The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena

5 authors, including:

Some of the authors of this publication are also working on these related projects:

Communication with horses View project

Intercultural Communication View project

phenomena are not detailed, because they only seek to capture features that are significant when studying interpersonal communication. However, the annotation of gesture shape and dynamics can be extended for specific purposes, for example to construct computer applications, without changing the functional level of the annotation.

The first kind of annotation considered is modality-specific, and concerns the expression types, the second concerns multimodal communication. For each hand gesture and facial display taken into consideration, a relation with the corresponding speech expression (if any) is also annotated. However, the scheme does not provide tags for the annotation of verbal expressions since the focus is on the facial displays and hand gestures which can be synchronized with spoken language.

Feedback

The production of feedback is a pervasive phenomenon in human communication. Participants in a conversation give feedback to show that they are willing and able to continue the interaction and that they are listening, paying attention, understanding or not understanding, agreeing or disagreeing with the message being conveyed. They elicit feedback to know how the interlocutor is reacting in terms of attention, understanding and agreement. While exchanging feedback, both speaker and listener can show emotions and attitudes. Both feedback giving and eliciting are annotated by means of the same three sets of attributes: *Basic, Acceptance*, and *Attitudinal emotions/attitudes*.

Function attribute	Function values
Basic	CPU, CP
Acceptance	Accept, Non-accept
Additional Emotion/ Attitude	Happy, Sad, Surprised, Disgusted, Angry, Frightened, Other

Table 1. Feedback Annotation Features

Basic features define hand gestures or facial displays in terms of whether they express or elicit i.

A Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing Phenomena

Jens Allwood
University of Göteborg
jens.allwood@ling.gu.se

Loredana Cerrato
TMH/CTT, KTH, Sweden
loce@speech.kth.se

Kristiina Jokinen
University of Helsinki
kristiina.jokinen@helsinki.fi

Costanza Navarretta, Patrizia Paggio
CST, University of Copenhagen
{Patrizia,Costanza}@cst.dk

ABSTRACT

This paper deals with the MUMIN multimodal annotation scheme, which is dedicated to the study of hand gestures and facial displays in interpersonal communication, with focus on the role played by multimodal expressions for feedback, turn management and sequencing. The scheme has been tested on the analysis of multimodal behaviour in short video clips in Swedish, Finnish and Danish. These preliminary results show that the categories defined are reliable, and points at a few necessary revisions.

Author Keywords

Multimodal annotation, feedback, hand and facial gestures

INTRODUCTION

The creation of a multimodal corpus often reflects the requirements of a specific application and thus constitutes an attempt at modelling either input or output multimodal behaviour. On the contrary the MUMIN coding scheme [4], developed in the Nordic Network on Multimodal Interfaces MUMIN (www.cst.dk/mumin), is intended as a general instrument for the study of hand gestures and facial displays in interpersonal communication, focusing on the role played by multimodal expressions for feedback, turn management and sequencing. It builds on previous studies of feedback strategies in conversations [9, 1], and on work where vocal feedback has been categorised in behavioural or functional terms [2,3,7]. In what follows, we briefly describe the annotation categories starting with the functional ones, and then deal with coding procedure, materials and results from three case studies. We conclude with a few reflections on the potential applications of the scheme.

ANNOTATION CATEGORIES

The main focus of the coding scheme is the annotation of the feedback, turn-management and sequencing functions of multimodal expressions, with important consequences for the annotation process and results. First of all, the annotator is expected to *select* hand gestures and facial displays to be annotated *only* if they play an observable communicative function. Moreover, the attributes concerning the shape or dynamics of the observed

continuation/contact and perception (CP), where the dialogue participants acknowledge contact and perception of each other; ii. continuation/contact, perception and understanding (CPU), where the interlocutors also show explicit signs of understanding or not understanding of the message. The two categories capture what [9] call *acknowledgement*. *Acceptance* indicates that the interlocutor has not only perceived and understood the message, but also shows or elicits signs of either agreeing with its content or rejecting it. Basic and Acceptance can be compared with process-related and content-related in [13]. Finally, feedback annotation relies on a list of *emotions* and *attitudes* that can co-occur with one of the basic feedback features and with an acceptance feature. The list includes the six basic emotions [111,5] plus an "other" value.

Turn management
The turn management system regulates the interaction flow and minimises overlapping speech and pauses. It is coded by the three general features *Turn gain*, *Turn end* and *Turn hold*. In addition, a turn gain is either a *Turn take* if the speaker takes a turn that wasn't offered, possibly by interrupting, or a *Turn accept* if the speaker accepts a turn that is being offered. Similarly, turn end can be achieved in different ways: the speaker can release the turn under pressure (*Turn yield*), offer the turn to the interlocutor (*Turn offer*), or signal completion of the turn and end of the conversation at the same time (*Turn complete*).

Sequencing
Sequencing concerns the organisation of a dialogue in meaningful sequences, corresponding to what in other frameworks has been described as sub-dialogues, i.e. a sequence of speech acts which may extend over several turns. In other words, sequencing is orthogonal to the turn system. *Opening sequence* indicates that a new speech act sequence is starting. *Continue sequence* indicates that the current speech act sequence is going on, for example when a gesture is associated with enumerative phrases such as "the first… the second… the third…". *Closing sequence* indicates that the current speech act sequence is closed, which may be shown by a head turn or another gesture while uttering a phrase like "that's it, that's all".

MULTIMODAL EXPRESSIONS
Under normal circumstances, in face-to-face communication feedback, turn management and sequencing all involve use of multimodal expressions, and are not mutually exclusive. For instance, turn management is partly done by feedback. A turn can be accepted by giving feedback and released by eliciting information from the other party. Within each feature, however, only one value is allowed, since the focus of annotation is on the explicit communicative function of the phenomenon under analysis. For example, a head nod which has been coded as CPU (continuation/contact, perception and understanding) cannot be assigned accept and non-accept values at the same time.

An example of a multifunctional facial display coded with ANVIL [12] is shown in the frame in Figure 1: the speaker frowns and takes the turn while agreeing with the interlocutor by uttering: "ja, det synes jeg" (Yes, I think so). By means of the same multimodal expression (facial display combined with speech utterance) he also elicits feedback from the interlocutor and encourages her to continue the current sequence.
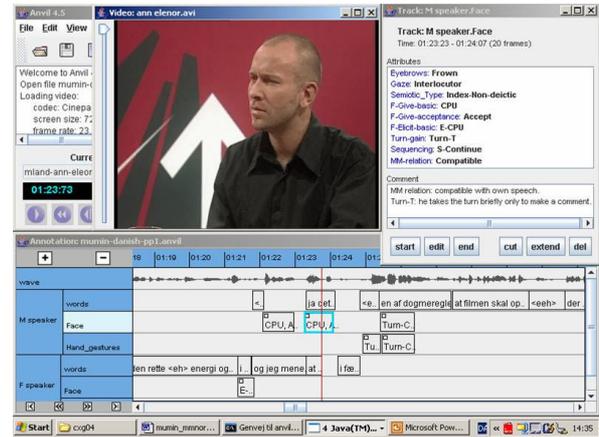

Figure 1: A multifunctional facial display: turn management and feedback

The components of a multimodal expression can have different time spans. For instance, a cross-modal relation can be defined between a speech segment and a slightly subsequent gesture. To define a multimodal relation, we make a basic distinction between two signs being *dependent* on or *independent* from each other. If they are dependent, they are either *compatible* or *incompatible*. For two signs to be compatible, they must either complement or reinforce each other, while incompatibility arises if they express different contents, as e.g. in ironic contexts.

FACIAL DISPLAYS AND HAND GESTURES
Facial displays and hand gestures are annotated with respect to the shape and dynamics of the movement. Although the categories proposed here, as already noted, are not very detailed, they should be specific enough to be able to distinguish and characterise non-verbal expressions that play a role in feedback, turn management and sequencing. They are concerned with the movement dimension of facial displays and hand gestures, and should be understood as dynamic features that refer to a movement as a whole or a protracted state. Internal gesture segmentation is not considered since it doesn't seem relevant for the analysis of communicative functions we are pursuing.

The term *facial display* [6] refers to timed changes in eyebrow position, expressions of the mouth, movement of the head and of the eyes. The coding scheme includes features describing *General face* expressions such as *Smile* or *Scowl*, features of *Eyebrow movements*, such as *Frown* or *Raise*, features referring to *Eye movement*, features for

Gaze direction, for movements of the *Mouth* and position of the *Lips.* Finally, a number of features refer to *Head* movements. The total number of different features is 36.

The annotation of the shape and trajectory of hand gesture is a strong simplification of the scheme used at the McNeill Lab [10]. The features, 7 in total, concern the two dimensions of *Handedness* and *Trajectory*, so that we distinguish between single-handed and double-handed gestures, and among a number of different simple trajectories analogous to what is done for gaze movement.

Finally, semiotic categories have also been defined common to both facial displays and hand gestures building on Pierce's semiotic types. They are *Indexical Deictic*, *Indexical*, *Non-deictic, Iconic* and *Symbolic*.

CODING PROCEDURE, TOOLS AND MATERIAL

The coding procedure was iteratively defined in several MUMIN workshops, and annotations have been carried out by means of the several coding tools, e.g. ANVIL [12]. The annotated material consists of a) one minute clip from an interview of a Danish actress for Danish television; b) one minute interview of the Finnish finance minister for Finnish television provided by the courtesy of the Centre of Scientific Computing; c) one minute clip from the Swedish film "Show me love".

The Danish case study

Two independent annotators with limited experience annotated gestures in the Danish clip using ANVIL. They started by annotating the non-verbal expressions of one of the interlocutors together to familiarise themselves with the coding scheme. Then they did the annotation task for the other dialogue participant independently in order to evaluate the reliability of the coding scheme.

In order to align the two annotations, it was decided that two segments referred to the same gesture if they covered the same time span, plus or minus ¼ of a second at the onset or end of the gesture. The first coder annotated 37 facial displays, and the second one 33. Of these, 29 were common to both coders. The agreement in recognition of facial gestures is thus 0.83. Concerning hand gestures, the first coder annotated 6, the second 4. Of these only two were in common. Therefore, only hand gestures have been considered for the κ-score evaluation.

The κ-scores obtained on the features concerning gesture shape and semiotic type are all in the range .83-.96 with the exception of those concerning *Gaze* (.54) and *Head* (0.2). This low agreement is partly due to the fact that one coder privileged head position over gaze (head up, no gaze), while the other in such cases ignored head movements and annotated gaze. There are also inconsistencies: in some cases the tag is *Gaze side* with the comment "away from the interlocutor", in others *Gaze other* with the comment "away from the interlocutor". Thus, the interaction of head

movement and gaze needs a more careful treatment in the coding manual.

In the coding of communicative functions, on the other hand (Table 2), the annotators achieved satisfactory κ-scores with the exception of *sequencing*, particularly the feature *Continue sequence*. The issue needs further investigation.

	P(A)	**P(E)**	**Kappa**
F-Give Basic	.79	.33	.68
F-Give acceptance	.86	.25	.81
F-Give Emotion	.86	.08	.84
F-Elicit basic	.93	.33	.9
F-Elicit acceptance	1	.25	1
F-elicit emotion	.93	.08	.92
Turn-gain	.89	.33	.83
Turn-end	.93	.33	.89
Turn-hold	.96	.05	.92
Sequencing	.69	.25	.59
MM-relation	.82	.25	.76

Table 2: κ-scores for classification of communicative function features

While they show a good reliability for most of the categories used, the κ-scores don't tell us anything about the coverage of the scheme. The material in the Danish case study is quite limited, so it is not surprising that many of the categories are not used. However, it is worth noting that one of the basic feedback features, *F-elicit-acceptance*, never appears (thus the κ-score concerns the default value "none"). The other case studies show that this is an idiosyncratic characteristic of this dialogue rather than evidence of empirical inadequacy of the feature. Concerning lack of necessary categories, on the other hand, it is obvious already from this limited study that body posture, which is not included in the scheme, is important for feedback: both coders noted in their comments that a relevant movement of the torso should have been annotated.

The Swedish and Finnish case studies

The Swedish video clip consists of a one-minute emotional conversation between two actors who interpret father and daughter. They are mostly filmed in close ups of their faces, so that the hands are rarely in the picture, making it impossible to annotate hand gestures. The actor that speaks is not always in focus, so in two cases in which the actors utters a feedback expression, the face cannot be observed.

Only one expert annotator coded the film scene, so the reliability of the coding scheme was evaluated only by means of an inter-variance test, which checks whether the same coder varies their judgments over time. The coder annotated the material once and after about six months repeated the coding. A total of 12 facial displays related to

feedback were coded both times, with complete intercoder agreement. The coded facial displays related to turn management functions were 12 the first time and 13 the second time, which means that the percentage of turn management identification was 95%.

Since the video-clip is extracted from a film, all the conversational moves are pre-defined and therefore only few turn-gain and turn-hold facial displays occur, moreover no sequencing facial displays or gestures were identified, probably due to the fact that the flow of discourse is pre-defined not leaving space to a spontaneous organisation of the discourse structure.

Given the emotional scene, it is not surprising that most of the feedback phenomena annotated have been labelled as *F-Give-emotion/attitude* (7, against 2 for *F-Elicit-acceptance*, and 1 for *F-Give-acceptance*, *F-Elicit-basic* and *F-Elicit-emotion/attitude*). The fact that *F-Elicit-acceptance* was used points to the fact that the category is useful, and that its absence from the Danish data is due to the different communicative situation. On the other hand, in the Swedish clip there are no examples of *F-Give basic*, which in spontaneous conversation has been found to be one of the most frequent feedback categories [8].

The distribution of turn management features was 10 for *Turn-end*, and 1 for *Turn-gain* and *Turn-hold*.

The Finnish 1-minute clip is similar to the Danish in that it is also an interview edited for broadcasting. The most important contribution of this study – still in the process of being analysed – again points to the fact that a broader selection of gestures are needed to cover the analysis of communicative functions. In particular, tilting of the head was recurrently used by the interviewee to elicit feedback from the interviewer.

CONCLUSIONS

The MUMIN coding scheme constitutes an attempt at defining a scheme for the annotation of feedback, turn management and sequencing multimodal behaviour in human communication. The preliminary results of the reliability test run in the Danish study case confirm the general reliability of the categories defined for the purpose of coding feedback and turn taking functions, although gaze, head and sequencing features seemed problematic in some cases, and not enough detailed in others (Finnish results). Body posture, which is not part of this version of the coding scheme, is a needed extension. Future revisions and extensions to the current version of the scheme will seek to accommodate these problems. We are now gathering additional experience by applying the coding scheme in graduate courses on multimodal communication.

The availability of such a scheme is an important step towards creating annotated multimodal resources for the study of multimodal communicative phenomena in different situations and different cultural settings, and for investigating many different aspects of human communication.

Examples of issues that can be investigated empirically by looking at annotated data are to what extent gestural feedback co-occurs with verbal expressions; in what way different non-vocal feedback gestures can be combined; whether specific gestures are typically associated with a specific function; how multimodal feedback, turn management and sequencing strategies are expressed in different cultural settings.

REFERENCES

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

2. Allwood, J. Dialog Coding – Function and Grammar. Gothenburg Papers. *Theoretical Linguistics*, 85. Gothenburg University, 2001.

3. Allwood, J. and Cerrato, L. A study of gestural feedback expressions. In Paggio et al (eds.) *First Nordic Symposium on Multimodal Communication*, 2003.

4. Allwood, J., Cerrato, L., Dybkær, L., Jokinen, K., Navarretta, C. and Paggio, P. *The MUMIN multimodal coding scheme*. Technical report available at www.cst.dk/mumin/stockholmws.html, 2004.

5. Beskow J., Cerrato L., Granström B., House D., Nordstrand M., Svanfeldt G. The Swedish PF-Star Multimoda Corpora. *LREC Workshop on Models of Human Behaviour*, 2004.

6. Cassell, J. Nudge Nudge Wink Wink: Elements of Face-to-Face Conversation for Embodied Conversational Agents, in Cassell, J. et al. (eds.), *Embodied Conversational Agents* (2000), pp. 1–27. Cambridge, MA: MIT.

7. Cerrato, L. A coding scheme for the annotation of feedback phenomena in conversational speech. *LREC Workshop on Models of Human Behaviour*, 2004.

8. Cerrato, L. Some characteristics of feedback expressions in Swedish, TMH.OPSR Vol.43 *Fonetik* (2002), pp. 101-104.

9. Clark H. and Schaefer E. Contributing to Discourse. *Cognitive Science* 13 (1989), pp. 259–94.

10. Duncan, S. *McNeill Lab Coding Methods*. Available from http://mcneilllab.uchicago.edu/topics/proc.html (last accessed 26/4/2004).

11. Ekman P. Basic emotions. In T. Dagleish and T. Power (eds.) *The Handbook of Cognition and Emotion* NY J. Wiley, 1999, pp. 45–60.

12. Kipp, M. Anvil – A Generic Annotation Tool for Multimodal Dialogue. In Eurospeech 2001. pp. 1367–1370.

13. Thórisson, Kristinn R. Natural Turn-Taking Needs No Manual: Computational Theory and Model, from Perception to Action. In B. Granström et al (eds.) *Multimodality in Language Speech Systems*. Kluwer Academic: Dordrecht, the Netherlands, pp. 173–207, 2002.