The MUMIN multimodal coding scheme

Jens Allwood[1], Loredana Cerrato[2], Laila Dybkjaer[3], Kristiina Jokinen[4], Costanza Navarretta[5] and Patrizia Paggio[5]

(1) University of Göteborg
jens@ling.gu.se

(2) Dept. of Speech Music and Hearing and Centre for Speech Technology,
KTH, Stockholm
loce@speech.kth.se

(3) University of Southern Denmark - Odense
laila@nis.sdu.dk

(4) University of Helsinki
kristiina.jokinen@helsinki.fi

(5) CST, University of Copenhagen
{costanza|patrizia}@cst.dk

Abstract

The MUMIN multimodal coding scheme was originally created to experiment with annotation of multimodal communication in short clips from movies and in video clips of interviews taken from Swedish, Finnish and Danish television broadcasting. However, the coding scheme also intends to be a general instrument for the study of gestures and facial displays in interpersonal communication, in particular the role played by multimodal expressions for feedback, turn management and sequencing. The first coding experiment was carried out at a workshop at KTH, Stockholm, on 21-22 June 2004. The version V3.4 of the coding scheme, presented in this paper, is the result of comments and discussions during and after the workshop.

Keywords: multimodal annotation, coding schemes for facial display and gesture annotation, feedback, turn-taking, sequencing

1. Uni-modal and multimodal annotation

Two kinds of annotation are considered. The first is modality-specific, and concerns the expression types indicated in Table 1, with the exception of those indicated in parentheses. For each expression type, levels of annotation and annotation tags are defined and exemplified below in Section 3.

Caveat: in this version of the coding scheme, no tags are defined for speech or dialogue act annotation. Several possibilities, including a reduced version of the DAMSL annotation tag set[1] or the tag set proposed by Allwood et al (2003)[2], have been taken into consideration and may be added later.

Modality	**Expression type**
Facial displays	Eyebrows Eyes Gaze Mouth Head
Gestures	Hand gestures (Body posture)
Speech	Segmental (Suprasegmental)

Table 1: Unimodal annotation level

The second kind of annotation concerns multimodal communication. For each gesture and facial expression taken into consideration, a relation with the corresponding speech expression (if any) is also annotated. Note that in a dialogue, gesture/facial display by one person may relate to speech by another. The correspondences foreseen for a two-party dialogue are shown in Table 2.

	Gesture/facial display speaker 1	**Gesture/facial display speaker 2**
Speech speaker 1	within-speaker	across-speakers
Speech speaker 2	across-speakers	within-speaker

Table 2: Multimodal correspondences in two-party dialogue.

[1] S*ee* *www.cs.rochester.edu/research/cisd/resources/damsl/RevisedManual/*

[2] Also available at*: www.gslt.hum.gu.se/~leifg/doc/allwood_long.ps*

2. Coding levels

For each modality expression, two levels of complexity are considered. One relates to the form of the expression, and the other to its semantic-pragmatic function. Note that these should not be understood as sequential with respect to each other, or leading an independent existence. They simply correspond to different aspects in the annotation matrix. The annotations for the first level are quite coarse. As for the second level, emphasis is put on the communicative function of the expression, and in particular its feedback, turn-managing or sequencing function.

3. Phenomena to be annotated

3.1 Communicative functions

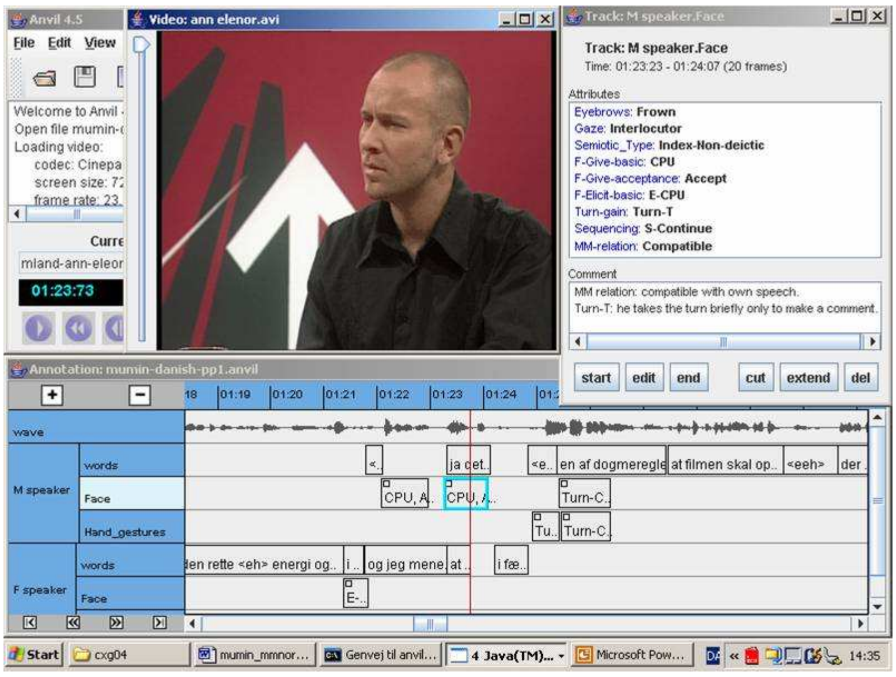
The main focus of the coding scheme is the annotation of feedback, turn-managing and sequencing functions of multimodal expressions, as well as the way in which expressions belonging to different modalities are combined. We focus then on three general communicative functions, one of which – feedback – combines the two aspects of feedback give and feedback elicit:

- feedback (give / elicit)
- turn-managing;
- sequencing.

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. For example, mechanical recurrent blinking of the eyes due to dryness of the eye 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, and only seek to capture features that are significant when studying interpersonal communication.

The three functions that constitute the backbone of the scheme, and which are intended to guide to selection of the gestures to be annotated, are not to be seen as mutually-exclusive. In other words, a communicative sign – whether uni- or multimodal – may well, and often does, play several communicative functions at the same time. It may be *multifunctional*.

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: 01:23:23 - 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.

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.

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.

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.

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. It may be argued that information structuring is also relevant for interpersonal communication, and that since gestures contribute to it, it should be included in the scheme. It would certainly be a relevant extension to the dimensions of communication considered here.

The specific tags for the annotation of feedback, turn-management and sequencing are shown in Table 3. Note again that these features are not mutually exclusive. For instance, turn managing 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.

In reality, some of the feature combinations allowed by the scheme may not be empirically meaningful, and some may be difficult to observe. However, we will leave it to empirical investigation to determine this. Another issue is how specific the annotator needs to be. This clearly depends on the specific interests, and an implementation of the scheme ought to allow for the possibility of either choosing a terminal value (e.g. a specific emotion like *anger*), or a more general one (e.g. *attitudinal emotion*, meaning that there is *some* emotion, without further specification).

Let us now look at the various features in more detail.

3.1.1 Feedback

Both Feedback Give and Feedback Elicit are described in terms of the same three sets of attributes, called **Basic**, **Acceptance**, and **Attitudinal emotions/Attitudes**.

Basic

- **Continuation/Contact**: indicates that the subject shows or elicits willingness to establish or maintain contact and to go on in the communication.
- **Perception**: indicates that the subject shows to have perceived or elicits signs of the interlocutor having perceived the message.
- **Understanding:** indicates that the subject shows to have understood or elicits signs of the interlocutor having understood the message.

The three basic feedback features are dependent on each other in such a way that Understanding presupposes Perception which in turn presupposes Contact. Therefore, three possible combinations of the three features could be envisaged. However, it is not totally clear if feedback can ever indicate pure Continuation/Contact without at least some degree of Perception, so only two combinations are allowed in the scheme:

- **CPU**: Most often a feedback sign can be characterised by all three of them at the same time.
- **CP**: Sometimes, a gesture or a verbal expression may convey Continuation/Contact and Perception without Understanding, as in the case of accepting an order one doesn’t understand.

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 interlocutors to signal that a contribution has been understood well enough to allow the conversation to proceed.

In using these categories, the annotator must not be concerned with whether the subject does or doesn’t perceive the message completely or correctly, nor is it relevant to worry about whether the subject doing a feedback understanding gesture has really understood what is being conveyed. What matters is whether the gesture that is being annotated seems to give or elicit feedback relating to one or more of the CPU categories.

Function feature		Specific function value	Short tag
FEEDBACK GIVE	Basic	Contact/continuation Perception Understanding	CPU
		Contact/continuation Perception	CP
	Acceptance	Accept Non-accept	Accept Non-accept
	Additional Emotion/Attitude	Happy Sad Surprised Disgusted Angry Frightened Certain Uncertain Interested Uninterested Disappointed Satisfied Other	
FEEDBACK ELICIT	Basic	E-Contact/continuation Perception Understanding	E-CPU
		E-Contact/continuation Perception	E-CP
	Acceptance	E-Accept E-Non-accept	E-Accept E-Non-accept
	Additional Emotion/Attitude	Happy Sad Surprised Disgusted Angry Frightened Certain Uncertain Interested Uninterested Disappointed Satisfied Other	

TURN-MANAGEMENT	Turn-gain	Turn-take Turn-accept	Turn-T Turn-A
	Turn-end	Turn-yield Turn-elicit Turn-complete	Turn-Y Turn-E Turn-C
	Turn-hold		Turn-H
SEQUENCING	Opening sequence		S-Open
	Continue sequence		S-Continue
	Closing sequence		S-Close

Table 3: Communicative Functions

Acceptance

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 accepting or rejecting its content, 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”).

- **Accept**: indicates that the subject shows or elicits signs of acceptance.
- **Non-accept**: indicates that the subject shows or elicits signs of refusal, non-acceptance of the information received.

Attitudinal emotions/attitudes

The scheme contains 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.

3.1.2 Turn management

Turn management has three general features:

1. **Turn gain**: when the speaker gains the floor. This can be done in two different ways depending on whether the turn is changing in agreement between the two speakers or not:
   - **Turn take**: when the speaker takes a turn that wasn't offered, possibly by interrupting.
   - **Turn accept**: when the speaker accepts a turn that is being offered.

2. **Turn end**: when the speaker gives up their turn. This can again happen in concordance with the interlocutor or not, and also without offering the turn. Thus we have three categories.
   - **Turn yield**: when the speaker releases the turn under pressure.
   - **Turn elicit**: when the speaker offers the turn to the interlocutor.
   - **Turn complete**: when 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, for instance by looking down to a newspaper.

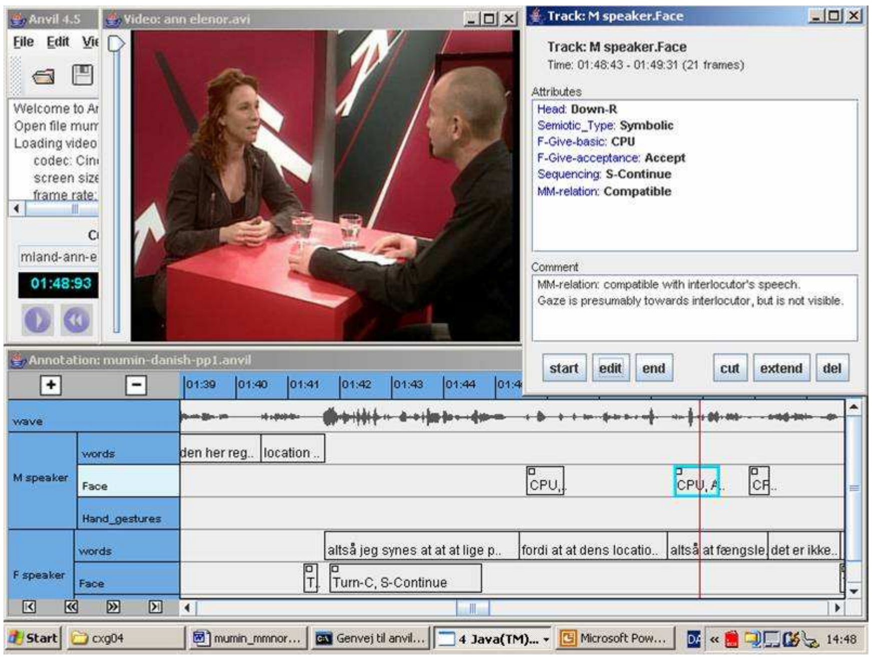
3. **Turn holding**: when the speaker wishes to keep the turn (this is usually done by rotating the head and the gaze away from the listeners).

3.1.3 Sequencing

The features of sequencing are:

- **Opening sequence**: indicates that a new speech act sequence is starting, for example a gesture occurring together with the phrase "by the way…".
- **Continue sequence**: indicates that the current speech act sequence is going on, for example a gesture occurring together with enumerative phrases such as "the first… the second… the third…".
- **Closing sequence**: indicates that the current speech act sequence is closed, for example a gesture occurring together with phrases such as "that's it, that's all".

Figure 2 shows a frame of a sequence in which several of the feedback categories defined above have been observed by an annotator. The 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.



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

3.2 Gestures

Table 4 shows the categories used to annotate gestures. A distinction is generally made between hand gestures and body posture. Body posture, however, has not be studied here: therefore, no relevant tags have been defined. The categories used to annotate hand gestures are taken mainly from McNeill (1992) and Allwood (2002), and build on Peirce's work with respect to the semiotic types.

Hand gesture annotation presupposes first of all that the so-called gesture phrases are identified, in other words that the annotator finds the gestures they want to annotate, and establishes where each gesture starts and ends. Selection is guided by the communicative functions we are interested in. Just as in the case of facial displays, which are treated in the next section, these are feedback-related, turn-management and sequencing functions. As far as start and end points are concerned, in order to simplify the work we do not try to capture the internal structure of a gesture phrase (preparation, stroke and retraction phases).

Gestures	**Shape of gesture**	
Hand gestures	Handedness	Both-H both hands Single-H single hand
	Trajectory	Up Down Sideways Complex Other
	Semantic-pragmatic analysis	
	Semiotic types	Indexical Deictic Indexical Non-deictic Iconic Symbolic
	Communicative function	Feedback give Feedback elicit Turn managing Sequencing

Table 4: Gesture annotation scheme

The tagging of the shape of hand gestures is quite coarse, and much simplified compared with the coding scheme used at the McNeill Lab, which has been our starting point. We only look at the two dimensions *Handedness* and *Trajectory*, without worrying about the orientation and shape of the various parts of the hand(s), and we define trajectory in a very simple manner, analogous to what is done below for gaze movement. There are thus a number of ways in which the coding of gesture shapes could be further developed for different purposes and applications.

The semantic-pragmatic analysis consists of two levels. The first is a categorisation of the gesture type in semiotic terms, the second concerns the communicative functions of gestures. Both levels also apply for facial displays, see below Section 3.3. Communicative functions have already been discussed above, whereas the semiotic types will be explained below. Cross-modal functions, which have not been defined specifically for gestures, are discussed in Section 3.5.

Below we describe each tag in more detail.

Handedness

- **Both hands**: both hands are involved
- **Single hand:** either right or left hand are involved alone

Trajectory

- **Up:** the stroke of the gesture is upwards
- **Down:** the stroke of the gesture is downwards
- **Sideways:** the stroke of the gesture is sideways
- **Complex:** the gesture is a complex combination of Up, Down and Sideways
- **Other.**

Gesture types

- **Indexical Deictic** gestures locate aspects of the discourse in the physical space (e.g. by pointing). According to Cassell (to appear), they can also be used to index the addressee. The example Cassel gives is when a teacher in the classroom says "yes, you are exactly right" and points at a particular student.
- **Indexical Non-deictic** gestures also indicate via a causal relation between the gesture and the effect it establishes. The small movements that accompany speech and underline its rhythm, and that some people have called *batonic* or *beat* gestures, fall into this category.
- **Iconic** gestures express some semantic feature by similarity or homomorphism. Examples are gestures done with two hands to comment on the size (length, height, etc.) of an object mentioned in the discourse. Some researchers distinguish *metaphoric* gestures as a separate type. Examples are conduit metaphors, which are often used in gestures accompanying concepts that refer to information and communication (as in a 'box' gesture while saying "in this part of my talk…"). In the MUMIN scheme we do not distinguish between iconic and metaphoric, since they can both be characterised by the fact that they express a concept by similarity.
- **Symbolic** gestures (emblems) are gestures in which the relation between form and content is based on social convention (e.g. the okay gesture). They are culture-specific.

3.3 Facial displays

The term *facial displays* refers, according to Cassel, to timed changes in eyebrow position, expressions of the mouth, movement of the head and of the eyes. Facial displays can be characterised by a description of the muscles or part of the body involved in the movement, or the amount of time they last, but they can also be characterised by their *function* in conversation.

The MUMIN coding scheme specifies features belonging to the movement dimension, and proposes to annotate the communicative function of facial displays in terms of the features defined in Table 5. The dimension concerning the movement expression uses rather coarse-grained features. All of them should be understood as dynamic features that refer to the movement as a whole or a protracted state, rather than punctual categories referring to different stages of the 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. Furthermore, we do not consider internal gesture segmentation since it doesn't seem very relevant for the analysis of communicative functions we are pursuing. However, nothing hinders annotators to extend the scheme in the direction of a more precise characterisation of the dynamics of gestures.

Remember that the goal of the coding scheme is only to annotate expressions that have a specific communicative function rather than the whole stream of facial displays throughout a conversation. Therefore, the annotator is not expected to code "neutral" facial displays or facial displays due to other factors, e.g. a frowning expression due to direct sun light. There may be cases in which a communicative gesture that requires annotation (see the next section) may occur together with a neutral facial display, or examples of communicative facial displays in which one part of the face may move in a characteristic way while other parts remain neutral. In such cases, only the part of the multimodal expression that shows a movement or a state different from the default neutral one should be annotated.

Facial displays can have phonological functions (for example articulatory gestures), they can have grammatical functions (for example eyebrow raising on pitch accented words), they can have semantic functions (for example nods and smiles to express feedback) and they can also have social functions (for instance politeness smile). As already mentioned, we will focus on feedback, turn-management and sequencing functions. We also propose a number of semiotic categories – the same as for gestures – in which facial displays can be grouped.

A coding scheme for the two levels of coding of facial displays is shown in Table 5 and Table 6. Tags concerning the relationship between the facial display and speech are defined and explained in Section 3.5.

The background assumption for coding is that we code those facial displays and gestures which have either a feedback or a turn-management function, or that we assign facial display values co-occurring with either a gesture or a verbal message that have a feedback or turn-managing function. Details on each tag are given below.

General face refers to the general impression that the coder gets from the facial expression of the subject under analysis. The general face can be labelled in terms of:

- **Smile:** when the facial expression shows pleasure, favour, or amusement, but sometimes derision or scorn. Smile is characterized by an upturning of the corners of the mouth and usually accompanied by a brightening of the face and eyes.
- **Scowl:** when the facial expression shows displeasure, scowl, anger**.** Scowl can be characterized by draw down or contract the eyebrows (i.e. frown) in a sullen, displeased or angry manner and may be accompanied by a down turning of the corners of the mouth and usually dull, grim face and eyes.
- **Laughter:** when the facial expression or appearance shows merriment or amusement, but also derision or nervousness and it is accompanied by an audible vocal expulsion of air from the lungs that can range from a loud burst of sound to a series of chuckles.
- **Other.**

Eyebrows movements are labelled in terms of:

- **Frowning**: when the eyebrows contract and move towards the nose.
- **Raising**: when the eyebrows are lifted.
- **Other.**

Facial display feature		**Form of expression/ Movement values**	
		Value	**Short tag**
General face		Smile	Smile
		Laughter	Laugh
		Scowl	Scowl
		Other	Other
Eyebrows		Frowning	Frown
		Raising	Raise
		Other	Other
Eyes		Exaggerated Opening	X-Open
		Closing-both	Close-BE
		Closing-one	Close-E
		Closing-repeated	Close-R
		Other	Other
Gaze		Towards interlocutor	Interlocutor
		Up	Up
		Down	Down
		Sideways	Side
		Other	Other
Mouth	**Openness**	Open mouth	Open-M
		Closed mouth	Close-M
	Lips	Corners up	Up-C
		Corners down	Down-C
		Protruded	Protruded
		Retracted	Retracted
Head		Single Nod (Down)	Down
		Repeated Nods (Down)	Down-R
		Single Jerk (Backwards Up)	BackUp
		Repeated Jerks (Backwards Up)	BackUp-R
		Single Slow Backwards Up	BackUp-Slow
		Move Forward	Forward
		Move Backward	Back
		Single Tilt (Sideways)	Side-Tilt
		Repeated Tilts (Sideways)	Side-Tilt-R
		Side-turn	Side-Turn
		Shake (repeated)	Side-Turn-R
		Waggle	Waggle
		Other	Other

Table 5: Coding scheme for facial displays: form

Semantic-pragmatic analysis	
Semiotic types	Indexical Deictic Indexical Non-deictic Iconic Symbolic
Communicative function	Feedback give Feedback elicit Turn managing Sequencing

Table 6: Coding scheme for facial displays: function

Eyes refer to movements of the eyelids and not to gaze, which is treated below. Those eye movements that do not carry a communicative function (such as biological blinking to keep the eyes wet) will not be annotated.
Eye movements are labelled as:

- **Exaggerated Opening:** when the eyes are wide open as in the case of surprise.
- **Closing-both:** when the eyes are both closed and this facial display is not a biological blinking. Closing both eyes can occur to underline when a word bears the focus.
- **Closing-one:** when one eye winks, that is opens and closes quickly. **Closing-both:** when both eyes wink, that is open and close quickly.
- **Other.**

Caveat: For the sake of simplicity we do not separate the coding for left and right eye.

Gaze direction*:* gaze refers to “an individual’s looking behaviour, which may or not be at the other person” (Knapp and Hall 2002, p.349). Gaze is used to regulate the flow of conversation, by managing turn regulation and monitoring feedback, but also by expressing emotions and communicating the nature of the interpersonal relationship. It is labelled as:

- **Towards interlocutor**: the person under observation appears to be looking towards the interlocutor. In a conversation, this corresponds to neutral, or normal behaviour. In fact, normally the two interlocutors will be looking at each other. In practice, however, it is often impossible in videos to actually see a mutual gaze, since the camera focuses on one speaker at time.

- **Up**: when the person looks up.
- **Down**: when the person looks down.
- **Sideways**: when the person looks at the side.
- **Other.**

Mouth: this group of features is intended to describe the position of the mouth related to facial displays other than “articulatory gestures”. This means that we annotate whether a person has their mouth open (or is opening their mouth), for example because they are surprised, but we do not annotate when the mouth is open because the person is uttering an open vowel. In other words, all of these features are mostly relevant to an annotation of the listener’s rather than the speaker’s mouth displays. Mouth expressions are labelled in terms of **openness** as open mouth vs. closed mouth and in terms of lips shape, where shape includes position of the mouth corners and lip rounding or protruded lips**.** The labels used are:

- **Open mouth:** when the mouth is open or opens as in the case of surprise. Note that there is no value for “closed mouth” as this seems the normal position if one is not speaking. The values “retracted” or “protruded” can be used if the mouth is closed in a “special” way.
- **Corners up:** when smiling**.**
- **Corners down:** in a scowl, sulk or sad expression**.**
- **Protruded:** when the lips are rounded and protruded.
- **Retracted:** when the lips are sucked-in, retracted in the mouth.

Head movements are coded as follow:

- **Single Nod**: a single head movement down-up.
- **Repeated Nods**: multiple head movements down-up.
- **Single Jerk**: a single quick head movement up-down.
- **Repeated Jerks**: multiple head movements up-down
- **Single Slow Backwards Up:** a single slow head movement backwards. (This movement differentiates from single jerk on the basis of the velocity. The term jerk implies quickness, while a single slow backward up refers to a slow movement.)
- **Move Forward**: is a movement of the head forward, this can either be a movement of the head only or can be a movement of the whole trunk. This movement occurs often as a turn elicit signal.
- **Move Backward**: is a movement of the head backward, this can either be a movement of the head only or can be movement of the whole trunk. This movement occurs often as a turn accepting signal.
- **Single Tilt (Sideways):** a single movement of the head leaning on one side.

- **Repeated Tilts (Sideways):** a multiple movement of the head leaning from side to side.
- **Side-turn**: is a rotation of the head towards one side.
- **Shake (repeated**): is a repeated rotation of the head from one side to the other.
- **Waggle**: is a movement of the head back and forth, side to side, it is like a mixture of shake and move backward or forward it is usually produced to show uncertainty, doubtfulness.
- **Other**: either a different type of movement than the three mentioned, or a combination of two or more of them.

3.4 Speech

This version of the coding scheme does not include features for the speech modality. Concerning the expression level, in addition to linguistic expressions of various granularity, filled speech pauses (sounds like *um* or *ehm*) and non speech sounds (like a laugh or a throat sound) should be considered. The last two categories are used in the orthographic transcription guidelines (Section 4.1).

3.5 Multimodal relations

Facial displays and gestures can be synchronized with spoken language at different levels: at the phoneme, word, phrase or long utterance level. In this coding scheme, the smallest speech segment we expect annotators to annotate multimodal relations for is the word. In other words, we do not expect them to take morphemes or phonemes into consideration. 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). They are shown in Table 7. 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.

Attribute	Value	Short tag
Cross-modal function	Non-dependent	Non-dependent
	Dependent-compatible	Compatible
	Dependent-incompatible	Incompatible

Table 7: Relationship between gestures/facial displays and speech

4. Description of required data source type

The coding scheme should be applied to orthographically transcribed video clips.

4.1 Conventions for orthographic transcription

These are a subset of the conventions described in Duncan (2004).

Punctuation

No punctuation is used in the transcriptions.
<…> filled speech pause
For sounds like <um> or <ehm>

%___ non-speech sound
For non-speech sounds like %laugh or %throat

{…} uncertain transcription
If a portion of speech is totally incomprehensible, write {…}; if you don't feel certain about what you hear, enclose the relevant part of the transcription in {}.

5. Coding procedure

The coding procedure described in this section was defined for the Stockholm MUMIN workshop, but it is presented here as a general procedure.

5.1 General task and annotators

At the workshop, three different short video clips, one in Swedish, one in Finnish and one in Danish have been annotated. Annotators were divided into groups of 2-3 people: all those belonging to the same group worked with the same video clip and with the same coding tool. In general, a group of annotators should work with the same video material and the same tool.

The MUMIN annotators were expected to have read this document and to have made themselves acquainted with the relevant literature (see below for a list of suggested references). Furthermore, they were given a tutorial on how to annotate by means of the three coding tools used in the workshop. These were ANVIL (Kipp 2001 and Kipp 2004), MultiTool (Gunnarsson 2002) and NITE (Bernsen et al 2002). Again, in general annotators are expected to be familiar with the coding tool they select for the task, since this manual does not provide any guidance for either coding tool choice or use.

5.2 Work distribution and organisation

The following steps were used in the annotation workshop held in Stockholm and are in general recommended.

First session
Annotators start by annotating a short sequence together in each group to assess their common understanding of the task. Each group works with one of the tools available. The result is saved in a temporary coding file.

Second session
Then each annotator continues coding the same video clip individually by means of the tool chosen by the group. The result is saved in a second temporary file.

Third session
The annotators in each group get together and compare their annotations. Problems are noted. Adjustments to the codings are made to reduce differences, and results are saved in a third coding file.

It may happen that after the first three sessions, changes to the coding scheme need to be made in order to ensure better inter-coder agreement. In such a case, session 3 will have to be repeated for the coders in a group to converge on the updated coding scheme. No modifications were done to the coding scheme at the workshop, although a number of suggestions were given on how to improve the scheme. These modification suggestions have been taken into account in the version described in this document.

Fourth session
If the group does not reach total agreement, the reliability of the competing codings should be calculated, for instance in terms of precision, recall and kappa score.

5.3 Coding passes

The following passes are recommended for an annotation session, and were followed at the workshop:

1. Watch entire video clip.
2. Correct transcribed speech if necessary.
3. Organise speech in short utterances and insert time stamps around the utterances if the tool does not do it for you. Intuitively, a short utterance corresponds more or less to a clause.

4. Identify gesture and facial displays related to the functions under observation.
5. Label facial displays and gestures with tags from the two levels provided.
6. Label the relationship between the facial display/gesture and the corresponding utterance; if necessary to express a correspondence between a gesture or facial display and a speech segment, break the utterances defined in (3) into shorter phrases.

Since understanding of phenomena and annotation tags usually changes as the coding proceeds, these passes should be gone through several times to ensure internal consistency.

6. Tag set declaration

A summary of the tags described in the preceding sections is shown in Appendix 1. We have used the term *dimension* to indicate the modality: in a coding scheme, a dimension will typically correspond to a track. Within each dimension, we then distinguish between attributes and specific values for each attribute. In a specific implementation, it may be desirable not to code at the most specific level: for instance if emotions are not in focus, the annotator may be interested in just coding that there is some emotional colouring attached to a face display without having to specify which one.

7. Annotated multimodal resources

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:

1. 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).

2. One minute clip from the Swedish movie “Fucking Åmal”, consisting of an emotional dialog between father and daughter.

3. One minute clip from an interview of the actress Ann Eleanora Jørgensen by Per Juul Carlsen from the Danish DR-TV (Danmarks Radio)

Since all of the videos are protected by copyright, they cannot be made publicly available, but can be inspected by contacting the authors of this manual.

References

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

Allwood J. (2002) Bodily Communication Dimensions of Expression and Content. In B. Granström and D. House (eds.) *Multimodality in Language and Speech Systems*. Kluwer.

Allwood J. and 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., Grönqvist L., Ahlsén E., Gunnarsson M. (2003) Annotations and Tools for an Activity Based Spoken Language Corpus. In van Kuppevelt J., Smith R. (ed.) *Current and New Directions in Discourse and Dialogue,* Kluwer Academic Publishers.

Bavelas J. B., Chovil, N. and Roe, L. (1995) Gestures Specialized for Dialogue. *Personality and Social Psychology Bulletin*. Vol. 21, No. 4, 394–405, April.

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 Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces,* Lisboa 25 May 2004

Cassel J., Sullivan J., Prevost S., Churchill E. (2000) *Embodied conversational agents*. The MIT Press.

Cassell J. (to appear) Embodied Conversation: Integrating Face and Gesture into Automatic Spoken Dialogue Systems. In S. Luperfoy (ed) *Spoken Dialogue Systems*. Cambridge, MA: MIT Press.

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

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

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

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

Duncan S., Jr. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2), 283-292.

Ekman P. (1999) Basic emotions. In T. Dagleish E.M. Power (eds) *Handbook of Cognition and Emotion.* NY J. Wiley.

Gunnarsson Magnus (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. *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, pp. 1367–1370. Aalborg.

Kipp M. (2004) Gesture Generation by Imitation – From Human Behaviour to Computer Character Animation. PhD Thesis, University of Saarland, under publication.

Knapp M. and Hall J. (2002) *Nonverbal Communication in Human Interaction*, Wadsworth.

MacNeill D. (1992) *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago Press, Chicago.

Paggio P., Jokinen K., Jönsson A. (eds.) (2003). Proceedings of the 1st Nordic Multimodal Symposium on Multimodal Interfaces, Copenhagen, September.

Poggi I. (2001) Toward the Lexicon and Alphabet of Gesture, Gaze and Talk. Available from http://www.semioticon.com/virtuals/multimodality.htm.

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

Serenari M., Dybkjær L., Heid U., Kipp M., Reithinger N. (2992) NITE Deliverable D2.1. Survey of Existing Gesture, Facial Expression, and Cross-modality Coding Schemes. Available from http://www.nis.sdu.dk/projects/pastProjects.php (last accessed 26/4/2004).

Appendix 1. Coding scheme tag set

Dimension	Attribute	Value	Short tag
Facial displays	General face	Smile Laughter Scowl Other	Smile Laugh Scowl Other
	Eyebrows	Frowning Raising Other	Frown Raise Other
	Eyes	Exaggerated Opening Closing-both Closing-one Closing-repeated Other	X-Open Close-BE Close-E Close-R Other
	Gaze	Towards interlocutor Up Down Sideways Other	Interlocutor Up Down Side Other
	Mouth-Openness	Open mouth Closed mouth	Open-M Close-M
	Mouth-Lips	Corners up Corners down Protruded Retracted	Up-C Down-C Protruded Retracted
	Head	Single Nod (Down) Repeated Nods (Down) Single Jerk (Backwards Up) Repeated Jerks (Backwards Up) Single Slow Backwards Up Move Forward Move Backward Single Tilt (Sideways) Repeated Tilts (Sideways) Side-turn Shake (repeated) Waggle Other	Down Down-R BackUp BackUp-R BackUp-Slow Forward Back Side-Tilt Side-Tilt-R Side-Turn Side-Turn-R Waggle Other

	Semiotic type	Indexical Deictic Indexical Non-deictic Iconic Symbolic	Index-Deictic Index-Non-deictic Iconic Symbolic
	Feedback give (F-Give) basic	Contact/continuation Perception Understanding Contact/continuation Perception	CPU CP
	Feedback give (F-Give) acceptance	Accept Non-accept	
	Feedback give (F-Give) emotion/ attitude	Happy Sad Surprised Disgusted Angry Frightened Certain Uncertain Interested Uninterested Disappointed Satisfied Other	
	Feedback elicit (F-Elicit) basic	E-Contact/continuation Perception Understanding E-Contact/continuation Perception	E-CPU E-CP
	Feedback elicit (F-Elicit) acceptance	E-Accept E-Non-accept	
	Feedback elicit (F-Elicit) emotion/ attitude	Happy Sad Surprised Disgusted Angry Frightened etc.	

	Turn-gain	Turn-take Turn-accept	Turn-T Turn-A
	Turn-end	Turn-yield Turn-elicit	Turn-Y Turn-E
	Turn-hold	Turn-complete	Turn-C
	Sequencing	Opening sequence Continue sequence Closing sequence	S-Open S-Continue S-Close
	Multimodal relation	Non-dependent Dependent-compatible Dependent-incompatible	Non-dependent Compatible Incompatible
Hand gestures	Handedness	Both hands Single hand	Both-H Single-H
	Trajectory	Up Down Sideways Complex Other	
	Semiotic type	Indexical Deictic Indexical Non-deictic Iconic Symbolic	Index-deictic Index-Non-deictic Iconic Symbolic

	Feedback give (F-Give) basic	Contact/continuation Perception Understanding	CPU
		Contact/continuation Perception	CP
	Feedback give (F-Give) acceptance	Accept Non-accept	
	Feedback give (F-Give) emotion/attitude	Happy Sad Surprised Disgusted Angry Frightened Certain Uncertain Interested Uninterested Disappointed Satisfied Other	
	Feedback elicit (F-Elicit) basic	E-Contact/continuation Perception Understanding	E-CPU
		E-Contact/continuation Perception	E-CP
	Feedback elicit (F-Elicit) acceptance	E-Accept E-Non-accept	
	Feedback elicit (F-Elicit) emotion/attitude	Happy Sad Surprised Disgusted Angry Frightened etc.	
	Turn-gain	Turn-take Turn-accept	Turn-T Turn-A
	Turn-end	Turn-yield Turn-elicit	Turn-Y Turn-E
	Turn-hold	Turn-complete	Turn-C
	Sequencing	Opening sequence Continue sequence Closing sequence	S-Open S-Continue S-Close
	Multimodal relation	Non-dependent Dependent-compatible Dependent-incompatible	Non-dependent Compatible Incompatible