases from the perspective of multimodal communication in an intercultural context.

Method

Data and analysis

This study is based on a combination of data from recordings of medical consultations. 63 recordings of medical consultations were used in the study; 34 recordings of foreign doctor - Swedish patient and 29 of Swedish doctor - Swedish patient consultations. The recordings were made after obtaining written consent from everyone involved. The recorded interactions were transcribed, using the Gothenburg Transcription Standard (GTS) and Modified Standard Orthography (MSO) (Nivre 1999, 2004) and the transcriptions were checked by two independent checkers. Relevant sequences were selected and analyzed.

The analysis was made by using the framework of Activity-based Communication Analysis (ACA) (Allwood 2000) for identifying the affordances related to the different phases of the doctor-patient interactions (see above). Against this background, gestures related to main communicative goals in the subsequences of each phase were identified by using a micro-analytic qualitative approach and typical examples were extracted in order to illustrate the functions of multimodality in the different phases..

Results

Phase 1. Greetings and establishing contact, case history and current health problem

Greetings are highly ritualized, especially in formal situations, such as doctor-patient interaction, and this ritualization applies to multimodal communication. Usually, the doctor takes the lead in the greeting and thereby determines the procedure. In an intercultural interaction, the doctor and the patient might have different conventions for greeting, but the patient most often adapts to the doctor. The doctor, can, however, if he or she is from another culture, hesitate when it comes to questions such as: Should the doctor and the patient approach each other (and how much should, in that case, each of them move, how close should they come)? Should doctor and patient shake hands or not? Should they establish eye contact or not? Should they smile or not? There can also be uncertainty as where each of them should sit, but this is usually resolved by the doctor indicating a chair and sitting down. This phase can take place in an office room, with the doctor sitting at a desk and the patient sitting either on the other side of the desk or on the same side as the doctor. It is also not uncommon that this phase takes place in a clinical examination room with both participants sitting on stools or the patient on the gurney and the doctor on a stool. In cases where more than one person have come for the visit, e.g. a mother with her child or a wife with her husband, there can also be some uncertainty of the role of these added persons, if they should be present, where they should be seated and how much and in what form they should participate. Family involvement in the treatment process of a patient is obvious, expected and essential in many cultures. It is especially common for more collectivistic cultures, i.e. the cultures in which an individual is viewed primarily as a member of a group (family), who supports them through lifetime. On the contrary, in Sweden, as well as in other more individualistic societies (cf. Hofstede et al 2010) , it is often the often patient-individual, who stands in focus, rather than the patient-family member.

Since greetings are highly ritualized, they are usually performed with a high degree of automaticity, except when there is some insecurity (which there might be in any doctor-patient interaction as well as in any intercultural interaction).

What is said is conventionalized, as well as the gestures. Conventional power differences between doctors and patients are likely to, to some extent, affect how the greetings are performed and how the participants take their respective places in the room. The doctors greet first and ask a conventional “How is it going?” “How are you?” in combination with direct eye contact.

Establishing rapport

During the first phase, it is usually important to establish some form of rapport and tension release in the interaction. This may be one of the hardest parts of the interaction when it is intercultural, since it is by no means easy to learn. Examples of ways of establishing rapport in this type of interaction are: making a

joke - responding to the joke, exchanging casual remarks about the weather and exchanging mutual smiles. Comments about the surrounding can also be used as well as reference to something that the participants might have in common. In intercultural medical consultations, establishing rapport is especially important, as it helps to alleviate uncertainty and stress, which might be caused by cultural differences.

While it is possible to learn the formal procedure of doctor-patient interaction in a different culture, some aspects of these rapport-establishing sequences can be very subtle and they are usually not subject to any explicit instruction or practice in language education. Their subtlety also make them highly "vulnerable", so that attempts can easily miss the point and be misinterpreted or just not understood.

Smiles, adequate and positive feedback by head movements and facial expression are important in this phase, especially if there are problems of language understanding and production. Single or repeated feedback words and some repetition of words are also adequate means for establishing rapport in general. Specific side sequences or jokes might not always be understood and it is here that, for example, facial expressions can help to clarify the sequence and possibly even in the end achieve the intended effect.

The case history and the current problem

This part of phase 1 consists of question-answer sequences, where the doctor asks the questions and the patient answers and of narratives produced by the patients, with feedback and follow-up questions from the doctor. The establishment of mutual understanding ideally involves showing an attitude of encouragement and interest/engagement by the doctor and this can, to a great extent be achieved using an expressive, but discrete, body language and short expressions. Some of the ways of doing this are maintaining mutual eye contact, leaning a bit forward and expressing calm, as in example 1, where a Russian doctor shows her interest and involvement from the first question she asks the patient:

Example 1)
D: < what is on your [heart what] are we going to do [today] >
< body movement: leans forward towards P, concerned, direct eye contact>
P: [yes]
P: [I] < / > I <I got so much pain in / in my shoulder blade here you know >
<sigh >, <body movement: left hand on right shoulder >
D: yes

[Transcription conventions:
P,D = patient, doctor; [ ] = overlap; ( ) = uncertain transcription; /, //, // = pauses of different length; + = incomplete word; CAPITALS = contrastive stress; : = lengthening; < > marks relevant sequence of comment in transcription and comment line]

Some touching can also be involved. In example 2, below, the Swedish doctor is reading the file, but when the patient expresses her concern and dissatisfaction with an earlier physical examination, she immediately turns her gaze to the patient. The doctor listens to the patient and touches her arm.

Example 2)
D: then it is so that you have been to the gynecologist first
P: <yes < (...) > >
<hand gesture: turning some papers and looking at them in the medical journal >, <a sound of disgust , gaze to the side >
D: in october //
P: < it was horrible >
< gaze: down to the side >
D: < it was horrible >
< empathetically>
P: <yes but you > / <he doesn't do them wrong >
<1 D: hand gesture: touching P:s left arm >, <gaze: looking up >
D: what was it what was it that was horrible
P: < eh he did an examination in eh the bladder / because I had so much pain [after]
< gaze: looking down >

Since the body is mostly in focus, pointing, demonstration and pantomime are means for the patient to use in his/her narrative and for the doctor to use in responses/ interpretations and clarification questions. Medical conditions, names of diseases, medications and treatments are linguistically difficult items and gestural information often has to carry parts of this type of information. Pantomime rendering of events, iconic gestures and indexical gestures in relation to body parts are useful in

this phase, which is very focused in factual information. In intercultural medical consultations use of gestures is especially important, as it helps to avoid lack of understanding and misunderstanding.

Consider example 3, below, an excerpt from an interaction between a German doctor and a patient who had undergone back surgery:

Example 3)
D: e:h // eh / (…) thrombosis oh oh oh oh oh // yes but why < >(fusion what was that) unstable < < (spoldiroristes) > //
< *gaze: looking down in the papers and reading* >, < *hand movement: waving illustrating instability* >
P: < spondylolisthesis > < i don't follow > < // >
<*head movement: shake* >, < *laughter* >
D: < you do < not > >
<*laughter: P* >, <*gaze: looking in the papers* >

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

Example 3 continued)
D: [you < had] surgery here >
< hand gesture: right hand on back >
P: < back surgery yeah >
< *hand gesture: right hand on back* >
D: yes
P: fourth fifth
D: why //
P: yes [it was unstable i suppose]
D: [what was it] < unstable >
< *head movement: nod* >
P: yes
D: < okay >
< *head movement: nod* >
P: the joints are fused you know [(…)] yeah
D: [1< that what i mean >1] unstable < it < flies > like this // front > // it is called < (spoldirolistes) >
< *head movement: nod* >, < *hand gesture: right hand in the air doing a sliding gesture* >
<*meaning: glides* >, <*meaning: spondylolisthesis, hand gesture: pointing at P with right hand* >4
P: [< m >]
< *head movement: nod* >

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

If rapport is established and the participants manage to take each other into consideration in this phase, the phenomenon of "mirroring", "alignment" and "co-construction of meaning" can develop, involving a mutual adaptation and flow of communication, which makes it smooth and efficient. It involves a visible coordination of body movements, so that the participants, for example, nod, smile, change body posture, perform similar gestures simultaneously or in rapid succession. Such sequences can be reported as very satisfactory by the participants, who feel that communication is fluent and easy. It might be more difficult to achieve this flow in an intercultural interaction with asymmetric power distribution, like the doctor-patient consultation, but, since it does involve so much of body language it is probably possible to bypass some of the differences, given that both parties show an open attitude. The co-construction of meaning in the interview is likely to be more easily achieved when there is this type of flow.

Phase 2. The physical examination

This is a physical-manual-instrumental phase, which is communicatively quite different from phase 1 and it is often clearly delimited from phase 1, for example, by moving into another room or another part of the room, by the doctor telling the patient to take of his/her clothes, by the doctor putting the case file away, standing up, picking up instruments and telling the patient to stand, lie or sit in a particular place and position. It can also be quite a sensitive phase for the patient. To prepare the patient for examination can be a challenge for a doctor. In example 4, the German doctor notices the patient's stress about the forthcoming physical examination and therefore attempts to console him:

Example 4)
D: mhm // okay // < yeah well i will examine you > a little
< *gaze: looking down in the papers* >
P: m
D: and you are ready < // < it was not that > dangerous < you are so > < /// > be afraid of it < not > < we don't < bi+ > bite > // < we don't inject you > >1

< laughing >, < hand gesture: pointing at P with right hand >, < hand gesture: both hands waving >, < gaze: looking down in the papers >, < hand gesture: both hands waving >, < hand gesture: illustrating biting with left hand >, < cutoff: bite >, < hand gesture: illustrating an injection with a syringe >
P: it will surely be
D: we just talk and // and examine you little and // try to help you

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

Often, there is also a phase introducing phrase from the doctor, such as *okay, now we will take a look*. The doctor naturally takes command in the examination phase, which linguistically can contain many requests/orders from the doctor. These can be very short and are mostly accompanied by indexical gestures, such as pointing or by demonstrations of something. Pointing and "handling", together with deictic (indexical) words like *here*, *there* and *this*, is common.

Specific for phase 2 is also that a number of medical instruments are in focus. In the interaction during examination between the Iranian oculist and his patient in example 5, the doctor puts eye glasses on the patient and points to the board:

Example 5)
P: now I see < < n k e y u d f n > //and then I see // > < m e k n c v f g < // b > > < I mean // no > [yes] maybe it is good // no < it didn't do that > is probably the same // as I said <7 was >7 // I have to have a prescription for
< gaze: D looks at the board >, < letters on the board >, < letters on the board >, < hand gesture: D takes another glass out of a box >, < hand gesture: D puts the glass into the glasses P are wearing >, < body movement: D stands up and takes off the glasses P is wearing >, < sigh >
D: [you see]

The phase is also different from the others in that the participants move around and change positions and it is one of quite few communicative situations, which involve legitimate tactile communication. The doctor uses his/her hands for example to palpate, e.g. feel a lump, to press in order to locate pain, or to direct or bend an arm into the right position for examination.

The physical examination can be a threatening situation for the patient, especially in a foreign culture and it is one which requires some openness, flexibility and understanding from the doctor. It is here mainly cultural differences in how a physical examination usually is carried out that have to be considered. In addition, gender is an influential factor. In Scandinavian cultures, the doctor has complete access to the patient's body, unlike in other cultures, when a number of requirements can apply.

Verbal explanations from the doctor of what is going on and why are valuable to ensure that the patient feels comfortable with the investigation. This can, however be taken care of already in phase 1, as preparation for the physical examination.

The examination is also an interesting situation from the communicative point of view, since it is in some ways similar to what Wittgenstein (1953) originally described as a type of "language game" – in his case bricklayers performing an activity where language was not in focus but just provided short instrumental words or phrases, like "here", "up", "more" "give me" etc. This is, thus, a situation, which potentially is possible to perform with very few words and, therefore, could be fairly successfully managed directly by a doctor and a patient speaking different languages. Manipulation, action and gesture can carry most of the communication in a situation like this, which can be compared to physiotherapy - a similar type of activity. Doctors also have different strategies, involving speech and or gesture/action for sharing direct results of the investigation with the patient. Some doctors chose not to say anything about results in this phase, others point to or hold up instruments or point to a

computer screen, usually showing numbers, and some say things like *your values are ok.*

Phase 3. Prescriptions, referrals, advice and closing the session

Phase 3 is, again, clearly distinguished from phase 2, by the doctor initiating movement to another area, perhaps a pause while the patient puts on his/her clothes, usually a return to the seating arrangements of phase 1, the doctor returning to the computer screen or picking up the case file, as in example 6, when the doctor says okay, stands up and returns to his chair in front of the computer:

Example 6)
D: < [yees] // and can you please open your eyes for a little while straight forward > > // m'm / < okay /// >
< *gaze: looking into the machine* >, < *body movement: D stands up and walks back towards the other chair in front of the computer* >
P: does it look good
D: < ye'es / it looks good > / you have cataract in your left < but it is not very much > (...) >
< *body movement: P rolls the chair back facing D* >, < *gaze: looking down and reading P's case file* >

In many respects, phase 3 resembles phase 1, i.e. the placement is usually the same and the activity is more dependent on speech. In phase 3, however, it is the doctor who mainly speaks and contributes the new information (whereas in phase 1, the patient provided the main new information).

The information given in phase 3 consists of the doctor rendering the results, providing prescriptions for medications and possibly referrals to other clinics or experts, explanations from the doctor of how medication should be taken, what will happen at the expert he/she is referring the patient to, advice to the patient about what to do and what not to do and possibly setting up a time for a new visit/check-up.

Phase 3 is critical from the patient's point of view and determines a great deal of his/her satisfaction and potential "compliance", i.e. whether the patient will follow the instructions given by the doctor or not. The patient very often has an opinion already when coming to the doctor about what is the problem and what could or should be done about it. If the doctor has managed to capture this opinion and reconcile it with or relate it in some other way to his/her own results and recommendations, the patient's trust in the doctor is likely to be higher. The active listening from the doctor during phase 1 is, thus, important for the outcome of phase 3. For this, the earlier establishment of rapport is extremely important, since it gives the patient more confidence in the doctor and makes it more likely that the patient will reveal his/her own opinion about the condition, which the doctor can then take into account in phase 3.

In this phase, it is extremely important that the doctor makes him/herself understood and that the patient shows whether he/she understands or not and can pose clarification question. All means of multimodal communication can be used to achieve this.
Some of the important gestures in this phase are spatial and temporal referents and demonstrations. There is often a time course, like a period of medication or other treatment, different times of the day when medication has to be taken, numbers involved in the dosage, manners of taking the medication, how it should and should not be kept etc., as well as instructions about things to do, like taking a walk, resting etc. This can be managed by establishing referents in space and time, either imaginary or by using objects and by representing time as a line, a clock circle or in some other way. For this, the use of gestures is very important. The doctor can enact how the medicine should be taken, how it should be locked in, what the physiotherapist will do and so on. In example 7, the Russian doctor points on the list with the patient's medications to show which ones will be renewed and also demonstrates by rubbing the patient's shoulder how the ointment should be used:

Example 7)
D: left [good] / < but then we will see then these I wait to renew / but these I must renew > // < and I will prescribe < a good gel > / a good > such gel for the shoulder / so when it like OUCH eh it burns then you must put it on / and it goes away // do you want it like that
< *hand gesture: pointing at the paper* >, < *hand gesture: P takes the paper* >, < *hand gesture: D touches and rubs P's shoulder* >
P: [yes]

Another important part is the explanation of why the medication, referral and/or advice are given. As important as the multimodally enhanced presentation of information is the

sensitivity to multimodal cues from the patient, that can provide information about whether the patient believes in the doctor's explanation, whether the patient understands the prescriptions and advice and whether he/she is likely to follow them. This can often be detected from small shifts in the patient's facial expression and in less than optimal feedback, e.g. few or hesitant expressions of contact, perception, understanding and attitudinal reactions. It can also be seen in the body posture, for example, if a patient is leaning back and perhaps averting his/her eye gaze, there is reason to believe that there is some skepticism or reluctance to follow the doctor's advice. Such more or less subtle signs are important to notice. The closing of the session is, like the opening, highly ritualized both concerning what is said and what is gestured.

Conclusion and discussion

The doctor-patient interaction is only one type of health care interaction, but a fairly typical one. There is also some variation in how this interaction is conducted in different cultural contexts. However, the account given in this chapter points to many functions of multimodal communication, especially the use of gestures (i.e. communicative body movements).

In an intercultural context, especially where the participants do not share a common language, iconic and indexical sign types and the establishment of rapport with the help of indicated and displayed communication of positive attitudes and interest in what the other participant is trying to communicate are perhaps harder to achieve and at the same time more important than in a monocultural interaction.

It is true that much of gestural communication takes places at lower levels of conscious control than spoken or written communication and therefore, in the usual case, is not in focus of the participants. Nevertheless, it is taken in and affects the interaction in important ways. In order to handle sensitive intercultural interactions, there is a point in health care personnel trying to take the three steps of 1) being consciously open to cultural differences and showing flexibility, 2) acquiring some awareness about possible intercultural differences in multimodal communication and, especially, in the possibilities of perceiving and producing gestural communication more consciously as a strategy, and 3) training towards and increasing intercultural and multimodal proficiency to be used in intercultural health interactions.

Acknowledgements

This research has been supported by the Swedish Research Council (VR) under grant agreement 2006-1598 "Semantic processes in word finding and gestures" and by STIAS (Stellenbosch Institute for Advanced Research).

References

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. (2002). Bodily Communication - Dimensions of Expression and Content. *Multimodality in Language and Speech Systems.* Björn Granström, David House and Inger Karlsson (Eds.). Dordrecht: Kluwer Academic Publishers, pp. 7-26.

Calbris, G. (2011). *Elements of Meanding in Gesture*. Amsterdam, John Benjamins.

Fiscella, K., Roman-Diaz, M., Lue, B.H., Botelho, R. & Frankel, R. (1997) "Being a foreigner, I may be punished if I make a small mistake": assessing transcultural experiences in caring for patients. *Family practice*, 14(2), 112-116.

Hofstede, G. & Minkov, M. (2010). *Cultures and Organizations: Software of the Mind*, 3rd ed. New York, McGraw-Hill.

McNeill, D. (1996). *Hand and Mind: What Gestures Reveal about Thought*. Chicago, Illinois, University Of Chicago Press

Nivre, J. (1999). Modified Standard Orthography, Version 6 (MSO6). University of Gothenburg, Department of Linguistics.

Nivre, J. (2004) Gothenburg Transcription Standard (GTS) V.6.4. University of Gothenburg, Department of Linguistics

Peirce, C. S. (1931-38). *Collected Papers of Charles Sanders Peirce*, 1931-38, 8 vols. Edited by Hartshornd, C., Weiss, P. & Burks, A. Cambridge, MA, Harvard University Press.

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.

Wittgenstein, L. (1953). Philosophical Investigations. Oxford, Blackwell.

Predicting the attitude flow in dialogue based on multi-modal speech cues

Peter Juel Henrichsen
Copenhagen Business School
pjh.isv@cbs.dk

Jens Allwood
University of Gothenburg
jens@ling.gu.se

Abstract

We present our experiments on attitude detection based on annotated multi-modal dialogue data1. Our long-term goal is to establish a computational model able to predict the attitudinal patterns in human-human dialogue. We believe, such prediction algorithms are useful tools in the pursuit of realistic discourse behavior in conversational agents and other intelligent man-machine interfaces. The present paper deals with two important subgoals in particular: How to establish a meaningful and consistent set of annotation categories for attitude annotation, and how to relate the annotation data to the recorded data (audio and video) in computational models of attitude prediction. We present our current results including a recommended set of analytical annotation labels and a recommended setup for extracting linguistically meaningful data even from noisy audio and video signals.

Keywords: attitude detection, prediction of attitude flow, attitude annotation, multimodal speech cues

Introduction

Sharing of content and alignment of attitudes are two of the basic features and goals of human communication, most clearly in face-to-face communication. These features and goals are also present in human-computer interaction, especially when the computer is represented by an ”embodied communicative agent” (ECA). To be a natural and smooth communication partner, an ECA has to be sensitive to the attitudes of its interlocutor, thus it has to have processes for recognizing and producing attitudes. This we will call attitude administration below. We here present an analysis of the acoustic features of attitude expression in the Swedish part of the Nordic NOMCO project database.

In this paper we first discuss the challenges of attitude administration in a simplified experimental setting, viz. the prosodic component of a typical TTS system (text-to-speech). We then approach the even more difficult realm of dialogue. We believe, models of attitude administration in man-machine dialogues should build on annotated recordings of human-human conversations. We present some ideas for detecting and exploiting the correlates between the acoustic features of the speech signals and the communicated attitudes, using a subset of the Swedish NOMCO data (audio files and anvil-annotated video-files). Based on recorded naturalistic examples, we discuss how to pre-process the raw audio files and the original annotation files (ANVIL-format) preparing an automatic attitude recognizer.

Attitude administration in monologue

It is a well-established experience among constructors of synthetic voices (TTS, Text-To-Speech systems) that an incoherent or unnatural prosodic contour is extremely disturbing to the listener. Human listeners will, in general, be fairly forgiving of clumsily spliced phonetic segments and sudden clicks and cracks to the sound image; after all, we are often exposed to badly encoded speech signals in our mobile phones, and as long as the prosodic contour is authentic and the words reconstructable, we manage to compensate without too much cognitive effort. In contrast, a speech signal with a prosodic encoding out of sync with the intended message cannot be compensated by subconscious means since it is no longer redundant, but contradictory, the reconstruction effort now depending on an intellectual decision procedure. For this reason, naturally sounding prosody has a high priority in any ambitious TTS project. Unfortunately, the principles of prosody assignment are anything but simple and mechanically applicable.

Prosody is the quintessential parameter for emotions and attitudes in speech; by a subtle change in prosodic outline, an utterance may shift it's psychological effect entirely, from earnest to ironic, happy to sad, tentative to confident, or even communicating several emotions-attitudes simultaneously.

Prosody assignment, then, is ultimately an AI complete enterprise. Since genuinely intelligent reasoning systems are probably still decades away, or centuries, we currently have no better option than *mimicry*. By simulating human behavior through prosodic models trained on conversational data, at least we may be able to avoid unwanted attention traps as discussed above. Modern commercial TTS systems invariably employ large databases of human read-aloud data (usually 100+ hours). The sound repositories are, of course, aligned with phonetic transcriptions, but may also be annotated for parameters like style, mood, voice (assertive/interrogative/imperative), discourse function, and so on. By analyzing an input text through these parameters and using the result as an advanced multi-dimensional search query, a best-match for each text element is identified in the sound database. When successful, the speech produced is thus composed of played-back sound instances where the human reader was in a state matching the requirements of the text, not only phonetically, but in a generalized sense reflecting even the attitude. The best modern TTS systems often approach a 'nature identical' prosody when the input text conforms to the style and vocabulary supported in their sound database. Recent examples of TTS projects with highly conscious approaches to the psychological factors of prosody assignment include Aylett et al (2008), Oparin et al (2008), and Henrichsen (2012).

Attitude administration in dialogue

In TTS systems, a naturally-sounding prosodic rendering of an input text can often be determined through rule-based text analysis and intelligent database querying, as explained above. When entering the realm of dialogue, however, prosody assignment becomes far more challenging. Speaker A's attitude pattern must now be determined by speaker B within a very short time frame based on a wealth of multi-modal sensory data, or speaker B will be at risk of producing bizarre feedback (or other attention traps). Such attitude administration may not be perceived by humans as a great challenge, but in spoken language agents, any rules for xyz must be made explicit. Inspired by the success of data driven TTS, one could suggest to compile xyz, but no (manageable) database could ever cover the potential attitudinal variation in live conversations. What cues, then, can be computerized and exploited by an automatic agent tracking the attitude of the human interlocutor?

One approach is to build computational models trained on human-human data, applying them to recordings of dialogues. We concentrated on a sub-part of the NOMCO material consisting of eight dialogues from the "first-encounters" corpus. Our reason for selecting these eight were that (only) these conformed to these requirements:

- video+sound recording
- two extra sound tracks using high-quality chin mounted mics
- individual anvil tiers including markup for a range of attitudes (introduced shortly)
- mixed population of male and female informants

The experimental setup

All of the eight recordings contained two students meeting for the first time. Their instructions were to get to know each other. For most interactions this meant that they exchanged information about names, present occupation and interests. Both participants were standing up about 50 cm away from each other, face to face at an angle of about 90 degrees, and were filmed against a white background. They could move freely in all directions. They were typically friendly and attentive to each other.

Multi-modal corpus data - a computational challenge

As is often the case with speech signals recorded under quasi-ecological conditions, the acoustic quality leaves something to be desired with respect to signal to noise ratio, channel separation, reverberation, and so forth. In our recordings, there are several instances of over-steering (clipped samples), and the reverberation is measured to about 250dB/s corresponding to an echo of approximately 400 ms. These facts combined with a modest channel separation at 20 dB makes it difficult to

perform pitch tracking for the individual speakers (see below). Regarding the functional coding, all files were checked by a separate person than the annotator. The synchronization of the audio and video streams were out of sync by >1% in some instances and in these cases had to be manually assessed.

To these circumstantial challenges come the tractability issues. As mentioned, computational attitude prediction must be quick and responsive. CPU-heavy decoding methods are therefore not feasible (e.g. automatic speech recognition) leaving us with the 'easy', low-level acoustic parameters such as F0 (pitch), intensity, spectral tilt, and Harmonicity-to-Noise ratio (HNR). We introduce each of them in the following section. They are all very well understood in a linguistic frame of reference.

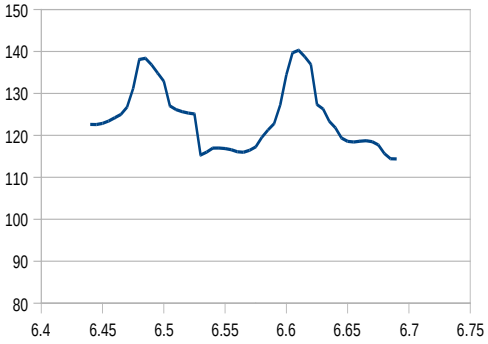

Figure 1. F0 tracing of a two-syllable word (NOMCO informant V8649L, t=6.44-6.69, utterance “eller”)

Among the acoustic features exploited by linguists, the fundamental frequency (*F0*) is probably the most popular, its interpretation in the audiological domain being so straightforward: The pitch. The difference between *pitch* and *F0* should however be noted, the former being a psychological quality and the latter, a physically well-defined2 property that can be determined by a measuring instrument independent of the human ear. A discriminating example is the so-called overtone singing. When listening to an overtone singer one experiences a succession of pitches corresponding to a certain melodic line; this line is however quite independent of the actual F0 progression and is achieved by the singer changing the filtering effect of his upper speech organs rather than changing the tension of his vocal cords. In ordinary speech, however, F0 tracings usually represent the experienced prosodic contour fairly reliably.

Fig. 1 above shows an F0 analysis of a NOMCO speech sample. The 250 ms sample represents the two-syllable Swedish word 'eller' (Eng. translation *or*) pronounced by a male speaker. This word consists of sonorants only so F0 is defined throughout. Usually, only a fraction of a speech signal will be defined for fundamental frequency since silent passages and passages without phonation (e.g. obstruents like [s] and [k]) do not produce meaningful F0 values. Observe in particular the 'wild' values, which have to be filtered away prior to the prosodic analysis. In our project, high cut-off points at 300Hz for male voices were used, 400Hz for female voices, and low cut-off points at 80Hz for both (fig.2). Even if most of the derived F0 values thus have to be abandoned as undefined or meaningless, the resulting data sparseness is not necessarily a problem for prosody analysis since the missing values can often be interpolated. Movements in the prosodic domain are, after all, relatively slow compared to the succession of phones.

Intensity is another parameter often used in acoustic-phonetic analysis. As a computational data type, this parameter has a quite different profile from F0 being *always* defined (even when the speaker is silent). In fig. 3 an intensity graph is shown for the same sound sample. Comparing the two projections it is obvious that most of the speaker's own activity is represented in the intensity range above 50dB (utterances around t=6.5”, t=11.0”, t=12.0”) while the activity of the other speaker (counting as noise in this audio channel) dominates the range 30-50dB. The limited channel separation adds to the challenges when interpreting the intensity data.

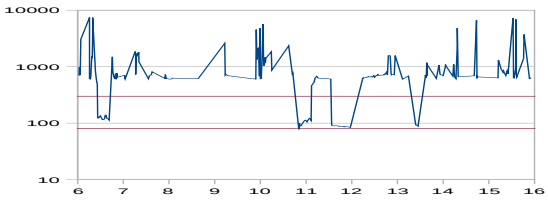

Figure 2. F0 graph, 10 seconds including the “eller” incident discussed above (t=6.45-6.70).

The two red bars indicate the filter for meaningful pitch values (80Hz<P<300Hz).

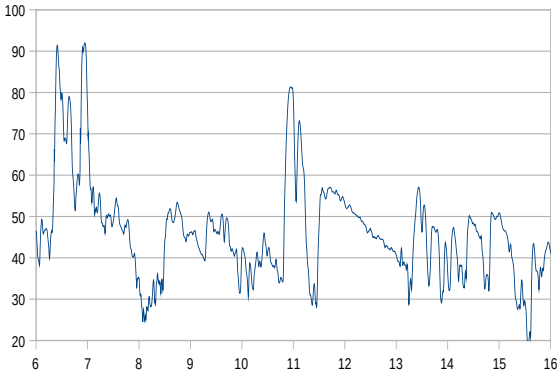

Figure 3. Intensity graph. 10 seconds' recording including the same “eller” incident as above.

The other acoustic parameters we have mentioned are both variants of the intensity parameter. *Harmonicity-to-Noise ratio* (HNR) corresponds roughly to the phonetic 'voicedness'. HNR calculation is performed by separating the harmonic components of the physical sound signal from its noise components, determining the ratio of their individual intensity (amount of energy per time unit). Language sounds with no harmonic components at all such as [s][f][h][p][t][k] and other obstruents produce very low values for HNR, due to their lack of harmonics in contrast to full vowels scoring high. The final acoustic parameter under consideration, the *spectral tilt*, may for instance be determined by comparing the intensity of a sound signal in two distinct frequency bands. Language sounds with much energy in its lower frequencies and less energy in the higher end will, under this interpretation, show a relatively large tilt. Depending on the instantiation of the filter values, various phonetic oppositions (e.g. front-back, open-close, labial-dorsal) and other features can be traced.

Acoustic parameters for prosodic analysis and attitude determination

Among the parameters we have considered, F0 is probably the most relevant for attitude detection, lending itself readily to prosodic interpretation. It should be supplemented by at least one other parameter, though, since data sparseness for F0 becomes a problem with declining acoustic quality (e.g. background noise or poor channel separation for overlapping speech). The other three candidate parameters are all robust in the sense of being defined everywhere, even for silent passages, so in a narrow sense they all serve well for data completion. However, after some initial experiments neither HNR nor spectral tilt proved suitable for our purposes. They both tend to respond more closely to the phonetic fluctuations than to the slower prosodic oscillation while of course the latter is the more important information source for attitude detection.

For these reasons we settled on a computational framework based on fundamental frequency and intensity measurements only.

The anvil annotation format for multi-modal transcription

The recordings were transcribed and annotated using the anvil annotation format for multimodal transcription (Kipp 2001). This format allows simultaneous viewing of the video recording , its transcription/annotation and listening to the audio recording. It also allows viewing of imported acoustic analysis of the audio recording from PRAAT (Boersma & Weenink 2005).

The purpose of the format is to allow analysis of different features of multimodal communicative behavior in synchronized relation to each other, e.g. the relationship of prosody to gestures and spoken words.

The annotation is done by a single annotator and then checked by another annotator. The annotators follow the GST+MSO transcription standard (Nivre 2001, 2004) and the MUMIN standard for multimodal annotation (Allwood et al. 2007).

Preparing the anvil annotations for machine learning

As mentioned, the study reported here used a sub-corpus of eight NOMCO recordings of Swedish first-encounters. The test material includes, for each encounter, one video+sound recording, two individual mono-recordings using good-quality portable microphones, and one anvil annotation tier per speaker.

The team of NOMCO annotators were, to a large degree, free to choose their own attitude labels and delimitation. As a result, the annotation material is extremely heterogeneous. The eight anvil files contain 439 reported

attitude events, the shortest lasting only a small fraction of a second (0.04”), and the longest stretching over almost three minutes (173”). The overall distribution of event durations is shown in fig. 4. Not surprisingly, the set of applied attitude labels is large and diverse: 55 English and Swedish terms, distributed over several grammatical categories.

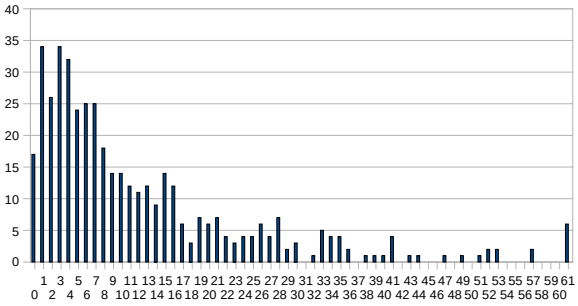

Figure 4. Distribution of attitude events as a function of their individual duration. Attitude events longer than one minute are accumulated at x=61”. Average duration = 12.7”; median = 8.0”.

A suitable subset of the attitude tags had to be extracted for machine learning purposes. As the effectiveness of learning algorithms stand or fall by the cardinality and consistency of data types represented in the training set, we excluded all sparsely used tags. In addition even some relatively densely populated attitudes (labels with many occurrences in the anvil files) had to be excluded due to the relatively low accumulated duration they represent (amount of acoustic data in terms of time frames). Since our investigations are based partly on F0 measurements, this data type being particularly fragile as discussed above, the accumulated duration for each attitude under investigation is thus at least as significant as a selection criterion as is the amount of associated events. Fig. 5 shows the set of applied attitude labels sorted by accumulated duration.

304486 Interested	124961 Friendly	121147 Casual	102014 Bored
99920 Thoughtful	65473 Confident	63998 Insight	55440 Amused
53247 Dominant	44826 Enthusiastic	42894 Uninterested	33589 Impatient
29758 Hesitant	27528 Unconfident	20371 Worried	19724 Happy
19655 Uncomfort-able	18061 Uncertain	16630 Sceptical	16518 Recogni-tion
13651	12113	9330	9210
Certain	Sad	Relaxed	Apprecia-ting
8249 Irritated	8111 Patient	7884 Nervous	7687 Surprised
7193 Describing	6688 Confused	6674 Condescending	5833 Nervous
4634 Suspicious	4553 Provocative	4290 Hopeful	2609 Sarcastic
2426 Compassion-ate	2409 Arrogant	2281 Restless	1274 Embarrasse d
1273 Questioning	1264 Explaining	1050 Confirming	1009 Amusing
967 Shy	920 Ironic	833 Convinced	433 Disappointe d
346 Understand-ing	328 Frustrated	328 Puzzled	313 Thoughtful-ness
225 Deliberative	160 Confirma-tion		

Figure 5. Applied attitude labels. The labels are sorted by accumulated duration (number of 5ms time frames).

After some formal considerations, semantic reflections, and initial experiments, we settled on a test set A10 of ten attitudes.

A10 =

{Interested, Friendly, Casual, Bored,

Thoughtful, Confident, Amused, Enthusiastic,

Uninterested, Impatient}

Each of the A10 terms is richly represented in the ANVIL files, both in terms of amount and accumulated duration. For reasons of dissemination the Swedish terms were excluded altogether (e.g. 'ifrågasättande'); also most of these were used very infrequently.

A Formal Model of Attitude Prediction

Relating the A10-based annotation data to the acoustic data based on F0/INT measurements, we arrive at the attitude profiles shown in Table 1. The profiles are based on three statistical parameters (I-III).

I. F0, standard deviation for each attitude event ('meaningful values' only, see fig. 2)3
II. INT, average for each attitude event (values relativized to the *most silent time slice* in the track)
III. INT, standard deviation for each attitude event

Average-based statistics (mean value and standard deviation) is a convenient way of minimizing the influence of irrelevant sound incidents caused by poor channel separation, echoic distortion, random acoustic events not related to the conversations, and other signal-to-noise problems. Also extreme variation in duration does not present an analytical problem in this perspective. On the flip-side, any contact is lost with the micro-structure of the attitude events when analyzing them as informational atoms, so the attitude model presented here must be a rather coarse one.4

Atti-tudes	**AM**	**BO**	**CA**	**CO**	**EN**
I	26.09	17.01	23.13	28.60	19.05
II	48.14	41.60	41.70	44.52	47.18
III	14.00	12.27	12.17	14.16	11.83
	FR	**IM**	**IN**	**TH**	**UN**
I	34.65	24.8	33.0	33.0	10.0
II	41.15	45.2	29.9	41.8	38.7
III	11.13	12.6	11.1	12.6	10.2

Table 1. Attitude Profiles. The A10 attitudes: AM=Amused, BO=Bored, CA=Casual, CO=Confident, EN=Enthusiastic, FR=Friendly, IM=Impatient, IN=Interested, TH=Thoughtful, UN=Uninterested.

Attitude Profiles as predictors

Each column in table 1 is interpreted as the formal profile representing the attitude in question. Consider a few examples. The A10 label 'Uninterested' is represented in the table by the vector (I, II, III) = (10.00, 38.77, 10.27), these values in turn representing a relatively low standard deviation for F0 ('little modulation') in conjunction with low values for intensity, both on average ('soft voice') and on standard deviation ('inactive articulation'). In contrast, the vector (26.09, 48.14, 14.00) for 'Amused' suggests a far more lively modulation, higher volume, and more active articulation.

Quantifying over all attitude events in the anvil files, we build a prediction table. Each event (i.e. its values for I, II, and III) selects its own attitude label among A10 as its nearest neighbor in the three-dimensional vector space. By way of example, consider the attitude event in anvil file v8649 from *t*=220.44 to *t*=224.12. Let us call it E'. This particular event – or rather, its vector – selects a label 'Enthusiastic' due to the relatively short geometrical distance between E' and 'Enthusiastic' in the three-dimensional data space spanned by I, II and III. No other attitude profile came closer to E' than 'Enthusiastic', this being the predicted attitude for E'.

We are now in a position to compare the annotated attitude for E' to the predicted attitude for the same event. In this case, the annotated and predicted attitudes were identical. Repeating this exercise for all attitude events, we arrive at the prediction table summarized in fig. 6.

Interested: Interested > Confident > Amused > Enthusiastic >> **Bored**
Friendly: Casual > Amused > Impatient > Confident >> **Bored**
Casual: Friendly > Confident > Amused > Casual >> **Thoughtful**
Bored: Uninterested > Bored > Thoughtful > Casual >> **Enthusiastic**
Thoughtful: Uninterested > Bored > Casual > Friendly >> **Confident**
Confident: Impatient > Interested > Amused > Friendly >> **Thoughtful**
Amused: Confident > Interested > Friendly > Impatient >> **Bored**
Enthusiastic: Enthusiastic > Interested > Confident > Amused >> **Bored**
Uninterested: Bored > Casual > Thoughtful > Impatient >> **Enthusiastic**
Impatient: Interested > Confident > Friendly > Casual >> **Thoughtful**

Figure 6. Attitude prediction table. Anvil labels are on the left, followed by the predicted labels sorted by geometrical distance.

The prediction table is best explained by an example. Attitude events labeled by the annotators as '*Interested*' are categorized by attitude predictor as 'Interested' (1st choice), then as 'Confident' (2nd choice), then 'Amused', et cetera, down to 'Bored' as the least likely choice. In a standard winner-takes-it-all regime, an automatic prediction algorithm would of course select the attitude minimizing the distance between the measured profile and the trained profile.

On a slightly more speculative note, one could read the interior of the prediction table as a set of 'gracefully declining' synonymy lists. Each line would then constitute a semantic theory about a particular attitude. The emerging relations between the various attitudes – 'Interested' associated with 'Amused' and 'Enthusiastic' and opposed to 'Bored', et cetera – seem to correspond fairly closely to our common sense understanding. Notice also that an intuitively weak predictor as 'casual' is also a statistically weak predictor. The suggested associations are the broadly-positive attitude qualities rather than any near-synonyms, in

contrast to the cases of e.g. 'Enthusiastic' and 'Bored' for which the suggested synonyms are much closer semantically related, and the semantic contrast to the antonyms at the other end much clearer (e.g. 'Enthusiastic' opposed to 'Bored'). In short, some generic knowledge on attitudes seems to have been transferred from the annotators to the trained model.

Conclusion

Building a conversational agent, we believe that attitude administration is indispensable. Since conversational partners are extremely sensitive to delayed or inadequate attitudinal response (e.g. showing indifference when presented with positive news, or enthusiasm when empathy was appropriate), attitude detection must be robust and effective within a short time frame. For these reasons we recommend that attitude predictions be based on acoustic measurements for F0 and Intensity for quick and robust data extraction under sub-optimal recording conditions (high-echoic and/or noisy surroundings).

An interesting off-spin of our investigation is the user-driven decision procedure in the design of the basic annotation scheme. As discussed, the annotators where allowed to select freely among all words in their vocabulary, unbiased by the academic purposes of the annotation activity. Based on our experience with the derived annotation scheme A10, we suggest this tag base for future annotation projects.

Finally, we have shown how anvil-transcribed video recordings of human-human dialogues can be used as data for training an automatic attitude detector. The trained attitude model even seemed to inherit some generic knowledge on attitudes from the human experts (the annotators) which is exactly what one hopes for in a data-driven competence model. As far as this preliminary experiment can tell, effective attitude prediction may hence be within reach even under sub-optimal recording conditions and extreme time pressure.

Acknowledgments

The research that has led to this work has been supported by the NOMCO project, which is funded by the NORDCORP program under the Nordic Research Councils for the Humanities and the Social Sciences (NOS_HS) and the European Community's Seventh Framework Programme (FP7/2007-2013), under grant agreement no. 231287(SSPNet).

References

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C. & Paggio, P. (2007). The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing. In J. C. Martin et al. (eds.) *Multimodal Corpora for Modelling Human Multimodal Behavior*. Special Issue of the International Journal of Language Resources and Evaluation. Berlin: Springer.

Aylett, M. P. & J. Yamagishi (2008) Combining Statistical Parametric Speech Synthesis and Unit-Selection for Automatic Voice Cloning; LangTech-2008, Rome.

Boersma, P., & Weenink, D. (2005). Praat: doing phonetics by computer (Version 4.3.01) [Computer program]. Retrieved from http://www.praat.org/

Henrichsen, P.J. (2012) Nature Identical Prosody; data-driven prosodic feature assignment for diphone synthesis, 4th Swedish Language Technology Conference (SLTC-2012), Lund

Kipp, M. (2001). anvil – a generic annotation tool for multimodal dialogue. In *Proceedings of Eurospeech*, pages 1367-1370.

Navarretta, C., Ahlsén, E., Allwood, J., Paggio, P. & Jokinen, K. (2011). Creating Comparable Multimodal Corpora for Nordic Languages. *Proceedings of the 18th Nordic Conference of Computational Linguistics*. Riga, Latvia, May 11-13. NEALT. pp. 153-160. See http://dspace.utlib.ee/dspace/handle/10062/16955

Nivre, J. et al. (2001). *Göteborg Transcription Standard* (GTS) 6.4. University of Gothenburg, Department of Linguistics.

Nivre J. et al. (2004). *Modified Standard Orthography* (MSO). University of Gothenburg, Department of Linguistics.

Oparin, I.; V.Kiselev; A.Talanov (2008) Large Scale Russian Hybrid Unit Selection TTS. SLTC-08. Stockholm.

Multimodal feedback expressions in Danish and Polish spontaneous conversations

Costanza Navarretta
University of Copenhagen
Copenhagen, Denmark
costanza@hum.ku.dk

Magdalena Lis
University of Copenhagen
Copenhagen, Denmark
magdalena@hum.ku.dk

Abstract

This paper presents a pilot comparative study of feedback head movements and vocal expressions in Danish and Polish naturally occurring video recorded conversations. The conversations involve three well-acquainted participants who talk freely in their homes while eating and drinking. The analysis of the data indicates that the participants use the same types of head and spoken feedback in the two languages. However, the Polish participants express more often feedback multimodally, that is through the two modalities, and they use more repeated multimodal feedback expressions than the Danish participants. Moreover, we found a stronger relation between repeated head movements and repeated speech tokens in the Polish data than in the Danish one. Our data also confirms that there is a correlation between familiarity and feedback frequency and familiarity and repetitiveness of feedback expressions as suggested in preceding studies (Boholm and Allwood 2010, Navarretta and Paggio 2012).

Keywords: multimodal communication, multimodal corpora, feedback, comparative analysis

1 Introduction

Many factors influence communication, inter alia the cultural and social situation, the communicative setting, the number of participants, their roles and relations (Allwood and Ahlsén 2008). Thus, it is important to investigate the relation between specific multimodal (speech and body) behaviors and the different communicative situations and cultures in which they occur.

While communicating, people are attentive to how their interlocutors react to what they say and, at the same time, they show their attention and provide response to their interlocutors' contributions. The function of giving (backchannelling) or receiving feedback involves both speech and body behaviors, especially head movements which are extremely frequent not only in face to face conversations, but also in interactions where the interlocutors are not able to see each other (Navarretta and Paggio 2010).

This paper compares the occurrences of multimodal feedback expressions in Danish and Polish video recorded naturally occurring conversations. The multimodal behaviors which we include in our study are head movements and speech. The conversations are comparable under many aspects comprising the number of involved participants, their age, gender and degree of familiarity. Furthermore, the settings of the conversations are similar, and the data have been coded following the same annotation scheme (Allwood et al. 2007) and annotation manual.

In this paper we focus on feedback expressed through speech and head movements in Danish and Polish conversations between three participants who are familiar with each other. We also investigate the repetitiveness of feedback expressions in these two languages, inspired by Boholm and Allwood (2010) who study repeated feedback in Swedish first encounters.

The paper is organized as follows. In section 2 we discuss background literature, and in section 3 we describe the Danish and Polish conversations and their multimodal annotations. In section 4, we present the comparative analysis of the annotated corpora, while in section 5 we account for the use of repeated multimodal feedback expressions in these data. In section 6 we discuss our results. Finally, we conclude in section 7.

2 Background

Numerous studies on video recorded conversations in different languages have shown that head movements, and especially nods and shakes, are often related to the communicative function of feedback (Yngve 1970, Maynard 1987, McClave 2000, Cerrato 2007, Paggio and Navarretta 2011a, Truong et al. 2011).

Paggio and Navarretta (2011a) analyze feedback head movements and facial expressions and their

relation to speech in the Danish NOMCO corpus of first encounters which was annotated according to the MUMIN scheme (Allwood et al. 2007). They find that 40% of the occurring head movements and facial expressions are related to feedback in the first encounters. The most frequently occurring visible feedback body behaviors in the Danish data are nods and smiles, but also tilts and forward and backward movements of the head are often related to feedback. Paggio and Navarretta (2011b) apply machine learning to investigate to which extent the various modalities can be used to predict feedback in the same data, and obtain promising results. Navarretta (2011) analyses multimodal feedback in Danish dyadic and triadic naturally occurring conversations. Her analysis confirms that head movements are the body behaviors that are most frequently used to express feedback, but she notices that also facial expressions and body postures often have a feedback function. Also in these data, nods are the most frequently occurring feedback head movement, but also side turns are often related to feedback. Finally, she finds differences in the frequencies of feedback body behaviors between the dyadic and triadic conversations.

Navarretta and Paggio (2012) investigate the effect of familiarity on the expression of verbal and non-verbal feedback in two types of conversations with participants having different degree of familiarity. They find that the degree of familiarity influences feedback body behaviors in those data. They also notice that not only the content of the conversations, but also the physical setting and the number of participants influence feedback behaviors.

For Polish, the only reported study on feedback in speech and gesture is by Malisz and Karpiński (2010). They investigate short verbal responses to instruction givers in an origami folding task. They study one- or two-syllable responses in terms of dialogue acts and intonation and analyse head movements, smiles and hand gestures co-occurring with those verbal expressions. In their data, the responses most frequently have a feedback function. Their analysis also shows that verbal feedback in Polish is often accompanied by head gestures. 90 % of nods in their data are produced with positive feedback expressions: *tak* (yes), *no* (yeah) and *mhm*, while head shakes co-occur with negative responses.

Comparative studies of video recorded first encounters indicate that there are both similarities and dissimilarities in the way different cultures express feedback through speech and especially head movements and facial expressions, inter alia (Rehm et al. 2008, Allwood and Liu 2010, Navarretta et al. 2012). In particular, Rehm et al. (2008) compare Japanese and German first encounters in order to generate culturally adapted software agents. Lu and Allwood (2010) examine the use of feedback multimodal expressions, comprising smiles, in Swedish and Chinese data and find a number of cultural specific differences. Navarretta et al. (2012) compare feedback expressing nods in Danish, Finnish and Swedish and find differences in the frequency of repeated up- and down-nods in these data. They also compare the most frequently occurring feedback speech tokens in Danish and Swedish, which are linguistically closely related, and conclude that the most frequent feedback speech tokens in the two languages correspond to each other.

Boholm and Allwood (2010) analyze repetitiveness of feedback head movements and vocal expressions in Swedish first encounters. They conclude that there is no correlation between repetitiveness in the two modalities. Furthermore, they suggest that familiarity can be a facilitator for repetition, explaining the low frequency of repeated feedback expressions in their data.

Differing from the studies which concern first encounters interactions, the conversations on which we work involve subjects who are well-acquainted. The age of the participants and the physical settings also differ from those in the first encounters. While in the first encounters corpora the participants were students recorded in a studio, our data feature participants over 50 years old recorded during free conversation at home. Similarly to Navarretta et al. (2012), however, we investigate feedback expressing head movements and their co-occurring feedback vocal expressions in comparable data. Here, however, we compare Danish and Polish, which culturally are not very distant, while linguistically they are not as strictly related as Danish and Swedish which were compared in (Navarretta et al. 2012). We also look at the relation between repetitiveness of the feedback expressions in speech and head movements, as Boholm and Allwood (2010), but we analyze this aspect in two languages.

3 The Data

In the following, we present the conversations and an overview of the annotations.

3.1 The conversations

The study is conducted on comparable video recorded conversations. The first involves three Danish native speakers while the second features three native speakers of Polish. The participants in the study are family members or near friends, thus they have a high degree of familiarity.
The pilot study covers 35 minutes interactions, approx. 17 minutes in each language. The interactions are comparable with regard to several dimensions: the participants are all female, aged 50+, and have similar degree of familiarity.

The social activity and the physical setting are also akin. The participants are video and audio recorded in their private homes while they sit around a table, drink, eat and talk freely about various subjects.

The Danish data were extracted from the MOVIN database (MacWhinney and Wagner 2010) and were orthographically transcribed and multimodally annotated as part of the Danish CLARIN-DK project (Navarretta 2011), while the Polish data were collected, transcribed and annotated under the on-going European CLARA project.

The Danish participants were filmed by one video-camera, and Figure 1 shows a snapshot from these data.

Figure1: Snapshot from the Danish data

Two cameras were used to record the Polish participants. Snapshots from the Polish conversations are in Figure 2.

The subjects were aware that they were videotaped, but the recording equipment was well incorporated in the space. The recorded Danish conversation is quite long and only the first part of it has been included in the study so that its length is approximately the same as that in the Polish conversation.

3.2 The Annotations

Both the Danish and Polish data were orthographically transcribed in PRAAT with word time stamps. The transcriptions were then imported in ANVIL where head movements were annotated according to the MUMIN scheme (Allwood et al. 2007).

Figure 2: Snapshots from the Polish conversations

This scheme provides predefined features describing the shape and the communicative functions of body behaviors. Furthermore, body behaviors can be linked to words if the annotators judge that they are semantically related. Body behaviors are multi-functional, but in the following we only focus on head movements signaling feedback. Table 1 contains the attribute and value pairs for annotating head movements.

Behavior Attribute	Behavior Value
HeadMovement	Nod, Jerk, HeadBackward, HeadForward, Tilt, SideTurn, Shake, Waggle, HeadOther
HeadRepetition	HeadSingle, HeadRepeated

Table 1: Features for Head Movements

Feedback is annotated via three features inspired by the work of Allwood et al. (1992), who define feedback as an unobtrusive behavior that has the purpose of either signaling or eliciting

signals of contact, perception and understanding. Table 2 shows the features describing the feedback function in the MUMIN scheme.

Behavior Attribute	Behavior Value
FeedbackBasic	CPU, FeedbackOther
FeedbackDirection	FGive, FElicit, FGiveElicit
FeedbackAgreement	FAgree, FDisagree

Table 2: Features describing feedback

Feedback is described by three attributes. The first attribute, FeedbackBasic is used to annotate if there is feedback and whether it involves all three aspects (Contact, Perception and Uderstanding), or only one or two of them (FeedbackOther). The second attribute, FeedbackDirection, indicates whether the behavior signals that the gesturer is giving or eliciting feedback, or whether the head movements signal both. Finally, the attribute FeedbackAgreement indicates whether the gesturer agrees or disagrees with the interlocutor.

Figure 3 and 4 show snapshots of the Danish and Polish multimodal annotations in ANVIL, respectively.

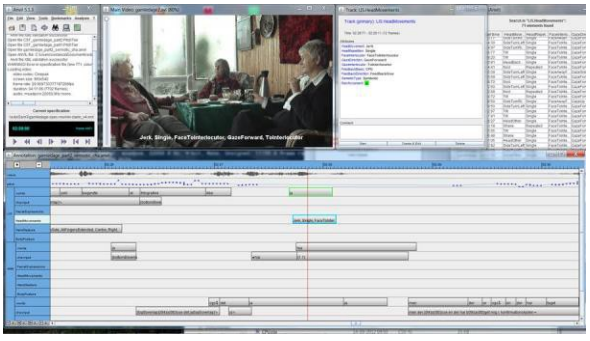

Figure 3: A snapshot of the ANVIL tool with the Danish data

4 Feedback expressions in the Danish and Polish data

In the following, we analyse feedback head movements and speech expressions in the two datasets.

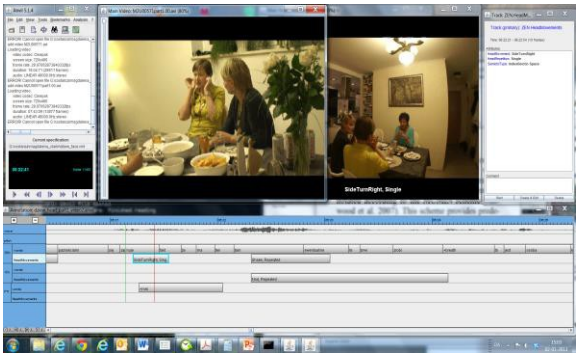

Figure 4: A snapshot of the ANVIL tool with the Polish data

4.1 Head Movements

The study comprises in total 965 head gestures. In the Danish data 476 feedback head movements are recognized, while there are 489 feedback head movements in the Polish data. The most frequently occurring feedback related head movement in both the Polish and Danish conversations is Nod (188 and 196 nods respectively), but other movements such as Tilt and SideTurn are also related to feedback in the conversations. In both the Danish and Polish data there are only 20 feedback shakes.

The occurrences of unimodal feedback head movements, that is head movements which occur without co-speech, and of multimodal head movements is in Table 3.

Feedback	Danish	Polish
Unimodal (head movements)	0.64	0.32
Multimodal (head movements and co-speech)	0.36	0.68

Table 3: Occurrences of uni- and multimodal feedback in Danish and Polish data

In Danish, 36% of the feedback head movements co-occur with speech, and 64% do not, while 68% of the feedback head movements in Polish co-occur with speech and 32% occur alone. Thus, there are much more occurrences of unimodal feedback head movements in the Danish than in the Polish data. The difference in the frequency of occurrence of unimodal and multimodal feedback in the two languages is significant. Significance in the article is measured with unpaired two-tailed t-test and the threshold for significance is p less than 0.05 while that for slight significance is p less than 0.1. In the case of unimodal feedback head movements we have

df=6 and p=0.00043, while for multimodal-feedback df=6 and p=0.00045.

In the following, we focus on nods and shakes as well as the most frequent verbal expressions with which they co-occur. As Navarretta et al. (2012) we distinguish between two types of nods: down-nods and up-nods.

Table 4 contains the occurrences of feedback nods and shakes per second in the two data sets. The table also specifies the occurrences of single and repeated nods and shakes.

Head Movement	Danish N/sec	Polish N/sec
Nod (Down and Up)	0.196	0.228
Single Nod	0.071	0.052
Repeated Nod	0.125	0.176
Down-nod	0.188	0.192
Down-nod single	0.062	0.046
Down-nod repeated	0.125	0.176
Up-nod	0.008	0.006
Up-nod single	0.008	0.006
Up-nod repeated	0	0
Shake	0.025	0.018
Single shake	0.009	0.005
Repeated shake	0.024	0.013

Table 4: Nods and shakes in the data

As table 4 indicates, the frequency of the category Nod, including Up-Nod and Down-Nod, calculated as movement per second in the two data sets is higher in Polish than in Danish, and the difference is slightly significant (df=9 and p=0.07). However, single Nod and Shake are more frequent in Danish than in Polish, while Repeated Nod and Shake are more frequent in Polish. The difference in frequency of single head movements in the two languages is slightly significant (df=9 and p=0.0875). Also the difference in frequency of repeated head movement is only slightly significant (df=9 and p=0.0793).

In the following analyses, we only consider single and repeated nods since they are the most common feedback head movements in the two corpora, and they are the head movements that were included in preceding studies on repetitiveness (Boholm and Allwood 2010) and feedback in Nordic first encounters (Navarretta et al. (2012) with which we will compare our data.

In Table 5, the percentage of unimodal and multimodal nods in Danish and Polish is given.

	Nods	Unimodal	Multimodal
Danish	Single Nods	0.70	0.3
	Repeated Nods	0.52	0.48
Polish	Single Nods	0.63	0.37
	Repeated Nods	0.34	0.66

Table 5: Unimodal and Multimodal Nods in Danish and Polish

Nearly two third of single nods and over half of the repeated nods are unimodal in the Danish, while most of the single nods are unimodal, and most of the repeated nods are multimodal in the Polish data. Thus, repeated nods are more often multimodal in Polish than in Danish and the difference is significant (df=6, p=0.039).

4.2 Feedback words

The verbal expressions with which nods co-occur most frequently in the two datasets are 'yes' expressions. In Polish they are: *tak* (yes), *no* (yah) and *aha* (yah) as well as *mm*, while in Danish they comprise *ja*, *jo* (yes), *jamen* (certainly) and *hmm*.

Shakes mostly co-occur with 'no' expressions (Polish *nie* and Danish *nej*) and with utterances containing negations, such as Polish and Danish equivalents of 'I don't know', 'I don't remember'.

5 The repetitiveness study

In tthe following, we present our study of repetitiveness of feedback nods and co-speech.

5.1 Single and repeated feedback in Polish

Table 6 shows the types of speech tokens which co-occur with the single and repeated multimodal nods in the Polish conversations.

The data indicate that 62% of the occurrences of multimodal Single Nod in Polish co-occur with single tokens, and out of these under 0.3% co-occur with the quasi-word "mhm". 35% of the multimodal single nods co-occur with more non-feedback words, and only 3% of them co-occur with repeated feedback words.

	Single feedback speech tokens	**Repeated feedback speech tokens**	**Multiple non-feedback speech tokens**
Single Nods	0.62	0.03	0.35
Repeated Nods	0.15	0.8	0.05

Table 6: Nods co-occurring with speech in Polish data

Out of the multimodal Repeated Nod occurrences, 15% co-occur with single feedback words, and out of these 16% co-occur with prolonged quasi-words such as "mhm". 80% of the repeated nods are related to repeated feedback speech tokens, and 5% are related to more non-feedback words. The vast majority of the occurrences of Repeated Nod co-occur with repeated 'yes' expressions, such as *tak tak tak* , *no no* and their combinations, e.g. *tak no no*.

5.2 Single and repeated feedback in Danish

Table 7 shows the types of speech tokens which co-occur with the single and repeated multimodal nods in these data.

	Single feedback speech token	Repeated feedback speech tokens	More non-repeated speech tokens
Single Nods	0.86	0.003	0.14
Repeated Nods	0.28	0.33	0.39

Table 7: Nods co-occurring with speech in Danish

In the Danish interactions, single 'yes' expressions are more frequent than repeated ones. 0ut of the multimodal feedback single nods, 14% co-occur with more non feedback speech tokens and only 0.3% co-occur with repeated feedback words, and 86% of the multimodal single nods co-occur with single feedback expressions.

More complex is the situation for repeated feedback expressions. Repeated nods co-occur with multiple non-repeated words in 39% of the cases, while 33% of the multimodal repeated nods co-occur with repeated feedback words, and finally 28% co-occur with single feedback words.

5.3 Danish and Polish

The analysis of the Danish and Polish data indicates that the feedback multimodal behavior in the two languages is not the same. The difference between the occurrences of repeated nods with repeated feedback speech tokens and multiple feedback speech tokens in the two languages is significant (df=6 and p=0.031 and p=0.002, respectively). The difference in frequency of single nods and single or multiple non-repeated speech tokens is also significant (df=06 and p=0.049 and 0.044 respectively), while the remaining differences are not significant. Summing up, the Polish participants use significantly more often repeated nods with repeated or multiple feedback words than the Danish participants, and the Danish participants use more often single nods with single speech tokens.

6 Discussion

Our study shows similarities between Danish and Polish speakers in terms of the type of feedback head movements and spoken expressions. The most common feedback head movements are nods, but also shakes, tilts and side turns are used to express feedback. Furthermore, feedback head movements co-occur with similar feedback speech tokens in the two datasets. Our analysis confirms the findings of preceding monolingual studies on Danish and Polish (Paggio and Navarretta 2011a, Malisz and Karpiński 2010), as well as studies of feedback on other languages (e.g., Lu and Allwood 2010), Navarretta et al. 2012, Rehm et al. 2008).

The frequency of feedback head movements is nearly the same for Danish and Polish speakers, but the Polish participants nodded more than the Danish ones. Given that the Polish nods are often repeated increases the difference between the two languages indicating that the Polish participants gave more body feedback in these data than the Danish participants . In general, the frequency of feedback head movements in both the Danish and Polish data is higher than in first encounters data reported in Navarretta et al. (2012). This is in accordance with the results in the study by Navarretta and Paggio (2012), who found that head gestures rate increases with familiarity and who compared this effect to the increase in speech flow among well-acquainted subjects (Campbell 2007).

Our analysis also shows a clear difference between Danish and Polish speakers in the occurrence of unimodal and multimodal feedback expressions. Overall, in the Danish interactions most feedback head movements are unimodal, while in Polish they more often co-occur with speech. While in both languages single nods are usually not accompanied by spoken feedback, the repeated nods are multimodal in Polish more often than in Danish. Differences in terms of uni- and multimodality of feedback expressions were observed for Swedes and Chinese by Lu and Allwood (2010). In the first encounters, although both groups gave multimodal feedback much more often than non-verbal unimodal one, the Swedes used gestural unimodal feedback twice as often as did the Chinese. It should be tested further whether this is a general characteristic of the different languages or it depends on other factors, such as the content of the conversations.

In both Danish and Polish data single nods co-occur most often with single words. While for Danish there is no clear-cut pattern for repeated nods in these data, for Polish repetitiveness in nodding co-occurs with repetitiveness of feedback words. Our data also show that in Polish repeated multimodal feedback expressions are more frequent than in Danish even though the degree of familiarity between the participants is the same and despite the fact that there are individual variations in both datasets. This suggests a difference between Danish and Swedish on the one hand and Polish on the other hand.

Overall, repeated feedback nods occur more frequently in our corpora of well-acquainted Danish and Polish speakers than in the first encounters corpora for Danish, Finnish and Swedish (Navarretta et al. 2012). The explanation might be that repetition is facilitated by familiarity, as suggested by Boholm and Allwood (2010).

To which extent the differences reported in our study depend on familiarity, the setting, the social activity and age of the participants or on the languages should be investigated further. Since our datasets are small, more data should be analyzed.

7 Conclusions

In this study, we compared feedback head movements and spoken tokens in Danish and Polish video recorded conversations between well-acquainted participants.

The analysis of the data indicates both similarities and dissimilarities in the two datasets. The types of multimodal feedback in the two corpora are similar, but the Polish subjects use more frequently feedback nods than the Danish subjects. There are significantly more repeated and multimodal feedback nods in Polish than in Danish and there is a stronger correlation in Polish between repetitiveness of feedback nods and speech tokens than in Danish. No correlation between feedback repetitiveness of nods and of speech tokens was found in Swedish by Boholm and Allwood (2010).

Finally, our data indicate that the feedback expressed by head movements is more frequent in the Danish and Polish conversations between people who know each other well, than in first encounters corpora (Navarretta et al. 2012), confirming preceding studies (Navarretta and Paggio 2012) which suggest that the level of familiarity influences the frequency of feedback expression.

Since our corpora are small and regard only one communicative situation, the results of our analysis should be tested on more data and on more types of conversation.

In future, we will also investigate to which extent the differences between uses of single and repeated feedback behaviors are language related or are connected to familiarity as hypothesized by Boholm and Allwood (2010).

Acknowledgments

The work described in this paper was partly funded by the Danish Research Council for the Humanities (VKK and NOMCO project) and the EU CLARA project.

References

Allwood, J. & Lu, J. (2010). Chinese and Swedish multimodal communicative feedback. Presented at the *5th Conference on Multimodality*. Sydney,1-3 December 2010, pp.19-20.

Allwood, J. & Ahlsén, E. (1998). Learning how to manage communication, with special refer- ence to the acquisition of linguistic feedback. *Journal of Pragmatics*, 31, 1353-1389.

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 Martin, J. C. et al. (eds) *Multimodal Corpora for Modelling Human Multimodal Behaviour*. Special issue of the International Journal of Language Resources and Evaluation. Springer.

Boholm, M. & Allwood, J. (2010). Repeated head movements, their function and relation to speech.

In Kipp, M. et al. (eds.) *Workshop on Multimodal Cor- pora: Advances in Capturing, Coding and Analyz- ing Multimodality*. LREC 2010.

Campbell, N. (2007). Individual Traits of Speaking. Style and Speech Rhythm in a Spoken Discourse. In. *Proceedings of the COST 2102 Workshop*, pp. 107-120.

Cerrato, L. (2007). *Investigating Communicative Feedback Phenomena across Languages and Modalities*. PhD thesis, Stockholm, KTH, Speech and Music Communication.

McClave, E. (2000). Linguistic functions of head movements in the context of speech. Journal of Pragmatics, 32, 855-878.

MacWhinney, B. & Wagner, J. (2010). Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository. In: *Gespraechsforschung*, Vol. 11, 2010, 154-173.

Malisz, Z. & Maciej Karpiński. M. (2010). Multimodal aspects of positive and negative responses in Polish task-oriented dialogues. *Proceedings of Speech Prosody 2010*. Chicago.

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

Navarretta, C., Ahlsén, E., Allwood, J., Jokinen, K. & Paggio, P. (2012). Feedback in Nordic First-Encounters: a Comparative Study. In *Proceedings of LREC 2012*, May 2012, Istanbul, Turkey, pp. 2494-2499.

Navarretta. C. (2011). Annotating Non-verbal Behaviours in Informal Interactions. In Esposito, A, Vinciarelli, A., Vicsi, K., Pelachaud, C. & Nijholt, A. (eds.) *Analysis of Verbal and Nonverbal Communication and Enactment: The Processing Issues*, LNCS 6800, Springer Verlag, pp. 317-324.

Paggio, P. & Navarretta. C. (2011a). Feedback and gestural behaviour in a conversation- al corpus of Danish. In P. Paggio et al. (eds.) *Pro- ceedings of the 3rd Nordic Symposium on Multi- modal Communication* May 27-28, 2011, University of Helsinki, Finland, NEALT Proceedings Series vol. 15, 2011, pp. 33-39.

Paggio, P. & Navarretta, C. (2011b.) Learning to classify the feedback function of head movements in a Danish Corpus of first encounters. In Proceedings of ICMI 2011 Workshop *Multimod- al Corpora for Machine Learning: Taking Stock and Road mapping the Future* , Alicante, Spain November, 8 pages.

Navarretta, C. & & Paggio, P. (2012) Verbal and Non-Verbal Feedback in Different Types of Interactions. In *Proceedings of LREC 2012*, May 2012, Istanbul, Turkey, pp. 2338-2342.

Rehm, M., Nakano, Y., André, E. & Nishida. T. (2008). Culture-Specific First Meeting Encounters between Virtual Agents. In *Proceedings of the 8th international conference on Intelligent Vir- tual Agents* (IVA '08), Prendinger (Eds.). Springer-Verlag, Berlin, Heidelberg, pp. 223–236.

Truong, K., Poppe, R., de Kok, I. & Heylen, D. (2011). A Multimodal Analysis of Vocal and Visual Backchannels in Spontaneous Dialogs. In *Proceedings of Interspeech 2011*, International Speech Communication Association, France, pp. 2973-2976.

Showing interest during first acquaintance

Gülüzar Tuna
University of Gothenburg
Gothenburg, Sweden

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

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

Abstract

The goal of this paper is to investigate the expression of *interest* as an affective-epistemic attitude in first acquaintance conversations. The study presents an analysis of how interest, as one of several different affective-epistemic attitudes, is shown through multimodal expressions between strangers meeting the first time. The results show that interest can be shown both as a single attitude and in combination with other attitudes. Interest as a single attitude occurs more often. The findings indicate that multimodal expressions connected with showing interest mainly include five body movements/gestures; gaze, head movements, holistic face, hand movements and body postures. Gaze occurs 70 times in the case of interest as a single attitude, and in combination with other attitudes 7 times. The corresponding numbers for head movements are 50 times for a single attitude and 7 times for a combined attitude. While smiles (classified as a general face movement) occur 24 times with a single attitude of interest, it occurs 9 times, in a combination with other attitudes. The difference between a single and a combined attitude was less pronounced regarding hand movements – 16 times for a single attitude and 8 times for a combined attitude. The numbers for body postures expressing attitudes were 11 for a single attitude versus only 1 for a combined attitude. The study suggests that there are signs of differences between how women and men show interest, even when taking into account that the number of women in the study exceeds the number of men. The difference between sexes regarding showing interest is bigger concerning the combined attitude type than the single attitude type.

Keywords: showing interest, attitude annotation, first acquaintance

Introduction

Communication is one of the tools humans use when socializing and interacting with others in any social community. Communication in general, but particularly face-to-face communication is multimodal, usually involving both speech and visible bodily gestures (Allwood, 2002:12). Further, bodily communication that is perceived visually has a central place in human communication. Navarretta et al. (2011) include attributes such as *facial expressions*, *gaze* and *hand gestures*, *head movements* and *body postures* as particularly significant for the expression of the emotional states of an individual (See also Meeren et al. 2005:16518).

In order to analyze how interest is expressed and perceived by interlocutors, we have to take into account the relationship between expressions of emotions using vocal verbal utterances, facial expressions and other bodily movements i.e. body posture and hand movements. It can also be important to take into account "touching and scratching" involving "self-manipulation", together with changes in body position, head and hand movements (Mehrabian, 1968:54). All of these types of bodily expressions can express the underlying positive or negative attitudes of a person (Ibid:54).

One of the motives for writing this paper is that in our data of audio/ and video recorded first encounters between two university students, *interest* turns out to be the most commonly expressed affective-epistemic attitude, followed by being *certain, casual (relaxed and/or informal), amused* and *reassuring*. Table 1 shows the five most

common affective epistemic attitudes that occur in our study. The table shows that an attitude can be simultaneously expressed by several bodily features, e.g. 171 particular features were involved in the 86 occurrences of the expression of the attitude of interest.

Affective Epistemic states - Single attitude	***Total Frequency of particular bodily features involved in the multimodal expression of an attitude*[1]**	***Total Frequency of the attitude expressed by participants***
Interested	***171***	***86***
Certain	96	44
Reassuring	74	44
Casual	72	33
Amused	77	29

Table 1: Most common affective epistemic attitudes in the analyzed first acquaintances

Showing interest

In this paper we focus on how *interest* is expressed multimodally, combining a study of the utterances and gestures made by the participants. As already mentioned, this can be done either by expressing interest alone or by expressing it in combination with other attitudes, for example showing *interest* and being *reassuring* at the same time. In both cases *interest* can be expressed using vocal verbal means, facial expressions or body movements.

Whether to include interest as an emotion or not has been a topic of discussion. Ortony et al. (1990) reports that some emotion theorists (Plutchik, Ekman and Arnold) claimed that whether interest can be considered an emotion is unclear since interest is something more of "a cognitive state and not an affective one" while other theorists like Frijda, Tomkins and Izard have called interest a "basic emotion" since interest "exhibits a distinctive facial expression" (Ortony et al. 1990:318).

There are several theories of emotion and also several lists of basic emotions, e.g. Plutchik lists 8 basic emotions, i.e. *joy, trust, fear, surprise, sadness, disgust, anger* and *anticipation* (Ortony et al. 1990:316). Ekman only identifies 6 emotions, i.e. happiness, anger, fear, sadness, surprise and disgust/(contempt) (Ibid:316). The lists vary between different theories. As already mentioned, some emotion theorists such as Frijda, Izard and Tomkins have included interest as an emotion, see Table 2 for an overview.

Basic emotions defined by theorists (based on Ortony & Turner (1990 & http://www.deepermind.com/02clarty.htm)	
Theorist	Basic Emotions
Plutchik, R. (1927 –2006)	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise
Ekman, P (1934-)	Anger, disgust, fear, joy, sadness, surprise
Arnold, M. B (1903–2002)	Anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness
Frijda, N. H. (1927-)	Desire, interest, happiness, surprise, sorrow, wonder
Tomkins, S. S. (1911 –1991)	Anger, contempt, disgust, distress, fear, interest, joy, shame, surprise
Izard, C (1924-)	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise

Table 2. A sample over emotion theorist's emotion position.

Showing interest as expressing a stance

Kiesling (2009) defines a stance as "a person's expression of their relationship to their talk /…/ or a person's expression of their relationship to their interlocutors (Kiesling, 2009:272/1).

A stance usually occurs when two (or several) contributors interact and communicate with each other face-to-face. In doing so, the contributors react to each other's facial and bodily attributes, which is a basic feature of a co-constructed stance (Allwood et al. 2012). Biber et al. (1999) in Clift (2006) have suggested 3 types of stance: *epistemic*, *affect* and *manner* (Clift 2006:579). Epistemic stances include certainty, doubt and particular viewpoints, while affective stances include more emotional attitudes and manner stances concern style of communication (Ibid:579).

Allwood et al. (2011:2) define communicative stance as an "attitude which is expressed and sustained interactively in communication, in a unimodal or multimodal manner". The difference between attitudes and stances, following this definition would thus be that a stance, but not an attitude, needs to be overtly expressed. The term *unimodal* is used when the stance is only vocal verbal or only gestural. *Attitudes* can be both of an

[1] The number stands for the total frequency of the occurring multimodal features i.e. gaze, face, head, hand, body.

epistemic and an affective kind. Epistemic stances can be exemplified by *believing* and *being bored* while expressing happiness is an example of an affective stance. However, there are many attitudes that are both "epistemic and affective" like *feeling of certainty* or attitudes that are related to expectation like *surprise* (Allwood et al. 2012).

Purpose

As already stated, the overall aim of this paper is to investigate interest as an affective-epistemic attitude. We do this by identifying different affective-epistemic attitudes shown between strangers when they meet the first time and by describing the multimodal features of these attitudes. The research questions for this paper are four: First, does the expression of interest express a single attitude or is it combined with other attitudes? Second, what kind of multimodal expressions are used in showing interest? Third, are there any differences between the sexes in how and how much interest is shown? And fourth, does showing interest qualify as a stance?

Data analysis and method

Transcription of audio

The data analysis in this paper is based on 14 video-recordings of first acquaintances between strangers, recorded at the University of Gothenburg during the period 2009 -2010.

The participants were Swedish university students from different disciplines at the University of Gothenburg who met the first time. In total, 14 of the participants were female (two of them occur twice in the recordings) and 10 were male (two of them also occur twice in the recordings). The 24 students were randomly selected in order to get as unbiased data as possible. The approximate length of each video recording ranges from 6 to 8 minutes. The video recording transcriptions were made using the Gothenburg Transcription Standard (GTS) with The Modified Standard Orthography version 6 (MSO6), (Nivre, et al. 1999:3). The transcriptions and recordings were imported into the Anvil-program (Kipp 2001). and annotated using the MUMIN annotation scheme (Allwood et al. 2007). A randomized selection, involving numbering all utterances and then using a numerically randomizing program to pick out sequences of 3 utterances in the transcription, exemplifying different affective-epistemic states, was used to select what was annotated.

In total, 97 out of 333 utterances were coded as expressing *interest*. For each expression of interest we determined exactly where the utterances started and ended on the video recordings. Each sequence of 3 utterances was for this purpose imported into Anvil for detailed analysis. The selected sequences were later on classified more carefully regarding what affective-epistemic states and speech acts they expressed. Besides classifying the utterances expressing a particular affective-epistemic attitude, we also classified the utterances immediately preceding these utterances in order to have an idea of the conversational context.

The annotation of video

The annotations are based on the MUMIN annotation scheme which is a tool for studying gestures in interactive communication (Allwood, et al. 2007:274) and in this study it has been used to annotate the communicative gestures Navarretta, et al. 2011:155) (i.e. gestural movements expressed through five manners) together with the vocal verbal expressions connected with these gestures, in the recorded video sequences.

In order to synchronize the annotated video files with the transcriptions, the computer program Anvil was used for annotating the video recordings (Kipp 2001). All in all, the empirical data for the study comprises 62 video sequences. The lengths of the video sequences vary from 4 to 19 seconds.

In this paper, we have included the following movements for the head: down nod, up nod, head backward, head forward, tilt, side turn, waggle, shake, and head other single movement or repetitive movement.

For facial expressions: smile, laughter, and other. Eyebrows are coded as: frown and raised. Eyes are coded as: x-open, close both and close one.

Gaze direction is coded as: gaze forward, gaze backward, gaze up, gaze down, gaze

side, gaze direction and other. Gaze to interlocutor or away from interlocutor.

Hand gestures are coded as: handedness (single, both), hands other. Body posture or body direction: body forward, body backward, body up, body down, body side, body direction other, body directed to the interlocutor or body away from the interlocutor, Shrug and shoulders other have also been coded.

An illustration of an utterance that expresses interest as an affective-epistemic attitude can be found below.

B: < ja så du läser / kursen vi{d} sidan om >
< yeah so you study /take the course on the side >
< eliciting > ; < gaze at interlocutor > ;
< attitude: interested >

(Transcription conventions:
< > = commented sequence of transcription and comments;
/ = short pause,
{ } = non-pronounced letter)

Result and discussion

In section 4.1, below we will present and discuss the results of identifying expressions of interest and the multimodal expressions through which this is done in conversations between university students meeting the first time.

General observations

Single attitude or combined attitudes

Our findings show that interest can occur alone and combined with other attitudes. By a single attitude we, thus refer to the case when no other attitude than interest is expressed. By a combined attitude we refer to the case where interest is expressed in combination with a second attitude. Interest as a single attitude occurs more often than as a combined state, i.e. 86 times, whereas interest combined with other attitudes occurs 11 times, as shown in tables 1 and 3.

Combined Affective-epistemic states	**Total 11**
Interested + surprised	4
Interested + casual	2
Interested + uncertain	2
Interested +ironic	1
Interested +unconfident	1
Interested +astonished	1

Table 3: Frequency of utterances expressing interest

Interest occurs mostly together with the *speech act* of *feedback* (both alone and multi-functionally) in combination with other speech acts The second and third most common speech act is *question* were interest occurs 14 times and *elicitation* were interest occurs 5 times.

The feedback utterances often involve overlaps, e.g. utterances such as [< m >] (occurs 6 times), [< m:>] (4), [< {j}a >] (4), [< okej >] (3), [< mhm >] (2), [< ja >] (2), [< {j}a >] (2), [< {j}a / {j}a >] (2).

Interest is also expressed in longer comments like: < okej men då då e de{t} fler än du som e jobbar me{d} precis samma sak / elle{r} e du unik på de+ [fronten så att säga >], <<m > // < va:{d} men då e1 > har du nån släkt kvar i < danmark > de{t} ha{r} du > << va:{d} e1 / m > men du va{r} lite{n} när du flyttade till < värmland > sa [du] > *<okay but then then are there more people like you um working with just the same thing / or are you unique on the + [front so to speak>] << m> / / < what but then um> do you have any family* left in *<denmark> have you > << what um / m > but you were small when you moved to <värmland> did [you] say>*.

We also have utterances expressing interest through direct comments such as < du då > (*you then*), <du > (*you*), < du har gått på andra] > (*<you have gone to the other>*),< då har du gått på < område tre > skolan också > < eller{r} >> *(<then you gone to <area three> school too> <or >>*), < har du jobbat hä{r} länge > *(<have you worked here for long time >*), < < ha{r} [du > några] syskon > *(<< do [you have > some] siblings >*), < ha{r} du plane{r} på att flytta < tillbaks till < värmland >>> *(<< do you have plans to move <back to <värmland >>>*), << å du plugga{r} inte här > / < eller >> *(<< and you don´t study here> / <or >>*). In general, the expressions of single

attitudes of interest are slightly longer than are those of combined attitudes.

Multimodal expressions connected with showing interest

The results indicate that the multimodal expressions connected with showing interest besides the vocal verbal part include five types of body movements/gestures; gaze, head movements, general face movements, hand movements and body posture, as shown in Table 4.

	Single affective-epistemi state of Interested 86			**Combined affective-epistemic state of Interested 11**		
Gesture	**Fem**	**Male**	**Tot**	**Fem**	**Male**	**Tot**
Face	66%	34 %	70	5	2	7
Gaze	66%	34%	24	8	1	9
Head	66%	34%	50	6	6	12
Hand	37%	67%	16	6	2	8
Body	27 %	73%	11	-	1	1
Total	104	67	171	25	12	37
	61%	39%		67%	33%	
	171			37		

Table 4: Frequency of Multimodal Communicative attitude.

The corresponding numbers for *head movements* are 50 times for a single state and only 7 times for a combined attitude. *Smiles i.e. holistic face* occur as much as 70 times expressing interest as a single attitude and 7 times as a combined attitude. The difference between a single and a combined attitude was less pronounced regarding *hand* movements 16 times for a single attitude and 8 times for a combined one. The frequency for *body posture* regarding a single attitude that was 11 times, versus only 1 time for a combined attitude.
Face (smile) is the most common multimodal feature, followed by *head* and *gaze and hand* and *body.*

Differences between genders

Our study suggests that there are differences between how women and men show interest, also when compensating for the fact that the number of women in the study exceeds the number of men. The results in the table have to be compared to the expected difference between men and women (57% vs. 43%) due to the fact that there were more female participants.

The results indicate that women show interest slightly more often than men, and that the difference between the genders is larger concerning the combined attitude type than for single attitudes. They also indicate a slight difference in how interest is expressed. Women use face(smile), gaze and head slightly more and men hand s and body.

As shown in Table 4, a single attitude of interest was shown 104 times and 25 times in combinations of multimodal attributes, by females and 67 times as a single attitude of interest and 7 times in combinations of multimodal attributes, by males.

Is showing interest a stance?

The participants who participated more than once enable us to explore the question of whether the expression of interest is a personality feature and thus a dispositional stance or if its occurrence is more dependent on the interlocutor and thus coconstructed.

As mentioned above, the participants in this study were 14 females, two of whom participated twice in the recordings and 10 males, where two participated twice in the recordings. This means that 57% of the recorded participants were female and 43% male. The reoccurring participants are two females, F1 and F2 and two males M1 and M2. If a particular attitude is shown frequently, in a similar way, we take it as an indication that what is expressed is a stance and perhaps a personality feature. If an attitude is shown less frequently, we take it as a sign that the expression of the attitude is perhaps not a personality feature and thus perhaps not a dispositional stance.

Using this criterion, our study indicates that showing interest can be considered a dispositional stance.

Showing interest - Difference between two males and two females								
Use of Behavioral feature	M1 -M2		F1-M2		F2-M1		F1-F2	
Single feature	4	9	7	11	9	4	13	
Combined features								2
Total	4	9	7	11	9	4	13	2

Table 6: Frequency of showing interest - Difference between only four participants, i.e. M1, M2, F1 and F2.

The reason is that when we compare the four encounters, M1 meeting M2, F1 meeting M2, F2 meeting M1, F1 meeting F2, as shown in Table 6, there is a similarity in how interest is shown by M1, M2 and F1 concerning both expression of interest both as a single and/combined attitude.

Conclusions

In this paper we have explored four questions i.e. does interest usually occur as a single attitude or is it combined with other attitudes? What kind of multimodal expressions are connected with showing interest? Are there any differences between the sexes in how much and how interest is shown? And lastly, is showing interest a stance?

Through our analysis we have illustrated that expression of *interest* is expressed both vocal verbally, by facial expressions and by bodily movements and gestures.

The expression of *interest* can occur both as a single attitude and in combination with other attitudes. A single attitude of interest occurs more often than a combined attitude. As shown in Table 4, the multimodal expressions connected with showing interest, besides the vocal verbal aspect, include five types of bodily expression features. Among these, *face*(smiles) is the most common, followed by, *head*, gaze *hand* and lastly *body*.

Turning to gender differences, females expressed interest slightly more often than males both as a single attitude and even more in combination with other attitudes. This is so also when we compensate for the fact that the number of women in the study is greater than the number of men i.e. 14 females and 10 males (where four of them participate twice in the video recordings). The study also suggests that there are differences in the way that women and men show interest. Since expression of interest often has long duration and is expressed repeatedly by the same person, our results suggest that showing interest can be a dispositional stance based on an emotional attitude, often expressed as a pattern in face, the body and hand, which according to (Scherer, 2007:158) also typically characterizes emotional features.

Limitation of research

There are some limitations in our study. First, the number of participants is small and only four of them occur twice. Thus our claims can only be of a tentative nature.

Secondly, we have not sufficiently analyzed how the background of the participants influences their communication. They are Swedish university students. Their cultural background, age and length of study at university may all play a role for their behavior.

And lastly, power and dominance are other features that could be further analyzed, i.e. Is there a “leader” and a “follower” in the conversation? Such relations may well influence the way the participants talk, respond and express *interest*.

Acknowledgments

This research has been supported by the European Community´s seventh Framework Programme (FP7/2007-2013) under grant agreement no.231287(SSPNet) and by the project: "Multimodal Corpora in the Nordic Countries" under the NORDCORP program, the Nordic Research Council for the Humanities and the Social Sciences (NOS-HS).

References

Allwood, J. (2002). Bodily communication dimensions of expression and content. In Granström, B,. House, D. & Karlsson, I. (eds) *Multimodality in language and speech systems*. Boston: Kluwer Academic Publishers, pp. 7–26.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C. & Paggio, P. (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. IN C. Martin et al. (eds) Multimodal Corpora for

Modelling Human Multimodal Behaviour. *Language Resources & Evaluation* 41, 273–287.

Allwood, J. & Lu, J (2011). Unimodal and Multimodal Co-activation in first encounters: A case study. In P. Paggio et al. *Proceedings of the 3rd Nordic Symposium on Multimodal Communication*, May 27-28, 2011, University of Helsinki. NEALT Conference Proceedings Series, Vol 15, 2011, pp. 1-9.

Allwood, J., Chindamo, M. & Ahlsén, E. (2012). Some suggestions for the study of stance in communication. *IEEE SocialCom*, Amsterdam 2012.

Clift, R. (2006). Indexing stance. Reported speech as an interactional evidential. *Journal of Sociolinguistics* 10/5, 2006, 69-595.

Ekman, P. & Kelner, D. (2008). Facial Expression of Emotion. In Lewis, M. a& Haviland-Jones, H. (eds). *Handbook of Emotions* 2nd edition. New York, Guilford Publications.

Griffiths, P. E. (2003). Basic Emotions, Complex Emotions, Machiavellian Emotions. In Hatzimoysis, A. (Ed.),*Philosophy and the Emotions* Cambridge, Cambridge University Press, pp. 39-67.

Kipp, M. (2001). Anvil – A generic annotation tool for multimodal dialogue. In *Proceedings of Euospeech 2001*, pp.1367-1379.

Kiesling, Scott (2009). Style as stance: Stance as the explanation for patterns of sociolinguistic variation. In Jaffe, A. (Ed.) *Stance: Sociolinguistic perspectives*. Oxford: Oxford University Press. Pp.171-194.

Mehrabian, A. (1968). *Communication without words.* No.P-89. Psychology today reprint series.

Meeren, H. K. M. (2005). *Rapid perceptual integration of facial expression and emotional body language*. Cognitive and Affective Neuroscience Laboratory, Tilburg University.

Navarretta, C., Ahlsén, E., Allwood, J., et al. (2011). Creating Comparable Multimodal Corpora for Nordic Languages. *Proceedings of the 18th Nordic Conference of Computational Linguistics*. Riga, Latvia, May 11-13. NEALT, pp. 153-160.

Nivre, J. & Allwood, J. (eds.) (1999). *Göteborg Transcription Standard.* Department of Linguistics Göteborg University. Version 6.4.

Ortony, A., Turner, T. J. (1990). What´s basic about emotions? *Psychological Review*. Vol. 97, No. 3, 315-331.

Scherer, K. R. & Ellgring. H. (2007). Multimodal Expression of Emotion: Affect Programs or Componential Appraisal Patterns? *Emotion*. American Psychological Association. Vol. 7, No. 1, 158–171.

http://www.deepermind.com/02clarty.htm 14:12, 2012-11-02

Prosodic Expressions of Emotions and Attitudes in Communicative Feedback

Gustaf Lindblad
University of Gothenburg
Gothenburg, Sweden
gustaf.lindblad@gu.se

Jens Allwood
SCCIIL (SSKKII)
University of Gothenburg
Gothenburg, Sweden
jens.allwood@gu.se

Abstract

This study investigates the communication of affective-epistemic states (AES) by means of prosody in vocal verbal feedback. The study was conceived as a pilot study to test certain methodological queries about investigating prosody as a part of multimodal communication, and as part of feedback in particular. We find that our method, with some slight adjustments, seems adequate to answer several interesting questions about prosodic features of feedback.

Keywords: prosody, emotion, attitude, communicative feedback

1 Introduction

We communicate emotions and attitudes using a few different means, perhaps the most well known and most studied being facial expressions. In vocal communication we also modulate different aspects of our speech, in particular voice quality and prosody, to convey our internal states.

Like facial expressions, speech modulations can both be voluntary and involuntary, as well as both innate and learned. In all cases, they need to be shared to a certain extent among their users to function as markers of internal states. In other words, there needs to exist some systematic mapping between expression and internal state that guides the interpretation of the expression. This mapping has not been subject of as much study for prosody as is the case for facial expressions.

The differences between emotions, attitudes and several other internal mental states are not always clear. Basically, the grounds for differentiating between these mental states are the experience and behaviour of the agent. Anger, happiness and fear are typically described as emotions, whereas being interested, sceptical or condescending are typically designated as attitudes. Similar internal states such as being hungry or tired are not generally regarded as either emotions or attitudes, but are in many ways similar to emotions and attitudes. All such mental states can be construed as having the function of making an agent more likely to behave in a particular way. For these reasons, we chose to use the inclusive term 'affective-epistemic state' (AES), to refer to any such internal mental state that can manifest itself both in an experience and a behaviour. This article is focused on the vocal expression of such affective-epistemic states.

Communicative feedback can be defined as unobtrusive vocal and bodily expressions informing an interlocutor about the feedback giver's 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 (e.g. Allwood, 1988; Allwood et al, 1992). Examples of vocal verbal feedback expressions in Swedish are words such as *m* ('m'), *ja* ('yes'), *nä* ('no'), and *okej* ('okay'), phrases such as *jag förstår* ('I see'), and repetition of what the interlocutor just said. Feedback often reflects the speaker's attitude or emotion with regard to the topic, interlocutor or context in general. Since vocal verbal feedback often consists of short one-word utterances of a limited number of words, vocal verbal feedback is a good candidate for also studying the communicative functions of prosody.

There are two basic features of speech that are modulated to convey affective-epistemic states: voice qualities and prosody. It is not possible to disassociate the two completely, as they can be partly dependent on each other. Voice qualities can be such thing as a raspy or nasal voice, and prosody correlates to intonation and stress, and often translates to what in common speech is called the tone of voice, e.g. a sharp or sarcastic tone.

Prosody is typically measured through three aspects of the vocal signal: pitch, intensity and duration. The pitch is usually identified with the fundamental frequency of the voice (F0) and measured in Hz, intensity is the volume of the voice measured in dB, and the duration is simply measured in milliseconds. The pitch and intensity varies along the duration of an utterance and can be visualized as a curve.

Pitch and intensity of an utterance can be measured, for example, by maximum, minimum or mean value, or by the shape of the curve, such as rising, falling or flat. The intensity of any utterance will always feature a rise in the beginning and a fall in the end as a natural part of the sound.

2 Method

The affective-epistemic states that we chose to use for the recordings were selected with the intention of getting a substantially different array of differently sounding samples. This list is not intended to be thought of as exhaustive or to reflect any specific opinion or statement on behalf of the authors. The affective-epistemic states have been translated into English, with the original Swedish word in brackets.

determined (bestämd)	surprised (överraskad)
factual (konstaterande)	interrogative (frågande)
hurried (stressad)	bored (uttråkad)
happy (glad)	uncertain (osäker)
irritated (irriterad)	neutral (neutral)

Table 1. List of the affective-epistemic states that the speakers were instructed to produce.

We recorded two different persons who were instructed to produce the five most basic feedback words in Swedish (‘ja’, ‘nej’, ‘m’, ‘okej’, and ‘jo’) in a way that they felt captured the different affective-epistemic states. Both persons were male, one in his sixties, and the other in his forties. They produced every feedback word three times in each affective-epistemic state (in total yielding a library of 300 samples).

The samples were recording using a studio grade condenser microphone (ADK A-51, fixed cardioid) in close proximity to the mouth of the speaker, with a bit depth of 24 and a sample rate of 48 kHz, in all giving a linear and clear (high signal-to-noise ratio) recording of the signal.

We selected one sample of each affective-epistemic state produced by either speaker, resulting in 20 samples in total, which we played back to an audience of 25 first year students in cognitive science. They were each given a form with 20 numbered lines and asked to write down what emotion or attitude they intuitively felt was being expressed by each utterance. The samples were played in random order with regards to the affective-epistemic state, but not in terms of speaker. That is, first the ten samples of the first speaker were played back in random order, and then the ten samples of the second speaker were played in random order.

The prosodic features of the samples were analysed using Praat (ref). Duration was measured in 10ths of milliseconds. Intensity was measured as the mean intensity between the start- and stop-point of the utterance. Peak intensity was also measured but not used for this study. Pitch was measured in several ways: the shape of the f0-curve was categorized into one of eight categories (Boholm & Lindblad, 2011). The mean pitch of the utterance was measured, as well as the averages of the highest and lowest 30 ms portions. The difference between the highest and lowest pitch of the utterance is calculated using the following formula ((hi pitch)-(lo pitch) / (hi pitch)), which gives a value between 0 and 1, where higher value means greater difference. The label we have given to this value is “pitch difference”.

Category	Description
Flat	The pitch curve describes a more or less flat f0, neither rising nor falling, with fluctuations smaller than 5%
Rise	The curve is increasing throughout
Complex-rise	The curve has an overall trend of increase, but contains smaller anomalies or fluctuations
Fall	The curve is decreasing throughout
Complex-fall	The curve has an overall trend of decrease, but con-

	tains smaller anomalies or fluctuations
Fall-rise	The curve has a distinct u-shape
Rise-fall	The curve has a distinct arced shape
Complex	The curve describes a more varied shape and does not fall in any category

Table 2. Categories of pitch curve shapes.

There are valid concerns that experimental speech has a low ecological validity for drawing conclusions about the intrinsic qualities of speech, and that ideally natural speech should be used in research such as this. We share this concern, but we find that we cannot get good enough quality recordings of natural speech to make reliable measurements. When measuring the fine qualities of prosody, it is very important to have a good control of the signal so you know what you are actually measuring: most importantly, the signal needs to be isolated from other sources of sound, and the subject needs to have a fixed distance and angle to the microphone. If these conditions are not met, the pitch measures often become distorted or void, and the intensity measures cannot be compared to each other.

The recordings in our study are experimental and the actors emulate the emotions, but the reactions of our panels are genuine. We believe that this approach gives us a suitable balance between ecological validity and clear data.

3 Results

Twenty-five respondents heard twenty samples each and were asked to write down their interpretation of the AES in the sample in Swedish (free choice), resulting in 500 answers. After correction of spelling errors and grouping synonyms and derivations together (e.g. if one respondent had written 'happiness' and another 'happy' these would be sorted in the same category, i.e. 'happy'), we found that the twenty most common reported affective-epistemic state's in English translation, were the following:

AES	Occurr-ences		AES	Occurr-ences
Hesitant	37		Certain	12
Determined	35		Harsh	9
Surprised	27		Neutral	8
Pensive	22		Tired	8
Uncertain	22		Reluctant	7
Happy	20		Stubborn	6
Interrogative	19		Interested	6
Hurried	17		Dejected	6
Positive	17		Bored	5
Agreeing	14		Sceptical	5

Table 3. Occurrences of most commonly reported AES's.

These are in total 302 of all 500 answers, all other responses occurred four times or less, 32 were blank, 44 were of a more inventive and un-categorizable nature, such as "condescendingly sympathetic" or "you are right, but I don't agree with you".

There is no one-to-one mapping of the answers of our respondents to the instructions of our actors, but there is considerable overlap. The difference in agreement between the respondents for individual samples ranges from almost unanimous for certain samples, to almost no agreement for others. We restrict ourselves to three typical examples, as there is too little data to be conclusive in any way.

One of the samples what was recorded with an 'interrogative' (questioning) AES got the following responses: 9 pensive, 7 interrogative, 2 happy, 2 agreeing, 1 blaming, 1 bored, 1 hesitant, 1 hopeful, 1 blank. We can see that even though there are some quite differing answers, the two most common answers have some similarity in meaning.

Another sample that was recorded with the AES 'happy' got the following responses: 11 happy, 8 positive, 1 enthusiastic, 1 exalted, 1 interested, 1 inviting, 1 sure, 1 determined. This shows quite a high degree of agreement between the respondents.

An example of a sample with very low agreement is one that was recorded with the AES 'hurried'. The responses were: 5 determined, 4 stubborn, 3 harsh, 2 afraid, 1 commanding, 1 formal, 1 negative, 1 offended, 1 sure, 1 unsettled, 2 blank and 3 incomprehensible. This sample had a duration of only 230 ms, which is likely to have

contributed to this sample being hard to interpret.

The samples that got the least agreement among the respondents on average show higher intensity and pitch, and wavering intensity- and pitch-curves. On contrast, the samples with the highest listener agreement have lower intensity and pitch, as well as more even intensity- and pitch-curves.

It should also be noted that among the samples with the highest listener agreement, several had quite distinct and audible non-prosodic voice qualities, such as creaking voice or audible breath sounds. This indicates that such features of the voice signal can have similar functions to prosody in indicating and displaying AES's.

For every category we calculated an average for the different parameters, based on every sample that was reported as belonging to that category. This means that every sample was counted as one instance every time it was reported as being a specific AES. E.g. if sample x was reported to be an instance of 'happy' by two different respondents and sample y was reported by one respondent, the average for any parameter of 'happy´ would be $((x_p + x_p + y_p)/3)$. This table presents four of these prosodic parameters.

Some interesting patterns emerge when the different categories are grouped together based on their averages on three key parameters, i.e. duration, intensity and pitch. For each parameter we split the range of the resulting values into three equal parts, e.g. if the range of the values on a particular parameter was between 1-30, all values between 1-10 would be designated as a low value, 11-20 medium and 21-30 high.

Term	Dur	Mean pitch	Pitch diff	Mean int
Hesitant	0,75	153	0,38	66
Determined	0,27	136	0,32	73
Surprised	0,54	200	0,63	76
Thoughtful	0,77	146	0,6	66
Uncertain	0,91	160	0,47	66
Happy	0,39	171	0,61	70
Interrogative	0,44	156	0,59	69
Hurried	0,24	140	0,27	76
Positive	0,34	164	0,52	73
Agreeing	0,31	144	0,52	67
Certain	0,31	134	0,37	73
Harsh	0,21	139	0,29	72
Neutral	0,31	131	0,44	70
Tired	0,46	124	0,26	67
Reluctant	0,69	140	0,29	66
Dejected	0,6	-	-	65
Interested	0,4	193	0,65	70
Stubborn	0,3	167	0,36	74
Bored	0,72	147	0,4	67
Sceptical	0,98	144	0,5	66

Table 4. Averages of the main prosodic features of the most commonly reported AES's.

	Low dur	Med dur	Hi dur
High int	stubborn hurried	surprised	
Med int	neutral certain determined harsh	positive happy asking	
Low int	agreeing	tired	sceptical thoughtful bored hesitant uncertain reluctant

Table 5. Grouping of the most commonly reported AES's in terms of duration and intensity.

	Low pitch	Med pitch	Hi pitch
High int	hurried		surprised stubborn
Med int	neutral sure harsh determined	asking	positive happy
Low int	reluctant tired	thoughtful bored hesitant uncertain agreeing	sceptical

Table 6. Grouping of the most commonly reported AES's in terms of pitch and intensity.

	Low pitch	Med pitch	Hi pitch
High dur	reluctant	thoughtful bored hesitant uncertain	sceptical
Med dur	tired	asking	surprised positive happy
Low dur	hurried neutral sure harsh determined	content agreeing	stubborn

Table 7. Grouping of the most commonly reported AES's in terms of pitch and duration.

There are three groups of labels that are not differentiated from each other in these dimensions, and they are 1) positive, happy; 2) neutral, sure, decided, harsh; 3) thoughtful, bored, hesitant, unsure.

In the case of the first group, it is hardly surprising to see that 'positive' and 'happy' fall close together. Looking more closely at their typical patterns, we can also find that both typically have a rise-fall pitch curve, and that both have a high pitch difference. Not very much set them apart in these data. We do find that 'happy' has somewhat longer duration, higher pitch and larger pitch difference, but lower intensity. The values for intensity are a little counterintuitive, as the word 'happy' would suggest a more intense affective state than 'positive', and the other three variables also indicate this. But emotional intensity does not equal vocal intensity in our data.

Term	Dur	Mean pitch	Pitch diff	Mean int
Happy	0,39	171	0,61	70
Positive	0,34	164	0,52	73

Table 8. Average values of the main prosodic features of 'happy' and 'positive'.

The second group has some cohesion between the labels, at least 'certain' and 'decided' seem to have some semantic similarity. Looking more closely at the data for this group, we also find that for three out of four values, 'certain' is closer to 'neutral' while 'decided' is closer to 'harsh', and that 'neutral' and 'harsh' are at the opposite ends of the spectra. The exception is intensity, where 'harsh' has a lower value than both 'decided' and 'certain'. But the differences are very small. 'Neutral' seems to be predominantly characterized by a fall-rise pitch curve, while none of the others have any typical pitch contour.

Term	Dur	Mean pitch	Pitch diff	Mean int
Neutral	0,31	131	0,44	70
Certain	0,31	134	0,37	73
Determined	0,27	136	0,32	73
Harsh	0,21	139	0,29	72

Table 9. Average values of the main prosodic features of 'neutral', 'certain', 'determined' and 'harsh'.

In the third group, 'thoughtful', 'hesitant' and 'uncertain' have some semantical similarity, but 'bored' seems to be a completely different thing. It should be noted that bored was only reported five times, whereas the others were reported more than 20 times each. All four typically have a rising pitch.

Term	Dur	Mean pitch	Pitch diff	Mean int
Bored	0,72	147	0,4	67
Hesitant	0,75	153	0,38	66
Thoughtful	0,77	146	0,6	66
Uncertain	0,91	160	0,47	66

Table 8. Average values of the main prosodic features of 'bored', 'hesitant', 'thoughtful' and 'uncertain'.

We also find clear indications that duration and intensity show a roughly linear correlation; shorter durations are correlated with higher

intensity and vice versa. Interestingly we also find a clear pattern that medium duration expressions are correlated with higher pitch and larger pitch differences on average.

4 Problems

Since this work was carried out as an extended pilot study to test the methodology, we will report some problems that should be avoided in future applications of this method.

The samples played back to the respondents were not of only one particular feedback word; rather it was random which AES was coupled with which feedback word. The idea behind this was that there should no systematic influence from the semantics of the word on the interpretation. The drawback of this is that the prosodic measures are not as comparable between different instances as they would be if the same word were used for all cases. Different sounds have different intrinsic qualities in terms of pitch and intensity, and different words have different durations. The latter is more notable in the case of the word 'okej' which has a longer duration than the others, which are more similar to each other in this regard. However, these differences between the different words are much smaller than the differences in focus in our results, and can be provisionally disregarded. By selecting to focus on only one word at a time, this problem is avoided. With enough data these intrinsic differences of the words could also be compensated for.

In two of the twenty samples the pitch was not detectable, because of a creaky voice quality. This might have had an influence on some of the composite pitch values. While creaking of the voice can be a very interesting signal with regards to communication of AES's, it does not fall within the scope of this investigation. In future studies, the problem will be handled by making sure that all samples can be analysed in all dimensions beforehand.

The small number of samples produced by only two different speakers and the relatively small number of respondents means that there is very little chance of making any extended statistical inferences, or calculating significance or doing variance analysis. This was an expected problem, as this is a pilot study.

5 Discussion

The fact that the respondents were in next to complete agreement about the AES of certain samples, while there was very little agreement for others, begs the question if there are certain qualities of these samples that produce these results. Further research to establish whether certain prosodic qualities are more easily identified with specific AES's, and the inverse for those that are difficult to identify, is under way. We are also interested in identifying if there are specific AES's that are easier to identify using vocal signal alone, and whether others are more reliant on other bodily expressions.

With regard to any specific measures presented here, we generally concede that we have to little data to draw any firm conclusions yet. What we have presented are preliminary findings, which at can be taken as indications of the kind of results that we hope to present later, and as indications of general directions of what these results might show. Even so, we are encouraged that many of the results that we do see are in agreement with our preconceptions of how these AES's are expressed in Swedish. Many of the categories seem to be distinguishable in terms of their prosodic qualities. The fact that some seem to cluster together can be seen as a call for more research into these specific AES's, using both our present methodology as well as other kinds. A possible hypothesis might be that these AES's are not distinguished very clearly in terms of prosody, but might instead rely more on facial expressions or other bodily expressions.

Acknowledgements

The research that has led to this work has been supported by the NOMCO project, which is funded by the NORDCORP program under the Nordic Research Councils for the Humanities and the Social Sciences (NOS_HS) and the European Community's Seventh Framework Programme (FP7/2007-2013), under grant agreement no. 231287(SSPNet).

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., Nivre, J. & Ahlsén, E. (1992) On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9(1), 1-26.
Boholm, M., Lindblad, G. (2011) Head movements and prosody in multimodal feedback. *NEALT Proceedings Series: 3rd Nordic Symposium on Multimodal Communication*, pp. 25-32.

Vagueness and Gestures

Jens Allwood
University of Gothenburg
Gothenburg, Sweden
jens.allwood@gu.se

Evelyn Vilkman
University of Gothenburg
Gothenburg, Sweden
gusvilkm@student.gu.se

Abstract

In communication, vagueness, unspecificity, approximation, uncertainty and hesitation (VUAUH) can all be expressed by words, prosody or gesture. This paper examines the relationship between the expression of VUAUH and gestures connected to them. A corpus consisting of political debates was used to select relevant data. The data enabled us to connect some of the VUAUH types, or combinations of types, to gestures. Findings indicated that, for example, approximation was expressed together with head waggle.

Keywords: vagueness, gesture

1 Introduction

In communication, vagueness, unspecificity, approximation, uncertainty and hesitation (henceforth VUAUH**)** can all be expressed by words, prosody or gesture. VUAUH phenomena are empirically not mutually exclusive, but can occur simultaneously in communicative behavior. It is possible to be vague, unspecific and approximate at the same time. This means that a particular vocal or gestural expression of one or other of the VUAUH phenomena could simultaneously be an expression of one or more of the other VUAUH phenomena. It is also possible that one or more of the phenomena occur(s) without the others, i.e. you can be uncertain without being vague or vague without being uncertain. In general, the purposes and reasons of a particular interlocutor in a particular activity will probably strongly influence what kind of VUAUH expression is used.

In face-to-face communication gestures and speech occur together (Allwood, 2001; Mc Neill, 1992). Gestures and speech can support each other so that the meaning is expressed in an appropriate way, e.g. making a statement more vague or less serious for reasons of lack of supporting evidence (Allwood, 2001). Gestures can be used both for strengthening and weakening a communicated message. For example, gestures for indicating a lack of knowledge could be a rocking head, shrugs (Allwood, 2001 & Poggi et al 2003) or lowering forearms with empty hands (Poggi et al 2003). Self-confidence can instead, e.g. in Italy, be communicated with hand open hand and palm up, like a flower (Poggi et al 2003).

Below we will identify different gestures depending on what VUAUH phenomenon, or combination of VUAUH phenomena that occur. We are also going to look at the two main reasons for expressing vagueness, (i) lack of information or (ii) unwillingness to express information, and see if this difference affects what gestures that are used.

2 Method

A corpus consisting of video recorded Swedish political debates was used to select relevant data. The corpus consists of 21 videos with a total duration of approximately 8.5 hours. The material is recorded by the Swedish television channels SVT 1, Kanal 5 and TV4 and has also been broadcasted on television. The debates concern different political issues that were relevant in the election 2010. This is also the period when the recordings were made.

The participants are politicians, journalists and also persons interested in or related to the different topics. Each debate was moderated by between one and four hosts, the TV-hosts were sometimes involved in the debates (expressed opinions, provoked, disagreed and so on).

Our data selection consisted in viewing the videos in order to excerpt utterances expressing vagueness. The videos were watched one by one

and when a vague expression occurred, the starting time for the utterance containing the vague expression was noted. After finishing one video the selected utterances were cut out, labeled with their associated vagueness expression (cf. table 1) and grouped based on this expression. A minimum frequency requirement (for the total number of occurrences in all videos) was set at 20, so only the most frequent vagueness words were included.

Two clips from each frequent vagueness expression were randomly selected to be included in this study. These clips were transcribed according to the GTS (Gothenburg transcription standard, Nirve et al 2004). Additional information about context was also described.

Two discussion sessions between two coders were held, during which the sample videos were watched and the utterance/vague words were annotated with regard to which VUAUH functions they expressed in context. There was also an attempt to interpret the reason for using the vagueness expressions.

13 types of vagueness related expressions with target words meeting the frequency requirement were used (see table 1) About 400 individual clips were totally collected. Some utterances in the video clips included several target words and where thus included in different categories leading to a total of 600 clips.

A closer investigation of the included expressions in context showed that they were not all used to express vagueness. Example (1) demonstrates this, where the word *något* (*any*) that often can be used to express vagueness, but in the example is not used to express any of the VUAUH functions. Instead it is used as a reinforcing strengthener.

(1) $FR: att <1 det inte ska >1 säljas / <2 något danderyds >2 sjukhus eller <3 något annat akutsjukhus >3 eller universitetssjukhus <4 | >4

that there will be not be sold any danderyds hospital or any other emergency hospital or university hospital

Frequent Swedish vagueness related expressions in the corpus	**English translation**
Liksom	like
Kanske	perhaps, maybe
Lite	a bit
Massor	a lot
Och/eller så/sånt	and so/such
Och så vidare	and so on
Så att säga	in a manner of speaking
Ungefär	roughly
Ganska/rätt	rather, pretty
En del/delvis	partly
Någon	some, somebody
Något	any, something
Någonting	something

Table 1: Common vagueness related expressions used (Swedish) and translation in English.

26 clips were randomly chosen from the 13 vagueness related expressions used (two clips from each type). All vagueness related expressions included in these 26 utterances were classified according to which VUAUH function they expressed. The classification is presented in table 2 (V = vague, UnS = unspecific, A = approximate, UnC = uncertain, H = hesitation).

All clips were also coded for which gestures cooccurred with the vagueness expressions.

3 Results

Table 2 below presents the VUAUH functions expressed by our chosen vagueness expressions in the context of the 26 utterances investigated

Frequent Swedish vagueness related expressions	**VUAUH function**
Liksom	V, UnS, UnC
Kanske	H, UnC
Lite	V
Massor	UnS
Och/eller så/sånt	UnS
Och så vidare	UnS
Så att säga	UnS, V
Ungefär	A
Ganska/rätt	UnS, UnC
En del/delvis	UnS, UnC
Någon	H
Något	
Någonting	UnC

Table 2: VUAUH functions expressed by frequent vagueness related words

Table (2) shows that all the selected vagueness related words except one, at least once, have been classified as expressing one of the VUAUH functions. The most common of these functions is unspecificity. Five of the words can be described as often expressing to at least two VUAUH functions (*liksom, kanske, ganska/rätt, en del/delvis, så att säga*). The expressions *ungefär* and *en del/delvis* have all been described as UnS and UnC. One category (*något/any*) did not express any of the VUAUH functions.

The different combinations of VUAUH functions that occurred in our sample material are shown in table 3.

Sequential and Simultaneous	**Sequential**
H + UnS, V + H H + H + H, UnC H + H, UnS + UnS H + UnS, UnC	H + UnC +H H + H + V + H H + UnS H + V H + H H + UnC UnC + H UnS + H + UnS H + UnS H + UnS + UnS H + Uns + UnS
Simultaneous	**Single**
	UnS UnS UnS H H H A A V

Table 3: Different combinations that occurred in the sample.

Simultaneous combinations occurred only in the surrounding of other VUAUH types. In all these cases, hesitation is present. Together with unspecificity they create the most common combination (in various settings). For the single VUAUH occurrences hesitation and also unspecificity are the most common types.

If we consider table 3, we see that hesitation seldom is bound to a particular word. It is often expressed by several words in sequence. It can also be expressed in various other ways such as through repetitions of words, sounds, different facial gestures; which are all possible expressions of the other functions as well, but hesitation seems to have the largest repertoire of expressions (see table 5).

The gestures that were connected to the words and reoccurred (at least twice) are shown in table 4. Words without any gesture connected to it in the table were either without any multimodal component or connected with gestures that did not reoccur. In spite of often expressing vagueness or unspecificity, the word *något* in this corpus did not express these functions and the gestures connected to this word are thus not included.

Frequent Swedish vagueness related expressions	**Gesture**
Liksom Kanske Lite	Gaze forward
Massor Och/eller så/sånt Och så vidare Så att säga	Eyebrow raise
Ungefär Ganska/rätt En del/delvis	Waggle with head
Någon Något Någonting	-

Table 4: Frequent Swedish vagueness related expressions and gestures that occurred more than twice together at the same time.

In table 5, we show how the gestures in table 4 and other gestures used with vocal verbal vagueness expressions are related to different VUAUH functions. Also gestures that only occurred once with a specific VUAUH function are included.

Uncertainty and vagueness had no specific gestures connected to them, whereas hesitation had many different gestures with no clear similarities between them. Unspecificity included shaking and tilting head. Approximation was the only category that had one consistent gesture. It has to be mentioned that the material is thus not so large (approximation was found only in two cases among the 26 video clips).

The reasons for expressing vagueness were not obvious. In almost all cases the choice between lack of information and unwillingness to express information) could not be made. Example (2) and (3) illustrate the problem.

VUAUH	Gesture
Vagueness Uncertainty Approximation Unspecificity Hesitation	 Waggle with head Shaking head Tilting head Eyebrow raise Gaze away (up, down, sideways) Gaze forward Beating with one hand Various coordinated movements with both hands (sideways and forward) Raising eyebrows Shaking head

Table 5: Gestures connected to each VUAUH type.

(2) $WP: ja nitti procent det rör väl sig<1 ungefär om / tolv tusen människor >1 <2 / åtta till tolv tusen människor >2 per år

yes ninety percent it is roughly about twelve thousand people from eight to twelve thousand people per year

The reason for the use of *ungefär* in this case could both be that the speaker doesn't know how many people that are relevant in this case and choses to be vague about it (*ungefär*/*roughly*). It could also be the case that the speaker knows but does not want to bore the audience with exact numbers.

(3) $GF: <1 Ö1 om vi ska nå >1 det så kallade två graders målet så måste vi <2 (b)/ minska koldioxidemissionerna >2 me{d} <3 / kanske >3 nittio procent till tjugo femtio 4< å0 sen fortsätta å1 minska >4

uh if we are to achieve the so-called two-degree target we need to reduce carbon dioxide emissions by perhaps ninety percent to twenty fifty and then continue to decrease

The same argument applies in this case as well. The speaker could believe or guess that we have to reduce the carbon dioxide emissions by ninety percent. It could also be the case that the speaker doesn't think is relevant to be to precise about the number, despite the fact that he knows.

4 Discussion

The connection between speech and gesture is complex. There is no "one to one" relationship between a word and a gesture. But gesture and speech support each other and are used together for different reasons.

In this material we have indications that approximation may be coupled with a waggle of head. In Swedish culture, this gesture can be thought of some kind of tentative behavior, whereas in India a similar head gesture is a kind of affirmative gesture.

Unspecificity was connected with gestures described as: tilting head, head shake and raise of eyebrows. The raise of eyebrows may be connected with surprise, but also occurs in expression of both hesitation and unspecificity. This is also the case with shaking head, which in addition can be connected with some kind of disagreement/unwillingness.

Unfortunately, since we did not have access to the speakers appearing in the video tapes we have examined, we have not been able to investigate the difference between lack of ability and lack of willingness to express information. To our video observation based interpretation both possibilities seem to exist in all the cases we have analyzed. Another study with possibilities of interviewing the speakers might enable us to investigate this further.

Hesitation is the most common VUAUH-phenomena in our sample, and is connected with a range of different gestures. Hesitation is also found in all combinations of VUAUH-phenomena (both simultaneous and sequential combinations), where it in most cases marks the start of the coming combination. This is no surprise as hesitation can be used in different ways; to hold the turn while thinking as well as to "signal important affect-related information" (Campbell, N. 2007).

It is more common that VUAUH phenomena occurs together (parallel and/or sequential) that alone (with the exception of Approximation that occurred only alone). This may explain why it is hard to operationalize and separate them from each other. We can also see this as an indication that we should not try to isolate one type but instead focus on VUAUH-clusters and look at their

properties. It also remains an open empirical question whether the clusters should include other members, such as tentativeness. Further investigation will decide.

Acknowledgements

This research has been supported by the European Community´s seventh Framework Programme (FP7/2007-2013) under grant agreement no.231287 (SSPNet) and by the project: "Multimodal Corpora in the Nordic Countries" under the NORDCORP program, the Nordic Research Council for the Humanities and the Social Sciences (NOS-HS).

References

Allwood, J. (2001). Cooperation and Flexibility in Multimodal Communication. In *SALSA - Symposium About Language and Society - Austin*. Texas Linguistic forum. Vol. 44, Nos. 1 & 2, 2002, pp. 21-17.

Campbell N. (2007). On the Use of NonVerbal Speech Sounds in Human Communication. In *Verbal and Nonverbal Communication Behaviours.* Lecture Notes in Computer Science 4775, Springer Verlag, Berlin/Heidelberg, pp 117-128.

McNeill, D. (1992). *Hand and mind: what gestures reveal about thought.* Chicago, IL: University of Chicago Press.

Poggi, I., Pelachaud, C. & Magnu Caldognetto, E. (2003). Gestural Mind Markers in ECAs. In *Gesture-Based Communication in Human-Computer Interaction: 5th International Gesture Workshop, GW 200.*A. Camurri, G. Volpe (eds.). Springer Verlag, Berlin Heidelberg New York.

Nivre, J., Allwood, J., Grönqvist, L., Gunnarsson, M., Ahlsén, E., Vappula, H., Hagman J., Larsson, S., Sofkova, S. & Ottesjö, C. (2004). Göteborg Transcription Standard. V 6.4. Department of Linguistics. University of Gothenburg

Annotating Attitudes in the Danish NOMCO Corpus of First Encounters

Anette Luff Studsgård
University of Copenhagen, Denmark
Lund University, Sweden
anette@studsgaard.eu

Costanza Navarretta
University of Copenhagen
Copenhagen, Denmark
costanza@hum.ku.dk

Abstract

This article describes a strategy for facilitating the annotation of emotions and attitudes in a corpus of video-recorded spontaneous conversations. The proposed strategy helps the annotators in reaching a common understanding of emotion labels, thus resulting in a better inter-coder agreement. The adopted annotation scheme combines a list of emotion labels with the bi-polar annotation of three categories Pleasure, Arousal, and Dominance according to Kipp and Martin (2009). In order to reach a common understanding of the proposed emotion labels, the PAD for each emotion has been established, and then each emotion has been placed in the PAD dimensional space where each dimension had 10 degrees of freedom.

Keywords: multimodal corpora, annotation, emotion

1 Introduction

The Danish NOMCO corpus of first encounters is the Danish version of comparable Nordic multimodal first encounters corpora (Paggio et al. 2010, Navarretta et al. 2011). These corpora have been transcribed and multimodal body behaviours, gestures henceforth, have been annotated according to the MUMIN annotation scheme (Allwood et al. 2010).

In the Danish corpus, the communicative gestures have been identified; their shape has been described and classified with respect to the functions of feedback, turn management and sequencing in the ANVIL annotation tool, (Kipp 2001). In the following, we focus on the annotation of emotions and attitudes, attitudes henceforth, in facial expressions. All annotations are done manually. Since manual annotation of these types of information is a difficult task, it is extremely important that the annotators have a common understanding of the annotation categories, and follow the same annotation guidelines.

Figure 1: Two attitudinal states

Figure 1 shows two pictures of one of the participants, M1, expressing amusement and then uncertainty.

Annotating attitudes is more difficult than annotating other communicative features. First of all, facial expressions can be perceived differently by the various coders. Furthermore, the attitude labels can be interpreted in various ways also because of the ambiguity of many of the labels as well as the subtle difference between some of the attitudes.

In the Danish NOMCO project, attitudes were annotated by combining two annotation models: a discrete and a dimensional one. The discrete model consists of an extension of the attitude list proposed in the MUMIN scheme (Allwood et al. 2007) and also included a number of attitudes used in the Swedish NOMCO corpus. The discrete model is taken from Kipp and Martin (2009) who simplify the complex three dimensional system defined by Mehrabian and Russel (1997). Following Kipp and Martin, + or – values are assigned to describe attitudes according to the three categories Pleasure (P), Arousal (A), and Dominance (D). One of the reasons for choosing the combined annotation scheme was the fact that two coders preferred using semantic labels while two coders preference the dimensional model. A complete discussion of the motivation behind the chosen model and a first analysis of the emotions in the corpus is in (Navarretta 2012).

In this paper, we focus on the annotation manual which was constructed by the annotators to reach a common understanding of the

Figure 2: The ANVIL annotation

annotation labels using their position in the PAD space. The main aim of this work is to facilitate the coding and improve the consistency and reliability of the annotations.

In the rest of the paper, we present the data (section 2), the annotation procedure (section 3), the construction of the annotation manual and an evaluation of its effect (section 4). In section 5, we conclude.

2 Data

The data are the Danish NOMCO corpus of first encounters (Paggio and Navarretta 2011) consisting of conversations between two subjects who do not know each other in advance. They are placed in front of each other and are asked to converse until interrupted by the experiment leader (approximately after five minutes). The participants are 6 female and 6 male subjects. Each participant converses with one participant of the same gender and one of the opposite gender. The 12 conversations are video recorded from three different angles (upper body of subject 1, upper body of subject 2, and whole body of both subjects captured sidewise in one frame). During the annotations of the conversations a merged and synchronized video of the upper bodies of the two subjects was used as main source, supplied in few cases by the third video. Figure 2 shows a screen shot of the annotation tool, ANVIL (Kipp 2001).

Four annotators participated in the first inter-coder experiments, but the annotation of attitudes in all conversations have been conducted by two student employees. These two annotators have been associated with the NOMCO project for two years. The annotation criteria and the revision of the annotations have been devised by two senior researchers at the Centre for Language Technology at University of Copenhagen.

In the following, we focus on the annotations of attitudes. The annotators use both visual and auditory inputs when considering which facial expressions express an attitude and thus choose the correct attitude taking into account the whole context. As with any of the other features of the corpus, the procedure is that the first annotator annotates the attitudes, then the second annotator reviews the annotation and sends her comments back to the first annotator, who will consider these. Cases of disagreement are then discussed with the senior researcher, and an agreed upon version is so defined. The final attitude is thus defined through agreement of the two annotators and the senior researcher.

3 Defining the annotation procedure for facial expressions

According to the annotation guidelines, only facial expressions which express an attitude must be annotated. Thus, the annotators must decide which of the communicative facial expressions previously annotated also express an attitude.

Two different annotation experiments were performed in the beginning. In the first experiment, four different annotators had to annotate independently the same conversation describing an attitude only in terms of the bipolar PAD values. The best inter-coder agreement between two of the annotators was of 0.25 in average in terms of Cohen's *kappa* (Cohen 1960). These two annotators found the PAD annotation easier to annotate than the other two annotators.

In the second experiment, the four annotators had to code attitudes in a conversation using the attitude list. The results of these experiments were parallel to the preceding experiment. The two annotators who liked the emotion label system mostly are the two who did not like the PAD system, and they obtained the best inter-coder agreement results in this test. The achieved scores were lower than in the first experiment (Navarretta 2012). The two annotators who disagreed mostly in the two experiments were the two annotators who had to code the emotions in the corpus.

In the next step, the two annotators and a senior researcher connected PAD values to each attitude label, and added this information to the

PAD +++
Amused, Excited, Happy, Ironic, Proud, Satisfied, Self-Confident, Supportive

PAD -+-
Awkward, Embarrassed, Puzzled, Uncertain

PAD +-+
Certain, Friendly

PAD ---
Disappointed, Hesitant, Uncomfortable, Unconfident

PAD +--
Docile, Thoughtful

PAD ++-

Figure 3: The list of 25 attitudes ordered by PAD values

annotation scheme. Since the list of attitude labels is open-ended, new attitudes could be added to the list when necessary and if all the three agreed on this value. So, facial expressions which are judged to express an attitude are assigned an attitude value (Happy, Uncertain, Surprised etc.) and a PAD combination (e.g., P-A+D-). Every attitude has a static PAD combination shown in figure 3.

One conversation was annotated according to the new combined scheme and the results were corrected and discussed according to the annotation procedure. At this point a new inter-coder agreement test was done with the two annotators. The results of these experiments were much better than those achieved in the preceding ones, but they were still not higher than 0.40 on 25 emotion labels (Navarretta 2012).

4 Improving intercoder agreement

The list of attitudes in Figure 3 includes the following attitudes:

those which are internal feelings evolving around one's self-esteem, e.g. Self-confident, Happy, Proud

those which are reactions and feelings occurring during the discourse, e.g., Amused, Irritated, Surprised,

those concerning the content of the discourse (e.g., Supportive, Certain, Hesitant).

The different nature of the attitudes complicates the choice between attitudes with the same PAD combination. Thus, the annotators decided to grade the PAD value for each attitude on a scale ranging from -5 to 5, (0 not included). The stronger the attitude, the closer to either -5 or 5 the mean value for the PAD will be. Similar, the closer to 0, the more neutral the attitude is.

The ranking is decided by the two annotators to make sure the different attitudes were understood in the same manner. We found that the attitudes can be ranked as shown in the list in figure 4.

5.0: Ironic

4.0: Amused, Happy, Proud

-4.0: Uncomfortable, Unconfident

3.7: Confident

3.3: Excited

-2.7: Disappointed

2.3: Engaged, Supportive

2.0: Satisfied

1.7: Certain, Surprised

-1.7: Embarrassed, Puzzled, Uninterested

1.3: Friendly

-1.3: Awkward

1.0: Interested

-1.0: Hesitant

-0.7: Uncertain

0.3: Docile, Thoughtful

-0.3: Irritated

Figure 4: One Possible Ranking of the Attitudes

According to this list Ironic is the strongest attitude in the present scheme, whereas Docile, Thoughtful and Irritated (in this corpus used for a mild irritation) are the less strong attitudes.

In the following the different PAD combinations will briefly be described, because there actually is a variation among the attitudes within the same PAD combination. Even though the PAD combinations are the same, the "strength" of Pleasure, Arousal, and Dominance varies among the different attitudes.

4.1 P+A+D+

Looking at figure 4 it might seem that e.g. Amused, Happy, and Proud are interpreted as the same attitudes since they all are ranked 4.0 on the PAD scale. However, this is not the case. The mean PAD value (4.0) is derived from the

	P	A	D	Mean
Ironic	5	5	5	5,0
Proud	4	3	5	4,0
Happy	5	3	4	4,0
Amused	5	5	2	4,0
Self-Confident	5	2	4	3,7
Excited	3	5	2	3,3
Supportive	3	3	1	2,3
Satisfied	3	1	2	2,0

Figure 5: Ranking fir attitudes with +++ PAD

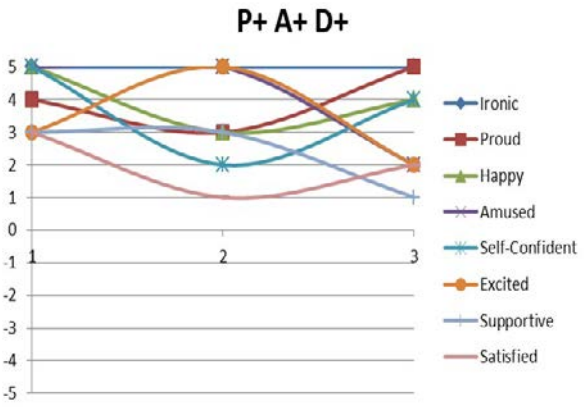

Figure 6: Graph for P+A+D+ attitudes

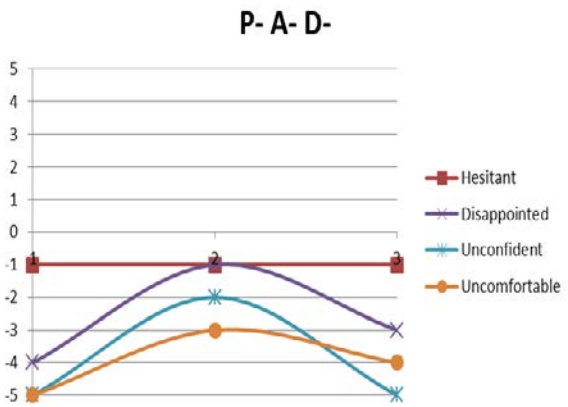

Figure 7: Graph for P-A-D- attitudes

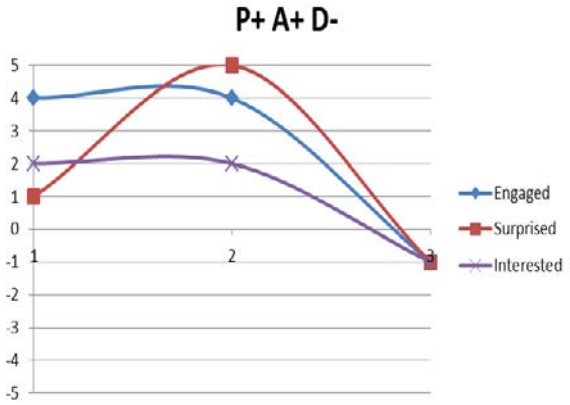

Figure 8: Graph for P+A+D- attitudes

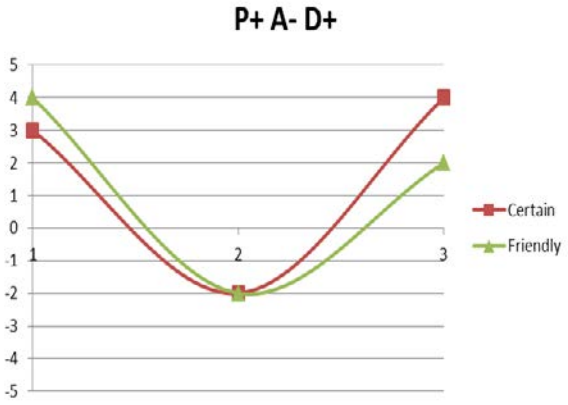

Figure 9: Graph for P+A-D+ attitudes

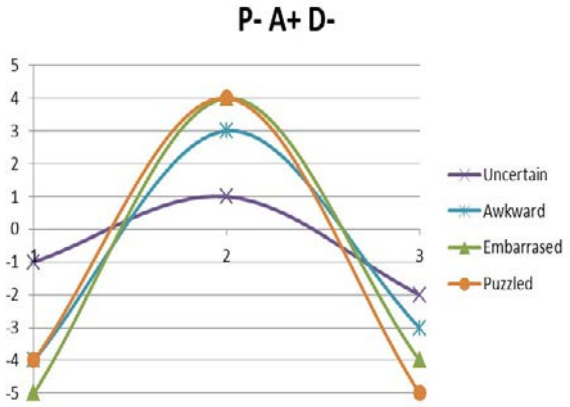

Figure 10: Graph for P-A+D- attitudes

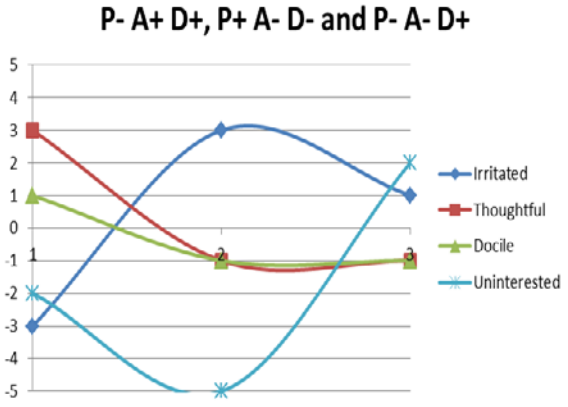

Figure 11: Combined graph for P-A+D+, P+A-D- and P-A-D+ attitudes

independent ranking of P, A, and D, and the rankings of e.g. D in these cases are 4, 5, and 2, respectively. All ratings for the attitudes with PAD combination +++ can be seen in figure 5.

Figure 6 shows the PAD values for all of the attitudes with the PAD combination +++. In the graph 1 resembles Pleasure, 2 Arousal, and 3 Dominance.

Figure 6 also shows how differently the eight attitudes are interpreted in spite of having the exact same PAD combination, e.g. is the curve for Excited, with Arousal having higher rating than both Pleasure and Dominance, opposite to the curve of Happy, Proud, Satisfied, and Self- confident where Arousal is rated lower than both Pleasure and Dominance.

4.2 P-A-D-

Similarly, a graph for the attitudes with the PAD combination --- can be created (figure 7). Here it can be seen how Hesitant is the most neutral attitude and does not vary in strength among the three features, whereas the three other attitudes, Disappointed, Unconfident, and Uncomfortable, all follow the same pattern with lower ratings on Pleasure and Dominance, and a higher rating on arousal.

4.3 P+A+D-

This category contains the attitudes Engaged, Interested, and Surprised. These attitudes all evolve around being an active interlocutor and thus are higher on Pleasure and Arousal, and lower on Dominance, however only just below the negative boarder. The graph for P+A+D- can be seen in figure 8.

4.4 P+A-D+

In this category there are only two attitudes, Certain and Friendly. When looking at the ratings (figure 9) for the two attitudes displayed in the graph they seem quite alike. The biggest difference between the two is that Friendly is rated higher than Certain concerning Pleasure, and that Certain is rated higher than Friendly concerning Dominance.

4.5 P-A+D-

All these rather unpleasant attitudes, Awkward, Embarrassed, Puzzled, and Uncertain, show a quite similar curve when rated (figure 10), though Uncertain is not as distinct in the rated difference between the three features as the other attitudes. Awkward, Embarrassed, and Puzzled are all rated very low in Pleasure whereas there is more variation concerning Dominance.

4.6 P-A+D+, P+A-D-, and P-A-D+

The four attitudes in figure 11 have been combined since they are all quite rare in the corpus. Irritated (P-A+D+) and Uninterested (P-A-D+) are only different on one feature, Arousal, whereas Docile and Thoughtful (both P+A-D-) only have one feature in common with the two other attitudes.

4.7 An evaluation

A new inter-coder agreement test was done after the annotators had produced the annotation manual. The inter-coder agreement improved with more than 0.20 on both emotion labels and PAD values compared to the preceding experiments, but the two results are difficult to compare given that fewer emotions were assigned in the second experiment and that the annotators had gained more experience. However, the improvement is high, thus we believe that positioning the attitudes in the PAD dimensional space helped the annotators defining a common understanding of the emotion labels.

5 Conclusion

Initial to this work an intercoder agreement with the result of a mean on 0.40 was conducted. Even though annotating attitudes is a difficult task regarding the subjectivity with which we understand attitudes of other people, the first intercoder agreement was not acceptable. However, this work improved the intercoder agreement on average with 0.20, ending up with a mean of 0.60 on intercoder agreement, which we regard satisfying in the annotation of attitudes.

So, combining attitude labels with combinations of + or – in the features Pleasure, Arousal, and Dominance, and assigning different degrees of strength to each attitude on the three scales has helped the annotators to get a similar understanding of how an attitude value should be interpreted and annotated. However, in order to actually test whether this improvement is due to a better procedure, or simply the gained experience of the two annotators, it would be necessary to let other less experienced annotators use the method for annotating attitudes in a similar corpus.

Acknowledgements

The authors would like to thank the three other coders of the corpus, Sara Andersen, Magdalena Lis, and Bjørn Wessel-Tolvig, and the Nordic and Danish councils of research that founded the research as well as the NOMCO project partners, Elisabeth Ahlsén, Jens Allwood, Kristiina Jokinen, and in particular Patrizia Paggio.

References

Allwood, J., L. Cerrato, K. Jokinen, C. Navarretta & P. Paggio (2007): The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. In *Lang Resources & Evaluation* (2007) 41, 273-287. Springer Science+Business Media B.V.

Kipp, M. (2001): Anvil – a generic annotation tool for multimodal dialogue. In *Proc. of Eurospeech*, pp. 1367–1370.

Kipp, M. &. Martin, J.C. (2009): Gesture and Emotion: Can basic gestural form features discriminate emotions? In *Proc. ACII Workshops*, 009, pp. 1–8.

Navarretta, C. (2012): Annotating and Analyzing Emotions in a Corpus of First Encounters. In *Proceedings of the 3rd IEEE International Conference on Cognitive Infocommunications*, Kosice, Slovakia, 2-5 December 2012.

Paggio, P., J. Allwood, E. Ahlsén, K. Jokinen & Navarretta, C. (2010): The NOMCO multimodal Nordic resource - goals and characteristics. *Proceedings of LREC 2010*, pp. 2968- 297.

Paggio, P. & Navarretta C. (2011): Feedback and gestural behaviour in a conversational corpus of Danish. In P. Paggio, E. Ahlsén, J. Allwood, K. Jokinen, C. Navarretta (Eds.) *Proceedings of the 3rd Nordic Symposium on Multimodal Communication* May 27-28, 2011, University of Helsinki, Finland, NEALT Proceedings Series vol. 15, pp. 33-39.

Russell, J. & Mehrabian, A. (1977): Evidence for a three-factor theory of emotions, *Journal of Research in Person* 11(3), pp. 274-294.

Attitudinal emotions and head movements in Danish first acquaintance conversations

Bjørn Wessel-Tolvig
University of Copenhagen
bwt@hum.ku.dk

Patrizia Paggio
University of Copenhagen
University of Malta
paggio@hum.ku.dk
patrizia.paggio@um.edu.mt

Abstract

There is wide agreement about the fact that different kinds of communicative gestures like head movements and facial expressions are used extensively in face-to-face communication for various purposes e.g. expressing feedback, emphasis, turn management and emotions, or supporting the speaker's own communication management. Moreover, studies show that individual differences in gestural behaviour can be attributed to different personality traits like extraversion and openness. But these findings do not model the fact that the same person may feel and act differently in different situations. We compare the results of post-experiment questionnaires consisting of self-rating questions about conversation experience, influence and attitudinal emotions, to gestural behaviour in a multimodal corpus of face-to-face conversations. Our findings suggest a slight positive correlation between the way conversations are perceived and the number of head nods produced. In turn, the degree of head movement in the specific conversation is probably a function of the subject's positive engagement in it.

Keywords: attitudinal emotions, .first-acquaintance conversations, head nods, gesture, Danish

1 Introduction

A variety of body behaviours are used in everyday communication. Especially head movements are used extensively in face-to-face dialogue. These movements encompass head nods, shakes, tilts, (side-) turns, forward and backward movements, and are used for various purposes e.g. to express feedback both in terms of affirmation and negation, to regulate turn taking, to emphasise certain parts of speech, to show emotions or simply control own communication management (Duncan 1972, McClave 2000, Paggio & Navarretta 2011a). The nodding of the head is a very effective form of communicative body behaviour. In many western cultures, head nods are used as signs of perception, contact and understanding (Allwood et al. 1992) or acceptance, approval and agreement (Helweg-Larsen et al. 2004). A term often used is *backchannel*. Yngve (1970) defines backchannels as vocal or gestural expressions of the listener meant to give feedback to the speaker without taking the floor. Head nods are considered backchanneling devices in Duncan (1972). Maynard (1987) studies head nods in dyadic conversations of Japanese and finds that nods are frequently used as backchannels. He also observes different distribution patterns of nod occurrences in Japanese and American speakers. McClave (2000), in an extensive overview and detailed qualitative study, claims that backchanneling is a hearer signal in line with words such as "yeah, um, hm". A nod indicates that the listener is taking note of what the speaker is saying. McClave finds that head movements in general have a variety of functions, and head nods in particular often function as backchannel requests, to which listeners are extremely sensitive. Many studies of Scandinavian and Baltic languages also describe the feedback function of head movements in general and nods in particular (Cerrato 2007; Boholm and Allwood 2010; Jokinen 2008; Paggio and Navarretta 2010a, 2010b, 2011a, 2011b and 2011c).

If these studies discuss the form, function and frequency of various head movements in speaker populations of specific language communities, other work focuses on the fact that personality traits are reflected in different kinds of gestural behaviour. Argyle (1975), for example, observes how gesture range appears to correlate with extraversion. Lippa (1998), in a study on nonverbal displays, extraversion and gender, finds that extraversion correlates with broader gestures among women, but not among men. In the same

study, extraverts tended to have faster speech, which led to higher gesture rates due to speech-gesture correlation. Brebner (1985) and Riggio (1986) both discuss how the frequency of hand movements correlates with extraversion. They seem to confirm the general intuition that gestural behaviour and traits like openness and extraversion correlate. Most of these findings, however, do not model the fact that the same person may feel and therefore act differently in different situations.

Indeed, we think context might play a role in gesture frequency. This is noted e.g. by Batrinca et al. (2011), who explain the difficulty in automatic classification of extraversion and agreeableness with the complex interaction of personality traits and characteristics of the situation. Our hypothesis is that a person's non-verbal behaviour is likely to be affected by that person's experience of the conversation, how much they felt in control, were affected by the surroundings and what mood they were in during the interaction. In particular, we are interested in the role attitudinal emotions play in gesture production, where by attitudinal emotions we mean emotional reactions towards the conversation itself – e.g. whether the subject is at ease or irritated – rather than the six basic emotions described and used in many studies (Ekman, 1972).

2 Corpus

The data we use for this analysis are found in a Danish multimodal corpus which is part of NOMCO (Multimodal Corpus Analysis in the Nordic Countries): a large scale corpus collection consisting of multimodal corpora in Danish, Swedish, Finnish and Estonian. The primary aim of the project is to provide comparative multimodal data in Nordic languages and study how speech and gesture are interrelated when expressing feedback, turn taking, sequencing and information structure (Paggio et al. 2010).

2.1 Recordings

The Danish corpus consists of 12 video recorded first acquaintance conversations of an average duration of 5 minutes each ~ 1 hour of data. The participants in the study were six females and six males, all university students or post grads. They had never met before and had no relation with each other beside the fact that they all knew one of the experimenters and could all relate to university studies. All subjects engaged in two conversations, one with a male and one with a female. The subjects were standing in front of each other and recorded from three different angles, one centered on both simultaneously in profile, and two focusing on each of the participants from a more frontal angle as shown in figure 1.

Figure 1: Recordings from the Danish NOMCO corpus. One center view (top) and one merged split view (bottom).

2.2 Annotation

Speech was transcribed with Praat (Boersma & Weenink 2009) and gestures with ANVIL (Kipp 2004). In annotating the multimodal data we followed the instructions and coding conventions from the MUMIN coding scheme (Allwood et al. 2007). In this particular study we focus on head movements, which were annotated for form and function, using the attributes and values listed in table 1 below.

Attribute	Value
Head movement	Nod, Jerk, HeadForward, HeadBackward,Tilt, SideTurn, Shake, Waggle, HeadOther
Head repetition	Single, Repeated
FeedbackBasic	CPU, FeedbackOther
FeedbackDirection	FeedbackGive, FeedbackElicit

Table 1: Annotation features for gestural behaviour

First form, or the movement direction and where applicable repetition, was annotated, then function. In this study we only focus on feedback functions, defined by Allwood (1992) as behaviour that has the purpose of giving or eliciting

signals of contact, perception and understanding. Note, thus, that feedback in the MUMIN scheme subsumes backchanneling: whilst backchannels are signals given by the listener (corresponding to FeedbackGive in our framework), we also annotate signals made by the speaker to elicit feedback (FeedbackElicit).

2.3 Questionnaire

After each conversation the subjects were asked to describe their experience in a questionnaire containing questions about setting, interaction and attitudinal emotions. They were given 12 questions inspired by Nezlek (2010), which had to be rated on a 1-5 point scale, five being the most positive result. The questions fall into four types concerning:

1) Perception of the conversation and oneself
2) Control of the situation
3) Surroundings
4) Attitudinal emotions

We were interested in knowing whether the participants enjoyed the experiment, if it was interesting, if they felt influential and free to express their ideas and if they were affected by the studio surroundings. The fourth category, *attitudinal emotions*, groups questions about personal attitudes relating to the subject's emotional state during the interaction and their response to the conversation context. Did the participants feel pleased or sad with the interaction/interlocutor, were they relaxed or tense in the situation, at ease or not at ease, content or irritated with the way the conversation proceeded?

Question type	Variable	Mean	SD
Perception of conversation and oneself	Enjoyable	4.42	0.72
	Intimate	2.71	1
	Liked	4.04	0.91
	Interesting	4.17	0.76
Control of situation	Influence	3.75	0.79
	Free	4.13	0.74
Surroundings	Not affected	3.46	1.06
	Natural	2.33	1.05
Attitudinal emotions	Pleased	4.58	0.58
	Relaxed	3.58	1.06
	At ease	3.83	0.82
	Content	4.46	0.88

Table 2: Questionnaire results

A summary of the results of the questionnaires is displayed in table 2.
In what follows, we will focus on the *difference* between the two conversations each subject took part in, to study i) if there is in general a difference between the way subjects experienced first and second conversation ii) if this difference correlates with a difference in gestural behaviour. We compared average results for the whole group, but also looked at individual variation to find if any of the subjects deviates significantly from the group norm.

3 Data analysis

The goal is to compare the results from the questionnaire with the participants' body behaviour. Thus a statistical analysis was carried out on the questionnaire results and the gestural annotation.

3.1 Questionnaire results

A comparison of the group's first and second conversation shows that the subjects on average rated the first conversation quite positively with 44.08 points (out of a possible maximum of 60). The second conversation was rated higher than the first, with an average score of 46.83 points. However, these results are slightly less homogeneous as can be seen by the score distribution for the two conversations in figure 2. The increase by 2.75 points *can* be explained by acclimatisation and familiarisation with the experiment, the task and the settings. However, still as illustrated in figure 2, some participants had a somewhat worse second experience.

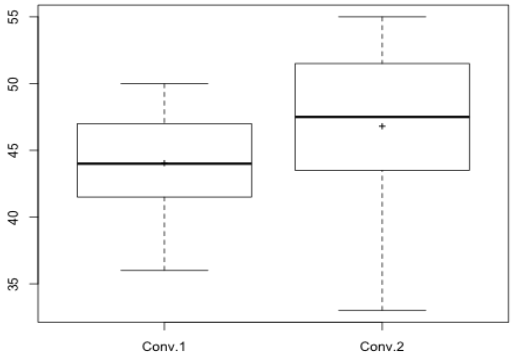

Fig. 2: Distribution of scores in the 1st and 2nd conversation, with an average increase of 2.75.

If we look more closely at the individual variation in the way the scores change between first and second conversation, we can observe that

three subjects stand out from the norm. Figure 3 shows the three outliers. The difference between first and second ratings for these three participants is remarkably high, for two of them in a positive direction (+11 and +10) and for the third one in a negative one (-6), whereas the average difference, as already noted, is +2.75. These three participants were singled out for further analysis.

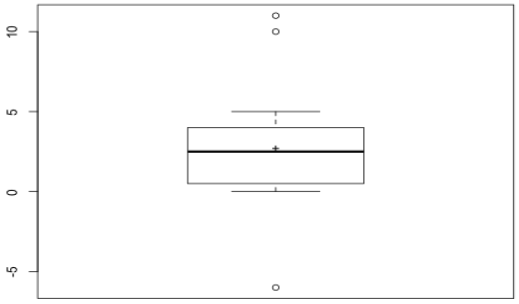

Fig 3: Comparing 1st and 2nd conversation scores and singling out outliers

3.2 Gestures and self-assessment scores

The only gestural annotation we consider here, as already mentioned, are the head movements (cf. table 1). The most frequent head movement types are nods, sideturns and tilts. The total number of head movements is 1427 in the first conversation, and 1563 in the second, with an increase of almost 10%. However, subjects differ substantially in this respect.

We decided to focus on head nods for several reasons. Nods are the most frequent head movement type; they are mostly tied to the functions of giving feedback and accepting the turn, both of which presumably relate to the subjects' experience of the conversation; finally, at first sight there seemed to be a slight correlation in the way the number of nods varied from conversation to conversation and the way the scores did.

The annotated head nods are 719 (single and repeated nods not differentiated). There are 338 head nods in the first set of interactions and 381 head nods in the second. This gives a +43 nod increase, in other words an increase of almost 13%. The head nod distributions in the two conversation sets are shown in figure 4.

Figure 5 shows the distribution of the *difference* in head nods between the two interactions. As can been seen, the mean individual increase is of 3.59 nods. The figure also shows that, as expected, some of the participants varied more than others. This time, no subject falls out of the general distribution.

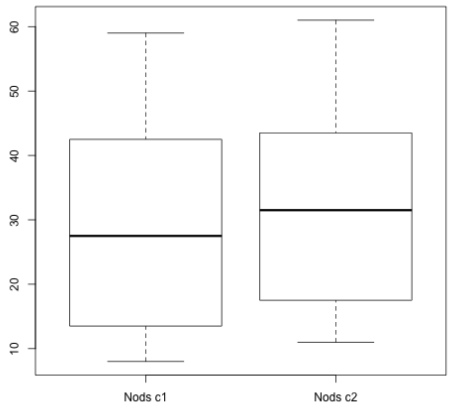

Fig 4: Head nods in 1st and 2nd conversation

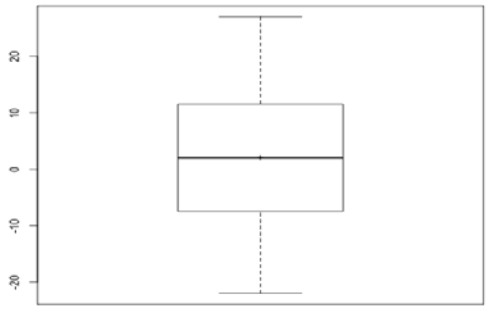

Fig 5. Distribution of difference in head nod number. The mean individual increase is +3.59.

A regression analysis was applied to the data to establish if there was a correlation between difference in questionnaire ratings and difference in the number of head nods produced by the participants in the two sets of conversations. Figure 6 visualises the regression line between the two parameters. In spite of the individual differences, the Pearson coefficient still indicates a weak positive linear association of .49 between difference in questionnaire ratings (self-assessments scores) and difference in number of head nods. These results seem to indicate that in general, a positive experience is associated with the production of a higher number of head nods. It is, however, only a slight tendency. There may be several other factors or individual differences affecting the production of head nods that are not accounted for in this study.

Fig. 6. Correlation between difference in scores and difference in number nods

In fact, the correlation is stronger (Pearson coefficient of 0.63) if the difference in head nods produced in the two conversations is compared to the difference in those ratings that only concern the last group of questions, in other words questions referring to the attitudinal emotions of the subjects during the interactions. This category comprises questions on whether the participants were pleased or upset, relaxed or tense, at ease or not at ease and content or irritated. The two sets of differences, and their relation, are shown in figure 7.

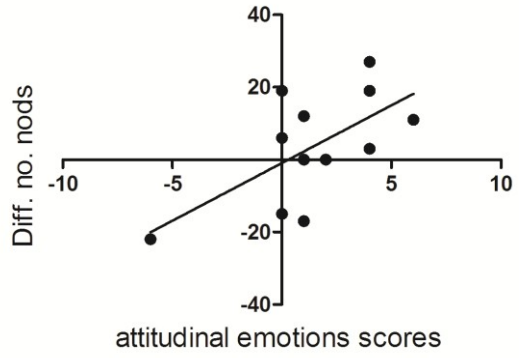

Fig. 7. Correlation between difference in scores referring to attitudinal emotions and difference in number of nods

3.3 Three outliers

Let us now turn to the three outliers we picked out from the analysis of the self-assessment scores because of the very different experiences they had in the two conversations. Figure 8 visualises the outliers compared to the average difference in self-assessment scores as well as number of head nods produced in the two conversations.

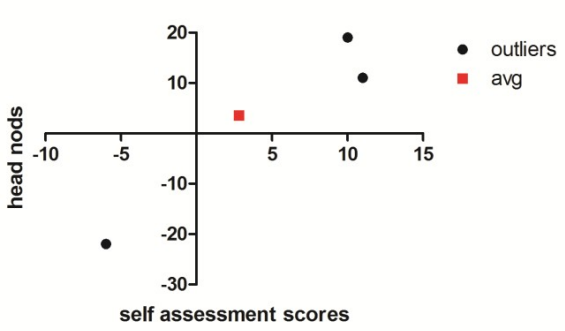

Fig. 8. Differences in number of nods and scores: three outliers and average

The figure shows that the three subjects' nodding behaviour is quite far from the average, and that it correlates with the polarity of the self-rating scores. This seems to suggest that for these three subjects, there is a strong correlation between the production of nods and the way in which they react emotionally to the conversational situation.. Figure 9 shows the difference in attitudinal emotions and number of head nods between the two sets of conversations. It again shows how the same three outliers differ markedly from average both in terms of rating their attitudinal emotions and in number of nods.

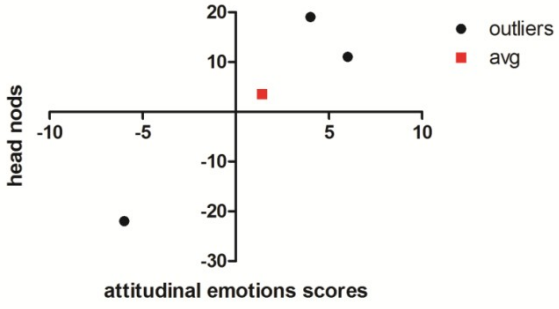

Fig. 9. Differences in number of nods and ratings of attitudinal emotions: three outliers and average.

As head nods in this corpus in many cases are tied to the function of feedback giving, it seems reasonable to predict that there may be a correlation between how positively a person perceives the conversation, and how much gestural feedback the person is giving their interlocutor, and that a difference in the first dimension between the two conversations should correspond to a difference in the second.

Looking at the feedback behaviour of the three subjects in the two conversations, however, does not provide clear indications of the fact that it should be directly related to the conversation experience. One of the three subjects, in fact, is quite close to the average in the feedback give

dimension, in other words this subject gives more or less the same feedback in the two conversations in spite of the difference number of head nods and the different self-rating scores. Therefore, the explanation for why a higher nod production seems to be associated with more positive emotional attitudes should probably be found in the complex of conversational functions that nods have, and which relate not only to feedback giving and eliciting, but also to the turn acceptance. In general, it could be said that head movements in general and nods in particular, signal a positive engagement in the conversation.

4 Conclusion

To sum up, the data provided seem to indicate a tendency to produce more head nods the more positive the subjects' experience of the conversation is and less head nods the more negative the experience is perceived. The tendency is strongest for three specific subjects, who showed a markedly positive or negative difference in the polarity of the attitudinal emotions they reported having in the two conversations. The main conversational functions of head nodding in the conversations studied here are giving or eliciting feedback as well as accepting the turn. Thus, the data seem to indicate that a more positive experience of the interaction is traceable either to smooth turn taking or to the participants' willingness and ability to signal that they want to continue the conversation, they are listening, paying attention to and understanding what's being conveyed.

The correlation observed in our corpus between attitudinal emotion ratings and head nodding is, however, not a strong one. To further confirm or disprove this indication, therefore, more data is needed. Other comparable corpora in the Nordic NOMCO corpus could be used to validate our results.

Another interesting avenue for further research that is related to the ideas discussed in this paper concerns emotions not only as they are remembered by the subjects after the experiment, but also as they are perceived by annotators based on the subjects' non-verbal behaviour. In the Danish NOMCO corpus, in fact, emotion labels have been added to the facial expressions using the P-A-D coding description (Mehrabian, 1996). Adding data from the emotional annotation to our analysis might help give a more precise explanation of the positive correlation we see between self-rating scores and non-verbal behaviour.

Acknowledgements

We'd like to acknowledge Costanza Navarretta, Anette Luff Studsgård and Sara Andersen, who all contributed to the annotation of the Danish corpus.

References

Allwood, J., Nivre, J. & Ahlsen, E. (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. Multimodal Corpora for Modelling Human Multimodal Behaviour. Special Issue of the International *Journal of Language Resources and Evaluation*, 41(3-4), 273-287.

Argyle, M. 1975. *Bodily communication*. London. Routledge.

Batrinca, L. M., Mana, N., Lepri, B., Pianesi, F. & Sebe, N. (2011). Please, Tell Me About Yourself: Automatic Personality Assessment Using Short Self-Presentations, (13th International Conference on Multimodal Interaction - ICMI 2011, Alicante, Spain).

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

Boholm, M. & Allwood, J. (2010). Repeated Head Movements, their Function and Relation to Speech. In Kipp, M., Martin, J.C., Paggio, P., & Heylen, D. (eds.) *Proceedings of the Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*. LREC 2010.Valletta, Malta, 18 May 2010, pages 6-10.

Brebner, J. (1985). Personality theory and movement. Individual differences in movement, 27-41.

Cerrato, L. (2007). Investigating Communicative Feed- back Phenomena across Languages and Modalities. Ph.D. thesis, Stockholm, KTH, Speech and Music Communication.

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

Ekman, P. (1972). *Emotion in the Human Face: Guide- Lines for Research and an Integration of Findings*. New York. Pergamon.

Helweg-Larsen, M., Cunningham, S.J., Carrico, A., & Pergram, A.M. (2004). To nod or not to nod: An ob- servational study of nonverbal communication and status in female and male college students. *Psychology of Women Quarterly* (28), 358-361.

Jokinen, K., Navarretta, C. & Paggio, P. (2008). Distin- guishing the communicative functions of gestures. In *Proceedings of the 5th MLMI, LNCS* 5237, pp. 38–49, Utrecht, September. Springer.

Kipp, M. (2004). Gesture Generation by Imitation. From Human Behavior to Computer Character Animation. Boca Raton, Florida: Dissertation.com.

Lippa, R. (1998). The nonverbal display and judgment of extraversion, masculinity, femininity, and gender diagnosticity: A lens model analysis. *Journal of Research in Personality* (32), 80-107.

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

Mehrabian, A (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology* (14), 261-292.

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

Nezlek, J. B. (2010). Multilevel modeling and cross-cultural research. In Matsumoto, D. & van de Vijver, A. J. R. (eds.) *Cross-Cultural research methods in psychology*. Oxford.

Paggio, P., Allwood, J., Ahlsén, E., Jokinen, K & Navarretta, C. (2010). The NOMCO multimodal Nordic resource - goals and characteristics. In *Proceedings of LREC 2010, Malta*, 17-23 May pp. 2968-2973.

Paggio, P. & Navarretta, C. (2010). Feedback in head gesture and speech. In Kipp, M. et al., (eds), *Proceedings of LREC-2010*, pages 1–4, Malta, May 17.

Paggio, P. & Navaretta, C. (2011a). Feedback and ges- tural behaviour in a conversational corpus of Danish. *Proceedings of the 3rd Nordic Symposium on Multimodal Communication. NEALT Proceedings Series* (15), 33–39.

Paggio, P. & Navarretta, C. (2011b). Learning to classify the feedback function of head movements in a Danish Corpus of first encounters. In Proceedings of ICMI 2011 Workshop *Multimodal Corpora for Machine Learning: Taking Stock and Road mapping the Future* , Alicante, Spain November 2011, 8 pages.

Paggio, P. & Navarretta, C. (2011c). Head movements, facial expressions and feedback in Danish first en- counters interactions: a culture-specific analysis. In C. Stephanidis, (Ed.), Universal Access in Human-Computer Interaction. Users Diversity. *Proceedings of 6th International Conference, UAHCI 2011*, Held as Part of HCI International, pp. 583–590, Orlando, FL, USA, July. Springer.

Riggio, R.& Friedman, H. (1986). Impression formation: The role of expressive behavior. In *Journal of Personality and Social Psychology*, 50 (2), 421-427.

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