

## **Determining what people feel and think when interacting with humans and machines**

### **Notes on corpus collection and annotation**

*Dirk Heylen, Anton Nijholt, Dennis Reidsma*

Human Media Interaction

University of Twente, The Netherlands

{heylen,anijholt,dennisr}@ewi.utwente.nl

### **Introduction**

Any interactive software program must interpret the users' actions and come up with an appropriate response that is intelligible and meaningful to the user. In most situations, the options of the user are determined by the software and hardware and the actions that can be carried out are unambiguous. The machine knows what it should do when the user carries out an action. In most cases, the user knows what he has to do by relying on conventions which he may have learned by having had a look at the instruction manual, having them seen performed by somebody else, or which he learned by modifying a previously learned convention. Some, or most, of the times he just finds out by trial and error. In user-friendly interfaces, the user knows, without having to read extensive manuals, what is expected from him and how he can get the machine to do what he wants. An intelligent interface is so-called, because it does not assume the same kind of programming of the user by the machine, but the machine itself can figure out what the user wants and how he wants it without the user having to take all the trouble of telling it to the machine in the way the machine dictates but being able to do it in his own words. Or perhaps by not using any words at all, as the machine is able to read off the intentions of the user by observing his actions and expressions. Ideally, the machine should be able to determine what the user wants, what he expects, what he hopes will happen, and how he feels.

The Human Media Interaction group is carrying out several studies that involve determining in one way or another what the mental state is of a person by analyzing various aspects of his behaviour. In developing intelligent tutoring systems, for instance, we like to know as much as possible of the mental state of the students as is useful to respond appropriately to their actions. A tutoring system should create the right conditions for the students to learn something. This means adjusting the teaching strategy or the specific instructions and feedback to the student establishing the right frame of mind for optimal learning. This not only involves determining whether the students are involved and attending, but also whether or not they are motivated and how they are responding to the mistakes they are making. Do students feel the right kind of frustration? Are they bored or challenged to do better? Are they understanding the instructions? Are things too easy for them or too difficult etcetera? These are some of the things that a skilled human instructor can deduce without effort from the students' behaviours. The question is: how can we build a machine with similar skills?

Also for other applications, it is extremely valuable if one could find out more about the mental state of a person. Besides intelligent tutoring systems, we are also interested in automatically interpreting the behaviours of people engaged in human to human conversation. In the research carried out in the context of the AMI project (<http://www.amiproject.org/>), we are looking at people involved in meetings. Observing people and interpreting their actions to find out their intentions and motivations, what people feel and think, is needed to develop intelligently searchable multimedia recordings of the meetings. Speech recognizers can help us to produce transcripts of the meeting automatically and natural language processing, information extraction and information retrieval techniques can help us to enrich the data with tags that can be used as metadata enabling semantic access. But what we need is not just access to what was said. We also have to find out how it was intended. Only to a limited extent do the actual words that were used express what was meant. Also, we might not just be interested in what the speaker had to say, but also what the listeners thought and felt about it. Was a proposal greeted with great enthusiasm or with a fair amount of scepticism? Who agreed immediately and who didn't?

For both the tutoring and the meeting case, we can deduce a lot about the intentions of students and participants in the meeting by analysing the actions they performed in the task they were given or the things they said in the meeting. However, much of the information about the mental state is not presented in these actions or the words that are spoken, but rather in the way the actions are carried out and the other behaviours that accompany them. In particular, the paraverbal and nonverbal signals may give us an indication of extra dimensions of the mental state of the users of a system or the participants interacting in a meeting or other form of conversation. Recognizing and interpreting such cues is an important research area. In order to understand the way the cues work in the actual cases that we are considering, we need to look at naturalistic data to find out what behaviours are being displayed and how they are associated with diverse functions. In the paragraphs that follow we will discuss several aspects of this undertaking, focussing on the collection and analysis of data. Both for the tutoring case

and the meeting case we have collected material to gain insight in the questions of what mental states are relevant to investigate in the context of the application and the kinds of behaviours that can provide cues with respect to these mental states. We will furthermore discuss in more detail our approach to the data collection and annotation procedure and the philosophy behind this. An important point to consider in this process is that cues and signals can work in different ways and give information about various elements of the mental state. Before we discuss the work on the collection and analysis of the data, we make a brief detour and say something about the semiotic modes of operation of nonverbal and paraverbal cues.

### **Expression Interpretation**

Understanding human communicative action beyond the verbal part of natural language, shares with computational linguistics the general goal of trying to recover meaning from some system of signs. However, the balance in the kinds of “meanings” expressed through the nonverbal or paraverbal channels is different from the balance in the verbal part. Although it is possible to express complete propositional content in gesture, for instance, it is not the main function of most gestures. On the other hand, many metacommunicative or interaction control functions are expressed in the nonverbal channel. One can think about head nods and other backchannels by listeners, for instance, as a clear point in case. In fact, the complete repertoire that listeners use, constitutes a constant source of semiotic signalling in various nonverbal and paralinguistic channels without having, by nature, a verbal counterpart.

Not only the balance in the kinds of expressed meanings is different, but also *the ways in which* the various channels “mean”: their semiotic mode. The symbolic mode is typical for natural language that uses arbitrary conventionalized symbols to denote objects, events, attitudes and whatever. Arbitrary in this case means that there is no natural connection between the sign and the signified. Iconicity is not found a lot in verbal language (though certain symbols may have an iconic origin – which does not mean that they are still used in an iconic sense) but with gestures that are used as illustrators, the coding is mainly iconic.

At first sight, then it may seem that much of what goes on in the nonverbal and the paraverbal channels is not particularly intended for communication in the first place but provides meaning as natural signs in Gricean terms, as symptoms, one could say. For instance, a trembling voice may lead one to assume that the speaker is nervous. One should not assume though, that all nonverbal and paraverbal communication follows this symptom mode of communication. It is, however, a popular point of view with researchers in computer vision that investigate facial expressions. When computer vision researchers analysing facial expressions write about “emotion recognition” they mostly mean recognising one of 6 facial displays. These are the well-known universal expressions of the basic emotions, according to Ekman (Ekman and Friesen, 1974). This, of course, can hardly be called emotion recognition. We know very well that (a) people experiencing a certain emotion may not express it on their face (b) people experiencing a certain emotion may express it on their face in another form than the expression that is studied by the computer visionaries and (c) people displaying an “emotional expression” may not experience the associated emotion. People may use an “emotional expression” not to “express their emotional state” but to “denote” or “portray” an emotional or mental state in some other way. In semiotic terms: the facial display is seldom used as a symptom or natural sign, particularly not in face-to-face conversations but rather as a symbolic event. As Bavelas and Chovil (1997) write: “We propose that the facial displays of conversants are *active, symbolic* components of *integrated* messages (including words, intonations, and gestures).”

So if the facial displays do not express “emotions felt”, what do they express and in what way is that helpful to our applications? This is one of the important questions to consider. The question we should address when interpreting facial displays is the way they function within the general conversational process. That is why we, in our analysis of the corpora that we are considering, are looking at displays in conversations and their communicative content as well. The same holds for human-computer conversations<sup>1</sup>: the facial expressions (emotional displays or others) may not be directly related to emotion felt, but may be communicative in one way or another just as they are in human-human communication.

When analysing and interpreting nonverbal expressions the mode of semiosis should be taken into account. It means, for instance, that it is not appropriate to model the “paralinguistic” resources for processing such as lexicons and grammars, on their linguistic counterpart. The difference between a facial expression dictionary and a dictionary of linguistic expressions lies not only in the elements (the difference between facial displays and words or the fact that facial displays express mostly different

---

<sup>1</sup> Note, that this would be a corollary of the Media Equation (Reeves & Nass, 1996).

kinds of things), but also in the way the expression and the expressed relate to one another. Given the multiplicity of behaviours that can mean different things in different ways in different dimensions (informational, emotional, cognitive), it is important for the corpus analysis and annotation to label the behaviours and the functions separately. Furthermore, it is important to consider the correlation between behaviours and several functional levels. In the following paragraphs, we describe some of our work on the collection and annotation of data of nonverbal behaviours and their cognitive and emotional meanings.

### **Corpus collection and Annotation**

For the tutoring case and the meeting case we are collecting and analysing a corpus with the aim to find cues that can tell us more about the mental states of the users or participants. For this we need to annotate the corpus both on a functional and a behavioural level. For the tutoring situation we looked at video recordings and collected clips of various expressive faces. We tried to determine what gave rise to the expressions in terms of the possible eliciting conditions or “stimulus evaluation checks” as described by Scherer (1987). In this way we tried to find out whether we could link features of the situation with features of the mental state. For the AMI data we are developing an annotation scheme for the cognitive and emotional state of the participants. This should allow us at a later stage to associate behaviours with mental states. However, keeping in mind that behaviours serve multiple functions in a conversation, we have also started to look at other correlations in the corpus. For the AMI data we have looked at the kinds of facial expressions that are associated with certain dialogue acts. Some of the findings will be discussed below.

#### *Tutoring*

For the development of intelligent tutoring programs, we want the system to understand the progress the student is making, the difficulties he is experiencing, his motivational levels, his frustration, etcetera. With this information, the tutoring system, can adjust the teaching strategy, the kind of feedback it provides, and select more or less difficult exercises. Part of such choices or adaptations can be made by building up user models based on the task-oriented interactions the student has with the program, the timing of the responses (how much time does it take the student to answer - or some other measure like this), the correctness of the answers and the history of previous interactions. Such measures were used in a version of the INES tutoring system we have built (Heylen et al. 2004).

For a possible second version, we ran a pilot to investigate the kinds of facial expressions that students show when they are engaged with the system and see whether these could tell us something more about the mental state of the student that is relevant for the tutoring program. The expressions on the face, the direction of gaze and the movements by the head are commonly assumed to provide information about what goes on in the mind of a person.

In our pilot experiment, we had several students interacting with the INES system while we videotaped them. The task consisted of a number of steps. First, the students had to ask a patient (represented by a virtual, 3d character) to roll up her sleeve and put her arm on the armrest using natural language commands. Next, they had to disinfect the region where they were going to inject and finally they had to give the injection. Both the disinfection and the injection were performed by manipulating a haptic device that gave appropriate force feedback. A virtual tutor would give instructions to the students. We collected about 1 hour of video data. From this we selected clips in which a particular facial expression was shown. One of the goals of our experiment was to find out whether there would be facial expressions that could be relevant for an intelligent tutoring system to detect and to make use of in its strategy to optimize the teaching process.

Patient asks for clarification	raised brows (2), frown (1), head pulled back (1)
Patient repeatedly does not understand what student says	smile (2)
Patient pulls up her sleeve	smile (4), raised brows (2), nod (1), pull head back (1)
Student is disinfecting the arm	smile (7), raised brow (1)

**Table 1 Situations and Expressions**

The main goal of the experiment, however, was to explore the use of a specific description method. Instead of labelling the mental state with categorical labels of emotions or mental states, we labeled them with values for the different stimulus evaluation checks (SECs) that Scherer (1987) postulates to make up the appraisal proces. In this way, we make conjectures about the appraisal proces, linking the presumed mental state to features of the situation. Scherer also made some suggestions as to which stimulus evaluation checks are associated with which facial expressions. If we could associate

situations with values for SECs and establish in more detail the association between facial expressions and the SECs, then we could try to figure out what the mental state of a student is (in terms of the values of the SECs) given a particular facial expression and a particular situation.

Novelty	High	Low
Sudden	1 + 2 + 5 + 26/27 (brow and eyelid raised, mouth open)	
Familiarity		4b + 7
Predictability		4b + 7

**Table 2 Association of Action Units with SECs**

Of course, one should take into account the fact that appraisals are individual and subjective and the correspondence between facial expressions and appraisals is also not one-to-one and deterministic. The inference process should take this into account and be at least probabilistic. However, given these caveats, we concluded after the experiment that it would be feasible to link the situations with the SECs. Also, the facial expressions, we found, overall corresponded to what one would expect from the suggested associations in the literature. However, students also display several facial expressions that have a communicative function rather than an expressive function. Also, for both associations (SEC/situation and SEC/expression), more data will be needed to establish a working model (Heylen et al. 2005).

**Table 3 Expressions and Checks in Corpus**

*Emotion annotation in AMI*

Trying to build software that can interpret human actions is not just a goal that serves the needs for intelligent human computer interaction. Understanding what people feel and think when they are engaged in conversation can help us to retrieve information from recordings of people engaged in conversation. Looking at human human interaction can help us to build models for software like the tutoring agent or other dialogue systems (such as embodied conversational agents).

In the AMI project more than 100 hours of meeting recordings are being collected and annotated. The goals are manifold. The prime goal is to develop several kinds of “meeting technologies”. The prime technology around which much of the research is centered in the first phase of the project is called the “Meeting Browser”. The Meeting Browser is a collection of programs that allows people to have access to the recordings that were made. It involves special techniques in multimedia indexing, multimedia retrieval and multimedia extraction. The major effort at this stage of the project is to annotate the recorded meetings manually with all kinds of information that can be used as meta-data for the recordings or, more importantly, for use by machine learning techniques that will automatically extract features from the data or that will generate annotations automatically. The recordings are of interest not just to the signal processing researchers, or the researchers dealing with multimedia information retrieval/extraction but also for people interested in face-to-face conversation for its own sake: conversational analysts, linguistics, social psychologists etcetera.

Annotations for level of interest, speech acts, individual acts of participants, focus of attention, gestures and of course manual transcripts are currently being produced. Another level of annotation that was deemed to be important was “emotion”. The development of an annotation scheme for emotion involves many issues. The most important of these are probably the following questions: What types of emotions do actually occur in the AMI recordings? How can we annotate effectively with the smallest effort? How reliably can these emotions be annotated? In a number of trials we have been experimenting with emotion schemes and procedures to develop a suitable emotion scheme.

It was expected that the AMI meetings would not contain many highly emotional episodes. To establish the kinds of emotions or affective dimensions one might expect in meetings we<sup>2</sup> asked people (33 participants) to select 20 ‘emotion’ terms that they thought would be frequently perceived in a meeting. The participants in this survey were presented with a list of about 200 emotion labels compiled from various sources. Participants could also suggest other emotions. The top 20 list (words that were mentioned most) were: bored (mentioned 23 times), confident, interested, attentive, serious, joking, friendly, curious, cheerful, at-ease, amused, relaxed, nervous, frustrated, decisive, uninterested, impatient, confused, agreeable, annoyed (mentioned by 9 participants). This list already shows that the terms that are being mentioned are not all emotions, strictly speaking. Is *curiosity* an emotion? What to

<sup>2</sup> This survey was carried out by Vincent Wan from Sheffield in collaboration with the authors and other AMI researchers.

think of *respectful*, *dominant* or *tired*, which are other terms in the top 80 list. Of course this depends on one's definition of emotion.

Besides labels for affective dimensions, it appeared from the reactions of the participants to the survey as well as from looking at the first recordings of AMI data, that other phenomena play an important role as well. Clearly, from the point of view of the relevance for meeting browsing and other techniques for building up memories of what happened in a meeting, it is obvious that what is relevant about what goes on in people's minds is not only what they "felt" about what was being said in the emotional meaning of the word, but also whether they were surprised by the things that were said, certain, skeptical or how clear or confusing certain issues were presented. This became apparent from our first trial as well.

In the first trial, people were asked to annotate 20 minutes of meeting data (involving one participant in a meeting) using the procedure defined by Cowie et al. (2000) using the FeelTrace tools from Belfast. This involves a continuous labeling of the emotional content on a plane involving two dimensions: arousal and valence. The agreement between annotators was very low, which is probably due to several reasons. First, the annotators may not have been trained adequately to proceed with the annotation of AMI data. Second, the annotators were asked to annotate the data real-time on a single pass through the data. This inevitably leads to delays, false starts, etcetera. Third, most of the observable changes in the mental states of the participants in the meeting do not directly relate to emotional dimensions. Annotators are noticing these changes and try to accommodate these some way or another in the annotation task. On the basis of these results we decided to make various parts of the procedure more discrete and to introduce a labelling framework that is more relevant to the mental state changes that occur in the meeting data (beyond emotions). The main finding of the trials, so far, has been that most of the states that one can identify in the meeting are fairly unemotional. People mostly show meta-cognitive or other mental states relating to thinking, believing, understanding, and attention.

Given the small amount of data we have annotated so far, we have not made an extensive analysis of how the behaviours team up with the mental states. From interviews with the annotators, however, it became clear that there are important differences between annotators with respect to the signals they pay the most attention to. Some pay more attention to the facial expressions and the gestures and other to what is being said. This has effects on the annotation, both on the segmentation and the labels.

So, in the process of annotation the first few meetings with "emotional labels" we noticed that most of the labels we use relate to meta-cognitive functions. As we remarked above, many of the nonverbal expressions, even expressions typically associated with emotions according to the literature (such as the six universal facial expressions associated with the six basic emotions by Ekman; see Ekman & Friesen, 1974, for instance) that we use are not directly expressing an emotion. We often found them in other contexts as well. Though we have not yet made a systematic analysis of the correlations between the behaviours correlated with the mental state annotations (as the annotations are still in a trial stage), we have started to look at the relations between facial expressions and other behaviours and the communicative actions participants take.

#### *Facial Expression and Conversational Functions*

We noted before that the face serves as a complex expressive medium that signals meanings in all kinds of ways. It is important, for automatic processing of the facial displays and their dynamics, to realise that there are multiple modes of semiotic signalling at work. Facial expressions in conversations are not a simple read out of our mental state. Conversations are interactive processes in which our actions are directed at persons for communicative purposes. Most of our expressions are therefore consciously produced to inform the other persons about things that we really want the other person to know. We do not simply "reveal" our mental state, but we choose expressions in the same way that we use natural language expressions. Most facial expressions could be said to functions as a kind of "speech act".

For her master's thesis and AMI traineeship in Twente, Petukhova (2005) looked at a number of meetings, that she had annotated with a refined dialogue act scheme proposed by Bunt (2000). Next, she collected a database of clips from the meeting which showed the kinds of nonverbal behaviours displayed with each of the functions. Table 4 shows the major behaviours that are associated with several kinds of communicative functions. When looking at the facial displays, one can see that typical expressions for emotions and mental states, like "surprise", "puzzled", "guilty" can be associated with specific communicative acts. These expressions could be said to make up the communicative act.

<b>Communicative act</b>	<b>Face</b>	<b>Posture</b>	<b>Hand/Arm gestures</b>	<b>Head movements</b>
Pause	neutral	Neutral	Gesturing stops	

Stalling	Thinking face/ uncertain	Turn to addressee	Iconic gestures, rotation movements, self touching	Waggle
Error signalling	Guilty face	Neutral	Gesticulation stops	Lowering head
Retraction	Neutral	Neutral	Gesturing stops. Retract to neutral.	Neutral
Completion elicitation	Uncertainty		Iconic gestures, rotation movements	
Self-correction		Neutral	Hand gestures stop. Retract to neutral.	
Completion	Neutral	Lean forward	Hand gestures start	
Correct misspeaking	Surprise Puzzled	Lean forward	Raise hand/finger Gesturing starts	Frequent head shakes

**Table 4 Nonverbal behaviours as Speech Acts**

By analyzing this data into more depth and combining it with the information provided by the emotional coding, we hope to gain more insight in the semiotics of facial expression, i.e. the way they operate in a conversation. Analysing the expressions in connection with the communicative acts gives us information about the way “emotional” expressions are used to communicate the “affective” impact of what is being said: agreement, acceptance, surprise, etcetera.

### Conclusion

We have illustrated several inroads that we have made with regard to the questions “How can we find out what people are thinking and feeling?” and “How do people express this?” We believe that it is important to realize that many nonverbal signals (and also verbal ones) express meaning in various ways and on various levels.

Once we have the annotation scheme working and have annotations for the meetings we can go back and look at the various behaviours that occur with the various classes of functional labels. We can also correlate the emotional state with the dialogue acts and other levels of annotation. But this is still future work. By a combination of these approaches we hope to gain more insight in the ways facial expressions (and other forms of bodily communication) contribute to the overall messages emitted and use this both for analysis (deriving inferences about mental state automatically) and generation (for our tutoring agents and other embodied conversational agents that we are building). A combination of such procedures is necessary to book reliable and sound conclusions about the relation between nonverbal behaviours and “mental state”.

When one starts collecting and analysing naturalistic data, it becomes immediately clear that one cannot rely on the standard associations between behaviours and functions that the existing research has focussed on (for instance the relation between facial expressions and emotions). Careful collection and analysis of data should enable us to construct a more accurate picture of the associations between expressions and meanings.

### References

- Bavelas, J. B., & Chovil, N. (1997). Faces in dialogue. In J. Russell & J.-M. Fernandez-Dols (Eds.), *The psychology of facial expression* (pp. 334-346). Cambridge, U.K.: Cambridge University Press.
- Bunt, H. (2000). Dialogue pragmatics and context specification. In *Abduction, Belief and Context in Dialogue: studies in computational pragmatics*, H. Bunt and W. Black (eds.), pages 81–105. John Benjamins, Amsterdam.
- Cowie, R. E. Douglas-Cowie, S. Savvidou, E. MacMahon, M. Sawey, M. Schröder (2000) ‘FeelTrace’: An instrument for recording perceived emotion in real time. ISCA Workshop on Speech and Emotion, p. 19-24.
- Ekman P. & W.V Friesen (1971) Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, Vol 17, p. 124-129.
- Heylen, D. & M. Vissers & R. op den Akker, & A. Nijholt (2004) Affective feedback in a tutoring system for procedural tasks. In E. André, L. Dybkjaer, W. Minker, and P. Heisterkamp, editors, *Affective Dialogue Systems*, Lecture Notes in Computer Science, pages 244–253. Springer-Verlag.
- Heylen, D. & Gijzen, M. & Nijholt, A. & Op den Akker, R. (2005) Facial Signs of Affect during Tutoring Sessions. In: *Affective Computing and Intelligent Interaction*, H. Tao, T. Tan and R. Picard (eds). 24-31
- Petukhova, V.V. (2005) Multimodal interaction of multimodal dialogue acts in meetings. Msc Thesis, Tilburg University.
- Reeves, B. and Nass, C. (1996) *The Media Equation*. CUP.
- Scherer, K. (1987) Toward a dynamic theory of emotion: the component process model of affective states. <http://www.unige.ch/fapse/emotion/publications/list.html>.