In Proceedings from the Second Nordic Conference on
Multimodal Communication, Göteborg, 2005

THE MUMIN ANNOTATION SCHEME FOR FEEDBACK, TURN MANAGEMENT AND SEQUENCING

Jens Allwood (1), Loredana Cerrato (2), Kristiina Jokinen (3), Costanza Navarretta (4), and Patrizia Paggio (4).
(1) University of Göteborg, (2) TMH/CTT, KTH, Sweden
(3) University of Helsinki, (4) CST, University of Copenhagen

Abstract

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

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

1. The MUMIN annotation scheme

The creation of annotated multimodal corpora is being recognised by a growing number of researchers, initiatives and organisations[1] as a prerequisite for the creation of more natural human-computer interfaces based on models of human behaviour. However, there is still a lack of

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

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

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

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

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

[2] We use "gesture" as a general term for non-verbal expressions, in our case hand gestures and facial displays.

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

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

2. Annotation categories

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

2.1 Categories of feedback, turn management and sequencing

The main purpose of the annotation is to capture the way in which facial displays and hand gestures, possibly in combination with verbal expressions, contribute to the general communicative phenomena of *feedback (give* or *elicit)*, *turn management* and *sequencing*. These three functions constitute the backbone of the scheme, and are intended to guide the selection of the gestures to be annotated. In defining the features for the annotation of feedback, turn management and sequencing, we have profited from an extensive number of references in which these phenomena are treated from the point of view of verbal expressions. We believe the features in the coding scheme are applicable to the annotation of non-verbal and multimodal expressions for which they have been designed, and the preliminary results described in this paper confirm our belief. However, these results will have to be validated by applying the scheme to more practical coding tasks.

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

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

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

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

Acceptance, which is a boolean feature, indicates that the subject has not only perceived and understood the message, but also shows or elicits signs of either agreeing with its content or rejecting it, e.g. by different head movements. Acceptance is treated as a separate dimension, different from understanding, also in coding schemes for dialogue annotation. For instance, the DAMSL coding scheme distinguishes between *understanding* (“Huh”, “What?”, “I see”) and *agreement* (“Yes”, “No”, “Sounds good”).

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

Table 1. Feedback giving annotation features

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

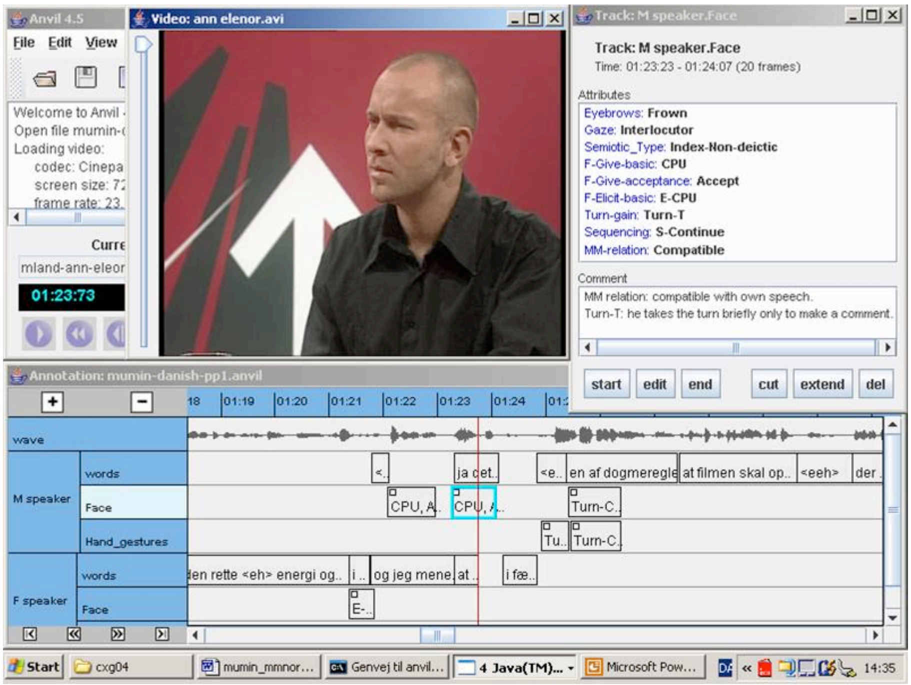
If feedback is the machinery that crucially supports the success of the interaction in interpersonal communication, the flow of the interaction is also dependent on the turn management system. Optimal turn management has the effect of minimising overlapping speech and pauses in the conversation. Turn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

Table 2. Turn management annotation features

Function attribute		**Function value**
TURN MANAGEMENT	Turn-gain	Turn-take
		Turn-accept
	Turn-end	Turn-yield
		Turn-elicit
		Turn-complete
	Turn-hold	Turn-hold

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

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



Track: M speaker.Face
Time: 01:23:23 - 01:24:07 (20 frames)

Attributes

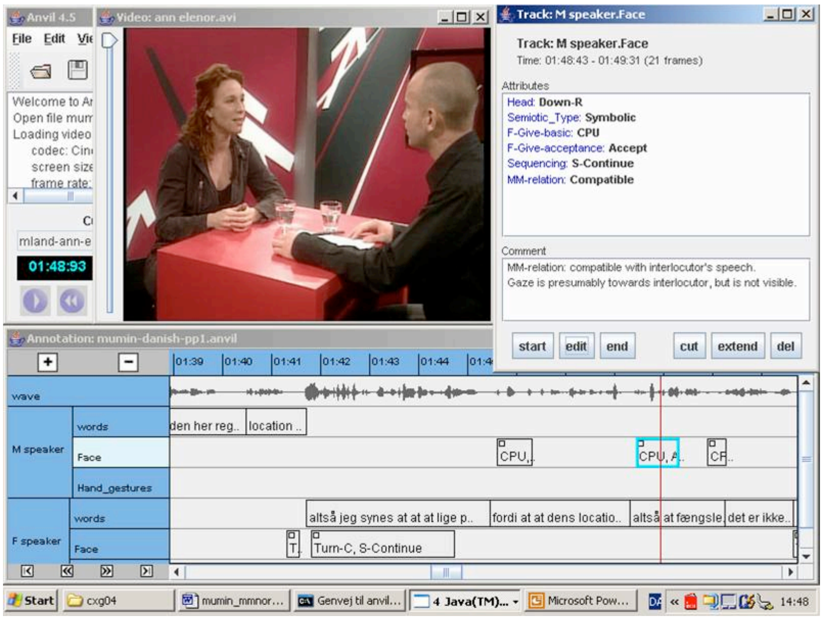
Eyebrows: Frown
Gaze: Interlocutor
Semiotic_Type: Index-Non-deictic
F-Give-basic: CPU
F-Give-acceptance: Accept
F-Elicit-basic: E-CPU
Turn-gain: Turn-T
Sequencing: S-Continue
MM-relation: Compatible

Comment

MM relation: compatible with own speech.
Turn-T: he takes the turn briefly only to make a comment.

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

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



Track: M speaker.Face
Time: 01:48:43 - 01:49:31 (21 frames)

Attributes

Head: Down-R
Semiotic_Type: Symbolic
F-Give-basic: CPU
F-Give-acceptance: Accept
Sequencing: S-Continue
MM-relation: Compatible

Comment

MM-relation: compatible with interlocutor's speech.
Gaze is presumably towards interlocutor, but is not visible.

Figure 2. Basic feedback and acceptance by facial expressions

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

2.2 Facial displays and hand gestures

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

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

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

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

2.3 Multimodal features

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

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

3. Annotation procedure and material

The coding procedure was iteratively defined in the MUMIN workshops and steering group meetings. Furthermore, the MUMIN annotators were given a tutorial on how to annotate by means of the three coding tools ANVIL (Kipp 2001), MultiTool (Gunnarsson 2002) and NITE (Bernsen et al 2002).

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

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

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

4. First case study: the Danish annotation

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

The annotation has been evaluated based on the strategy described by Carletta *et al* (2004). First of all, a method for aligning the annotations of the coders had to be established: it was decided to accept a difference in time coding of under one fourth of a second per segmentation. In other words, if both coders annotated a gesture within the same time span apart from a possible difference in start and/or end of under ¼ of a second, it was assumed that the two segments described the same expression. In all the cases where both coders annotated the same gesture, there was agreement of segmentation, with the exception of one case in which one coder recorded one facial display as a unit, while the other split the same display into two (i.e. the two segments in one annotation covered temporally the same time span of one segment in the second annotation).

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

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

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

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

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

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

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

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

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

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

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

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

The reason for the low agreement on gaze features was partly due to the fact that one coder encoded gaze relative to the head position (head up, no gaze), while the other coder chose to annotate the gaze instead of the head when the head movement was little (no head movement, gaze up). Furthermore, the two coders used different strategies for gaze. In some cases they coded "gaze:side" with the comment "away from the interlocutor", in some cases "gaze:other" with the comment "away from the interlocutor". Thus, the interaction of head movement and gaze is an issue that the manual does not seem to treat satisfactorily.

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

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

5. Second case study: the Swedish annotation

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

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

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

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

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

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

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

6. Conclusion

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

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

References

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

Allwood, J. (2001). *Dialog Coding – Function and Grammar*. Gothenburg Papers in Theoretical Linguistics, 85. Department of Linguistics, Gothenburg University.

Allwood J., & Cerrato L (2003). A study of gestural feedback expressions. In Paggio *et al* (Eds) *Proceedings of the First Nordic Symposium on Multimodal Communication*, Copenhagen.

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

Bernsen, N. O., Dybkjær, L., & Kolodnytsky, M. (2002). THE NITE WORKBENCH - A Tool for Annotation of Natural Interactivity and Multimodal Data. *Proceedings of the Third International Conference on Language Resources and Evaluation* (LREC'2002), Las Palmas, May 2002.

Beskow J., Cerrato L., Granström B., House D., Nordstrand M., & Svanfeldt G. (2004). The Swedish PF-Star Multimoda Corpora. *LREC Workshop on Multimodal Corpora: Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces,* Lisboa 25 May 2004.

Carletta J., Isard A., Isard S., Kowto J.C., Doherty-Sneddon G., & Anderson A. H (1997). The Reliability of a Dialogue Structure Coding Scheme. In *Computational Linguistics* 23(1), 13–31.

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

Cerrato, L. (2004). A coding scheme for the annotation of feedback phenomena in conversational speech. In *Proceedings of the LREC Workshop on Mulitmodal Corpora: Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces*, Lisboa, 25 May 2004.

Clark H., & Schaefer E. (1989). Contributing to Discourse. In *Cognitive Science* 13, 259–94.

Core, M., & J. Allen (1997). Coding Dialogs with the DAMSL Annotation Scheme. Presented at *AAAI Fall Symposium on Communicative Action in Humans and Machines, Boston, MA, November 1997. ftp://ftp.cs.rochester.edu/pub/papers/ai/97.Core-Allen.AAA12.ps.gz*

Cowie R. (2000). Describing the emotional states expressed in speech, in *Proc. of ISCA Workshop on Speech and Emotion*, Belfast 2000, pp. 11–19.

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

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

Gunnarsson, M. (2002). *User Manual for MultiTool.* Available from /www.ling.gu.se/~mgunnar/multitool/MT-manual.pdf.

Kipp, M. (2001). Anvil – A Generic Annotation Tool for Multimodal Dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, pp. 1367–1370. Aalborg.

Martin, J. C., Den Os, E., Kühnlein, P., Boves, L., Paggio, P., & Catizone, R. (2004). *Proceedings of the LREC workshop on Multimodal Corpora: Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces*, held in conjunction with LREC04, Lisbon, May 2004.

Poggi, I., & Magno Caldognetto, E. (1996). A score for the analysis of gestures in multimodal communication. In: *Proceedings of the Workshop on the Integration of Gesture and Language in Speech*. Applied Science and Engineering Laboratories. L. Messing, Newark and Wilmington, Del, 235–244.

Siegel S., & Castellan N.J.jr (1988). *Nonparametric Statistics for the Behavioral Sciences*, second edition. McGraw-Hill.