Feedback Loops in Communication and Human Computing

Rieks op den Akker and Dirk Heylen

Human Media Interaction Group, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands,infrieks@ewi.utwente.nl

Abstract. Building systems that are able to analyse communicative behaviours or take part in conversations requires a sound methodology in which the complex organisation of conversations is understood and tested on real-life samples. The data-driven approaches to human computing not only have a value for the engineering of systems, but can also provide feedback to the study of conversations between humans and between human and machines.

1 Introduction

An important aim for research that puts itself under headings such as Affective Computing, Ambient Intelligence, or Human-(Centered)-Computing, is to build systems that are able to interact with humans based on capabilities that are similar to those humans use to interact with each other. Being able to interpret human behaviour and determine the rational and affective concerns, motives and goals that lie behind it is a central capability of humans that are naturally inclined to take an intentional stance and have developed complex signalling systems to communicate their beliefs, intentions and to show or hide their attitudes and emotions.

With these central concerns, the research in such fields as Affective Computing can be seen to fit into the tradition of Artificial Intelligence and has in several respects complementary goals of research as Natural Language Processing. Whereas, NLP traditionally restricts its scope to language and the interpretation of utterances in semantic and pragmatic terms, Affective Computing focusses mainly but not exclusively on inferring information on the affective and the mental state of a person from physiological signals or forms of nonverbal communication.

For the creation of ambient intelligent systems a combination of approaches is needed that goes beyond the individual disciplines. Extending the work in NLP to include other modes of communication and other functional variables such as affect and attitude will also require rethinking methodology, theories and models. Similarly, the work in affective computing as it is currently practiced will need to re-orient itself if it is to be succesfully incorporated in a broader initiative that considers the full complexity of human communication. It is important for the success of visions such as ambient intelligence or affective computing to

overcome the many restrictions that are self-imposed by disciplines or simplifications that are assumed due to the divide and conquer strategies that are common in scientific and engineering practice. One of the important methodological shortcomings to date is the reliance on non-naturalistic data that is studied out of context in many disciplines. In particular, affective computing suffers from this restriction as the summary of the state of the art and research programme as presented in [Pantic et al., 2007] shows. Moreover the theoretical background of the programme is based primarily on psychological studies that themselves study (affective) behaviours outside natural contexts of occurrence. In this paper, we sketch a complementary view that studies human interactions as they occur naturally as the basis for computational modelling and ambient applications.

Ultimately an ideal system would know how to deal with 1) the full gamut of human communicative behaviours, 2) the full range of meanings produced, and 3) the full complexity of human communication. Currently, the focus of the various research disciplines on one or more modalities, fails to address the intricate ways in which communicative behaviours are composed into complexes that are highly context dependent. The notion of communicative behaviour is often restricted to the typical 'expressive behaviours' such as language or non-verbal communication. However, intention abduction also takes place when someone observes another person simply carrying out an action. The focus on either semantic/pragmatic issues or on affective parameters ignores the various relations between them and aspects of the mental state of persons that are expressed through their behaviours such as propositional and interpersonal attitudes. The complexity of human communication shows, amongst others, in the different ways in which behaviours can express meanings that may shift depending on the context. For instance, facial expressions that first appear as symptoms of an emotional experience may shift to iconic expressions in a conversational context.

In several projects we have worked on the boundaries of disciplines or have attempted to move beyond them. In our work on affective dialogue systems, for instance, we explored the option to interpret facial expressions and other behaviours in the context of the ongoing dialogue using the appraisal checks as labels mediating between the contextualised action of the student on the one hand and the emotional appraisal on the other ([Heylen et al., 2005]). In the same context we looked at the generation of the appropriate responses of the dialogue act system taking into account the estimated affective state of the student that determined the general teaching strategy, the choice of dialogue act and the wording of the utterance ([Heylen et al., 2004]). In the AMI and AMIDA projects (Augmented Multiparty Interaction, http://www.amiproject.org)), we take human face-to-face conversations as the domain of study. We will use this project to illustrate our points in this paper.

In this paper we present some of the challenges for building human-inspired interactive systems, pointing out the complexity of the task that is partly determined by the nature and the structure of communicative processes, the many variables that play a role, and the many forms that communication can take.

But most importantly, the complexity of the task arises because of methological restrictions.

In Sections 2 and 3 we will concentrate on what it means to develop human computing systems that are able to converse with humans. Theories of conversation provide a model of conversations and the terms that are used to name the important distinctions, structures and phenomena in conversations. Based on models such as these the parameters that make up an information state in a dialogue system are defined. Similarly, the theory provides us with the labels to use in annotation schemes for the description of actions and events in multimodal corpora. This is what we will focus on next. We discuss, how reliability analysis of the annotations play a central role in the evaluation of human computing systems. Without it a corpus is not of much use. Reliability analysis forms the interface between the quantitative measures of the phenomena and the qualitative and subjective measures of the same phenomena. They form the contract between systems designer and end user of the system. We illustrate these issues by discussing our analysis of feedback in the AMI dialogues in Section 5. More about the AMI corpus and the place of the scenario based meeting recordings in the methodology can be found in [McCowan et al., 2005] and on the website: http://www.amiproject.org.

The aim of this paper is to understand the *possibilities* and *boundaries* of human computing technology. Human computing, we believe, means building interactive systems that have the same capabilities that humans have in communication. We provide an analysis of what it means to "engineer natural interaction". This analyses may inspire new approaches to design such systems, and they may inspire new directions of research in human-human and human-computer interaction.

2 Engineering Natural Interaction: preliminaries

Every student after building his first natural language dialog system knows the disappointment when he demonstrated his system to his teacher and the first sentence that was typed in or spoken did not work. "Oh, yeah, sorry, if you want to say something like that, you have to say ...". Sometimes followed by a promising: "I can easily build that one in." A teacher feeling sympathy for the students attempt will respond with: "Well, you can't have it all. It's the basic idea that counts. How does your system work? How is it build up? And, how did you design it?"

Montague considered English a formal language ([Montague, 1970]) and in natural language processing this kind of view on the equation between natural and formal is the basis for developing theories, algorithms and applications. If we put off our formalist glasses we see that people don't speak grammatical sentences; fragmented speech, restarts, incomplete sentences, hesitations, insertions of seemingly meaningless sounds, are more the rule than the exception. It is important to remember that the formalisations in NLP, inspired heavily in the first years by Chomsky's view on language as a system of rules, are based on an

idealisation as well as a simplification. The early Chomskian view entailed abstracting away from performance to competence and from the actual language user to an "ideal" speaker. It focussed on formal aspects such as syntax and ignored language use ([Chomsky, 1965]).

Computational engineering, by nature, always follows a similar path of abstraction and formalisation. NLP systems rely on dictionaries and grammars, that try to capture the rules and regularities in natural language that make communication possible. It is this "static side" of language that we can build natural language technology on. This can be complemented by *data-driven*, *statistical* approaches that try to bridge the gap between what users do and the functioning of our natural language processing system. But natural language is not something fixed, something that is understandable in the form of a formal system. Communication is build on conventions and regularities but at the same time it is essentially a way to negotiate meaning and to establish new conventions for interaction. Technology therefore will always fall short.

In technology, the conventions are present in two ways. They are fixed, by the engineer (just as dictionaries and grammars fix a language in a particular state) and they are communicated to the user of the system in the form of a user manual which is essentially a *contract* between the user and the technology. The contract specifies the interaction with the user; what it affords, what the semantics of the expressions of the user in communication with the systems is, in terms of the effects the systems brings about in response to this. The contract further specifies how the user should interpret the systems outputs. In short, the contract defines the user interface with the system in the practice of use. When the user asks a question in a certain way, the contract says that the system considers this as a question and that it will respond to it in a certain way. In systems that aspire to natural interaction the forms of interactions are made to look as much as possible like the forms of interactions the user already knows from the experience of interacting with natural systems.

We will repeatedly use the term contract to identify a set of interactions between the phenomena as they 'naturally' occur and the scientific study of them, between the scientist and the engineer that uses the knowledge to construct systems, and between the engineer and the user of the system. If all works out well engineered interactions will become second nature.

3 The nature of communication

In this section we will outline how human-human communication works and what it would mean to build machines that can make sense of what is going on when humans communicate and what it would involve for a machine to participate in a conversation. The phenomena and processes that make up conversations are presented in the form of variables in so-called information states in dialogue systems and enter in the guise of labels used in annotations of recordings of dialogue, as we will see later.

To start the discussion, imagine a group of people having a conversation. Look around you or turn on the television and think about what you are seeing. Participants in face-to-face conversations move their heads, their bodies, their hands, their lips, their eyes. They produce sounds among which there are sounds that one might identify as expressions of some natural language. When we look at the whole collection of behaviours we segment them in all kinds of ways and label them with a variety of categories: a left eye-brow is lowered, a cup is grasped, a person is speaking, another is listening, somebody asks a question, the speaker is embarrassed, people shake hands, they interrupt each other, they threaten each other, they enjoy themselves. Often the same behaviour can be classified in many ways; we can describe the action in physical ways (mouth corner pulls up), or categorize the behaviour using a common word for the action (smile), characterize the function (she relaxes) or the effect (she made him think she likes him).

There are a couple of things we do when we are involved in a conversation. The behaviours that we see or hear displayed by our conversational partners are not only identified and classified in all kinds of ways. We also consider what caused them, what function they served, what intention (if any) was behind them. What we call communicative behaviours are typically those that were intended to be recognized as serving the intention they were intended to serve. One of these intentions is that the action would be recognized as a communicative behaviour but typically there are many others besides this one. A way to group theintentions is given by the list of levels that [Clark, 1996] distinguishes. Being recognized as a communicative behaviour is, one could say, the second one on this list: the level at which a person producing some communicative behaviour intends it to be identified as a signal (a meaningful act intended to be taken as meaningful by the receiver).

1. Joint[A executes behavior t for B to perceive; B attends perceptually to behavior t from A]
2. Joint[A presents signal s to B, B identifies signal s from A]
3. Joint[A signals to B that p, B recognizes that A means that p]
4. Joint[A proposes a joint project to B, B takes up the joint project]

This list of levels at which a communicative behaviour functions shows both the actions performed by the communicator (the speaker, in case of speech) as well as the corresponding action by the addressee (the hearer, or whatever). The notion of communication as a joint action is one of the ways to view an inherent feature of communication: the fact that behaviours, actions and intentions of participants in a conversation are tightly coupled and directed at each other. This fact leads [Schegloff, 1982] to view a conversation as an interactional achievement. This way of putting it stresses another point: that the way a conversation evolves, the way actions of a participant unfold are contingent on the actions and responses of the other participants. Clearly if A executes some behaviour for B to perceive, but notices that B is not attending he will have to take measures in order to ensure the success of the communication. As [Kendon, 1967]

remarks: participants function in two modes (at the same time): an expressing and a monitoring mode.

Monitoring the other participants is an important part in the process of conversation as a person who makes a contribution should check whether his action has succeeded and this is dependent on the changes effected in the state of the other person to which the contribution was directed. [Clark and Schaefer, 1989] consider what it takes to make a contribution to the discourse to consist of two participatory actions by the contributor and the addressee: the contributor presents his utterance (presentation phase) and the addressee provides evidence that he has understood the utterance (acceptance phase). The following main types of providing evidence are distinguished in [Clark and Schaefer, 1989]:

1. Continued attention by the addressee
2. Initiation of the relevant next contribution
3. Acknowledgement
4. Demonstration
5. Display

The first type shows that behaviours that are produced to monitor and to attend can by themselves be indicative and expressive. The most typical example of this is the pattern described by [Goodwin, 1981] who notices in the conversations he is studying that a speaker as he makes a contribution will make sure that the addressee is looking at him and if not will pause till the addressee does so. By initiating a next contribution the addressee shows that he has at least understood the contribution by the previous speaker as an invitation to make some contribution. Of course, this next contribution may make it clear that the addressee did not understand the utterance as intended at all. Acknowledgments consist typically of nods and backchannels such as 'uh huh' and 'yeah'. If an addressee performs or starts to perform the action that the contributor was inviting him to perform, than this is a typical demonstration of understanding. Finally, a display of understanding is taken to be a case where the addressee displays verbatim all or part of A's presentation.

One of the important factors that determines which kind of behavior is displayed as evidence of understanding, is the precise setting and the task that has to be performed. In face-to-face conversations continued attention is often shown through gaze or nonverbal acknowledgments. In contexts as described in [Nakano et al., 2003] and [Kraut et al., 2003] on the other hand, where one person instructs another on physical tasks, the right or wrong execution of the task provides information on understanding.

An important lesson to draw from these and many other examples is that contexts and setting (physical context, task context) determine to a large extent the precise kinds of behaviours that are being performed, whereas the underlying functions that require behaviours of one kind or another remain the same.

The behaviours that recipients of a communicative behaviour engage in as a response to show the acceptance (or failure of acceptance) constitutes a communicative behaviour in its own right and will be responded to by the other

participants. A typical case is where a speaker sees that a listener is paying continued attention to what the speaker is saying to which the speaker responds by continuing to speak.

There are many cases where observers and participants in a conversation fail to interpret signs and signals or whether acceptance and understanding is only partial. The perception of the behaviours displayed may be faulty. The behaviour may not be recognized as bearing a meaning or may be interpreted different from intended. And of course: understanding what someone wants to achieve does not necessarily lead to agreement and acceptance of the joint project that is being proposed. Typically, the acceptance may be partial in other ways as well. For instance, in the case where one is told something with the purpose to change one's beliefs, one may find the message implausible, unconvincing, untrustworthy, or highly plausible but with some scepticism remaining. So besides the fact that acceptance proceeds on different levels it can also proceed to different degrees. It depends on the context and the goals of the interlocutors what degree is acceptable. For instance, a belief may have been communicated by a speaker for the purpose of informing the addresse about the speaker's belief set or it may have been communicated to actually convince the addressee of the truth of this belief.

Having the addressee update his beliefs (as is the case for the class of speech acts called *assertives* by Searle) or showing how one feels (as with the class of acts called *expressives*), are not the only kind of purpose that a communicative act may serve and the kind of participatory behaviours that interlocutors display in response may differ accordingly. Note also that many conversations involve more than two participants. Communicative behaviours may be directed at multiple participants at the same time with the aim to have a different effect on each of them.

Not all the behaviours displayed in a conversation are intentionally produced to communicate. Eye-blinking or breathing are some of the stereotypical behaviours that go on automatically, unconsciously. Most of the time they also go by unnoticed. Nevertheless they may work as natural signs, providing information about the mental and physical state of a participant. Yawning and sneezing are behaviours that are mostly produced without the intention to communicate as such, but they do have an important impact on the conversation.

Another typical case of unintentional information production is the case of leakage which is often discussed in the context of non-verbal communication when particular behaviours that are difficult to control by a person provide information contrary to the intentionally produced utterances. It is foundational for human communication that we are able to distinguish the semiotic status of a behaviour. If we see someone raise his hand we can understand this behavior in three different ways ([Buytendijk, 1964]):

1. as an *expression*, of anger or condemnation, for example
2. as a part of some *action*, the killing of a gnat
3. as a *representative gesture*, a greeting for example.

We therefore have to consider the whole situation and how it develops in time. The distinction between an act and an expression is that an act is a movement that is directed towards some specific endpoint, its goal, whereas an expression is a movement which is an image, in which a meaning becomes visible. As far as a behavior is an expression it shows some inner state of the subject, affect or mood; as far as it is an act, it is a movement that is performed to establish some outside state of affairs. Eye blinking, for example, can be an act (to remove something from the eye) or an expression of nervosity, and it can also be a representative gesture. In the latter case the act has become a sign (a wink) and the movement looses its primary function. In gesturing, such as speech and writing the relation between the movements and their senses, the meanings they intentionally refer to is indirect, whereas in the expressions the inner state is immediately revealed. If someone stamps his feet on the ground this can be an expression of anger as well as a gesture to indicate that one is angry, or both. Machines only act, and even that only in a metaphorical sense; they don't have an inner state that is revealed in expressive behavior. Animals show expressions and can act as well, but they don't make gestures. It should be noted that it is sometimes hard to tell whether a certain behavior is expression or act. Almost all acts and all gestures (such as speaking) have the character of an *expressive* behavior[1] also, through the *way it is performed.*

Being able to identify the physical characteristics of a behaviour is a necessary prerequisite for interpretation but the challenge of interpretation resides in knowledge about the situation in which the behaviour takes place.

The interpretation of (communicative and other) behaviours involves a search for the determinants that caused the behaviour. Intentions and rational goals that are obviously important determinants, but there are many other kinds of determinants that play a role in conversation. Expressing or hiding how one feels, the need for affiliation and contact, social obligations and commitments that need to be taken care of, are some of the concerns that [Goffman, 1976] would classify under the ritual concerns. One way to classify the various motives for communicative behaviors is by the following needs.

1. The need to get something done: business, tasks, goals.
2. The need to communicate and build up rapport.
3. The need to express oneself.
4. The need to make conversations go smoothly.

Besides the task goals, the interpersonal and the expressive side, there is also a concern with the way the communication proceeds as such: metacommunicative goals that may involve concerns with turn-taking, channels of communication etcetera. The system constraints that [Goffman, 1976] lists provide a

[1] Studies of behavioral signals in affective computing research often consider only the expressive function, ignoring the fact that most signals (such as facial expressions) in dialogue constitute discourse-oriented actions, i.e. linguistic elements of a message instead of "spillovers" of emotion processes ([Bavelas and Chovil, 1997]).

good indication of what these meta-communicative actions involve or what is necessary for smooth communication.

1. A two-way capability for transceiving readily interpretable messages
2. Back-channel feedback capabilities "for informing on reception while it is occurring"
3. Contact signals (sigalling the search for an open channels, the availability of channel, the closing of a channel, etcetera)
4. Turnover signals (turn-taking)
5. Preemption signals: "means of inducing a rerun, holding off channel requests, interrupting a talker in progress"
6. Framing capabilities: indicating that a particular utterance is ironic, meant jokingly, "quoted", etcetera.
7. Norms obliging respondents to reply honestly in the manner of Grice's conversational maxim.
8. Constraints regarding nonparticipants which should not eavesdrop or make competing noise.

Besides such system constraints that tell how individuals ought to handle themselves to ensure smooth interaction, an additional set of constraints can be identified "regarding how each individual ought to handle himself with respect to each of the others, so that he not discredit his own tacit claim to good character or the tacit claim of the others that they are persons of social worth whose various forms of territoriality are to be respected." These are ritual contingencies that need to be taken into account and that may also take up a couple of exchanges. Also back-channel expressions may let a speaker know whether or not what he is conveying is taken to be socially acceptable besides signalling understanding. Because conversations are joint actions, the desire of one person to communicate must be matched with the will of the other to participate in the conversation as well. Conversation thus involves a complex structure of negotiating rights and obligations regulated by norms and social conventions.

Modeling conversational action for automatic processing (both analysis and generation) requires not just the modeling of how rational actions are performed through language, how the mechanics of language and conversations work, but also of the personal, interpersonal concerns, emotions and attitudes play a role.

In the next section we look at the way these aspects of conversation are taken into account in current systems that analyse and reproduce natural dialogue.

4 Ways to engineer natural interaction

The main challenge for ambient intelligent systems is to be able to determine what is on the mind of a person, inferring this from the behavior displayed in the particular context, prior knowledge about the person and about the behavior of humans in general. Although conversations are a particular type of action and other interactions between humans and the ambient technology need not be

conceived of as following this model in every detail (for instance, by having embodied conversational agents or humanoid robots all over the place) they are well suited to illustrate the issues of natural interaction as conversations display the full gamut of processes and modes of interaction. The structures and patterns in interaction, the processes as they have been identified above, by presenting a view on conversation and the way it is modelled in theoretical and practical frameworks, provide an overview of all the things to take into account when modeling and implementing human-system interaction in a natural way. As we said before, though, in different contexts the precise forms the contributions take and how they are organised will differ. It is therefore important to use methods that can deal with this contextual dependence. The development of spoken dialogue systems often proceeds by starting with collecting Wizard of Oz data and taking the communicative behaviors that people deploy in these types of interactions as representative for future interactions with the system.

We will now point out two kinds of research areas which formalise the phenomena that make up conversations. One is the practice of building dialogue systems, where the phenomena turn up as variables in an information state and the second is the study of algorithms for automatic analysis of human-human interaction, where the phenomena turn up as labels to describe the data.

Consider again the four levels described by [Clark, 1996]. If we think of a system that is engaged in conversations in a similar way the system should: 1) be able to execute particular behaviors, 2) that count as signals 3) with an intended meaning 4) in an effort to propose a joint project, and 5) be able to perceive behaviors from others, 6) identifying them as signals 7) and recognizing their meanings 8) so as to figure out and take into consideration the project that is being proposed.

Algorithms developed for specific applications may focus on one or more of these aspects. Current dialogue systems offer an example of the way in which the elements and processes that make up a conversation can be conceived of in terms of data structures and algorithms that keep track of the most important variables. In many dialogue systems what goes on in a conversation is captured in an information state that is updated as the conversation proceeds, often with a stack of states capturing the history of the conversation. One of the more complex instances of such a state is presented in [Traum and Rickel, 2002]. A multiple layer approach is taken in this paper towards modelling and managing the complexities involved in multi-party multi-modal interactive systems, "including who is accessible for conversation, paying attention, involved in a conversation, as well as turn-taking, initiative, grounding, and higher level dialogue functions".

The central Information State, a store of information, that is updated by functions, that are the interpretations of the Inputs received by the Interpreters. The Generator module uses to updated Information State to decide for the actions to be performed and generated.

The following layers are distinguished in [Traum and Rickel, 2002]. This list of layers and parameters bears close resemblance to the list of system constraints we have presented in the previous section.

1. Contact layer: whether and how individuals can communicate: the modalities that can be used and the media that can be used. (*make-contact, break-contact*)
2. Attention layer: the focus of attention of each of the participants (*give attention, withdraw attention, request attention, release attention, direct attention*)
3. Conversation layer: model of the various dialogue episodes going on throughout the interactions (there may be several conversations going on in parallel)
   (a) Participants: active speakers, addressees, overheareres, etc.
   (b) Turn: the participant with "the right to communicate" using the primary channel (take-turn, request-turn, release-turn, hold-turn, assign-turn)
   (c) Grounding: how is information added to the common ground
   (d) Initative: the person who is controlling the contributions: take-initiative, hold-initiative, release-initiative
   (e) Topic: start-topic and end-topic
   (f) Rhetorical connections between content units
4. Social commitments
5. Negotiation layer

Contributions to dialogue will typically perform several functions at the same time and will thus be multiply determined. The central research question is whether we can determine how the many behaviors determine the update of the various functions and how we can use this knowledge to analyse human behavior and generate appropriate responses.

The definition of the layers and the variables in the information state of a dialogue system is an instance of the formalisation that turns natural processes into a formal architecture. The same kind of objectivation of theories and assumptions about conversation takes place when corpora of human-human interaction are collected and annotated. The specification of the coding schemes puts down how terms and concepts will be applied to specific instances of real data. We have been working in particular on the AMI data, as mentioned in the introduction. In the case of the AMI corpus, this resulted in the following levels:

1. Named Entitities
2. Dialogue Acts
3. Meeting Actions
4. Emotion and Mental state
5. Topic Segmentation
6. Text Transcription
7. Individual Actions
8. Argumentation Structure
9. Focus of Attention
10. Person Location
11. Various kinds of summaries

There are several important research questions for human computing in these contexts that are typical for data-driven, corpus-based research. First, how can we derive metadata automatically from the raw signals; for instance, automatic transcripts using speech recognition, or descriptions of facial expressions like action units from the video using computer vision techniques). Second, using the hand-made and (semi-)automatically derived metadata to infer further information about what happened in the meetings. We will illustrate this second question below. Third, the corpus and its metadata can be used to derive certain statistics about behaviors as they occur in the corpus that can be used to test certain hypotheses about human behavior or as input for the construction of artificial entities that need to respond to or display similar behavior. Testing the assumptions about conversations derived from theories, does not only happen after the annotation has been performed but also at the development stages of the annotation scheme and the initial tests.

The collection of data and metadata can serve different purposes depending on the context. For instance, in one case information that is present may be used in the algorithm as one of the parameters on which the algorithms bases itself to further classify and label. In other cases, the metadata derived from manual annotations may be used as the ground truth with which automatically derived data of the same kind is compared for evaluation purposes.

An important methodological issue in the collection of corpora and the construction of the annotations is the specification of the labels to use for description and the definition of their use: to what kinds of objects do they apply and how should an annotator decide what counts as what. This issue is addressed mostly by using an iterative approach, where initial drafts of schemes are tested on subsets of the data by several annotators to find out the fit between theory and data and the precision of the specification by measuring the intersubjective agreement. We illustrate these steps in the remainder of this section.

We build three types of models. The first type comprises the *qualitative* models, in which we *describe* what happens in meetings using terms from the various scientific vocabularies to express the important concepts, ideas, the phenomena and processes that we observe in meetings. The second type consists of the *quantitative* models which can be rule-based or statistical. The third type contais the *computational* models, software implementation and their implementations. In each of these models the words and notions always keep referring to the intuitive semantics of the primary concepts that we know from our practical experience with meetings.

What we will point out in the next sections is how the various models are connected. The first model provides data from which to derive the second kind of models. The second kind is used to develop the computational models. Both of the latter kinds, can provide us with insights that make us change our theories and models of the data, which leads to an update of our theories and possibly our data annotations in an incremental way.

5 From data analysis to system integration

In our work on building affective dialogue systems or other applications in which human communication is processed automatically, corpus collection, corpus annotation, and reliability analysis of the annotation procedures play a central role. In this section we present this *methodology* and how the various steps fit in the proces of designing and evaluating a human computing system. We will show how this method is based on a number of feedback loops. To illustrate this we present some details about our studies on *feedback* behavior of listeners in face to face meetings where the analysis of the data leads us to rethink the notions that we started out from.

5.1 The Method

The method that is usually followed basically consists of steps that are motivated by the specific application that one has in mind. This can be to make a system that recognizes facial expressions of affective states of learners in a face-to-face tutoring situation, a system that recognizes certain backchannels and turn-taking behavior in face-to-face conversations, or a system that has to generate rich expressive speech for a virtual story teller. In all of these cases, what we are aiming at is to model natural human behavior, to implement it so that our system behaves as humans would do in similar situations. The steps we take are the following:

1. data-collection
2. data-analysis and modeling
3. model-implementation
4. system evaluation
5. reconsideration

Data-collection can either be done in natural situations, but often happens in controlled situations, similar to experimental physics. A major point of concern here is that the situation is such that we can reasonable expect that the results of our analysis can be transferred to the situation in which we want to apply the model. Ecological validity is very important in this kind of research. Based, in part, on intuition and state of the art theoretical findings, we design annotation schemes in which we define labels for the relevant phenomena of interest in our data.

The annotation procedures are used by human annotators to produce hand-annotations. Other features are computed automatically, such as for example the F0-contour of speech signals, or the movements of facial units, or the words that occur in the realisation of some dialogue act. Hand-annotation by human annotators is essential here since not all features can be automatically identified.

Reliability analysis is then carried out on the annotations to see whether different annotators agree sufficiently in the way they labeled the phenomena. If the measure of agreement is too low we can either throw away that part of the

annotations that show low agreement or we will redesign the annotation proces. Reliability analysis is an essential step because we need a reliable relation between the features that quantitatively describe a phenomenon and the qualitative label that is assigned to it. Only then, as engineers can we offer the users a contract that is the basis for the application of the system we build and the way the user interprets the outcomes or expressions that the system produces.

The models that we build based on the data analysis can either be statistical or analytical. If we use automatic classifiers trained on the annotated data we hope that the classifier performs accurately on unseen data. Note that a high inter-annotator agreement and a high accuracy score of the classifiers is an indication that we have succeeded to model the phenomenon in an accurate way. Evaluation with users, other subjects than annotators, should see whether the outcomes of the machine conforms the way the subjects assess and label the outcomes.

For all practical purposes reliable annotations and a good classification method, provide us with a contract that guarentees the soundness of our design and a manual for the potential users, as long as they use it in a context that satisfies the conditions of our experimental situation.

Still, since we are dealing with data from a limited number of humans annotated by a limited number of annotators our evaluations can only give us statistical outcomes: the best we can offer is saying something like "in 95% of the cases what the facial expressions of the ECA in this type of situation will show is a grin". It is not possible to pin this down to a statement about this unique situation. Moreover, the behavior that is shown by the system will be some statistical mean, representing the "average" behavior of the subjects that happened to act in the data collection proces.

5.2 Feedback in conversations

Our analysis of feedback or bachchannels[2] in the AMI corpus provides us with a good example of how the various steps in the method outlined before relate to each other: how one level feeds into another and back.

[2] The term backchannel was coined by Yngve (1970) and is derived from the notion of a "back channel" through which the listener sends the speaker short messages, such as "yes" and "uh-huh", that are not a bid for the floor. Which types of utterances can be considered backchannel activity is often debated. The very short messages like "mmm," "yeah," "right," -which are common in English- clearly qualify because they add a great deal to the quality of the interaction without really adding meaning to the conversation. However, Yngve also considers questions such as, "You've started writing it, then, ... your dissertation?" and short comments such as, "Oh, I can believe it," to be backchannel utterances. Duncan [Duncan and Niederehe, 1974] added other types of utterances to the list, such as sentence completions, requests for clarification, and brief restatements, since their purpose is not actually to claim a turn, but to provide the speaker with needed feedback.

The starting point of our research was technological but accompanied by other research questions as well. In multi-party interactions it is not always obvious who is addressed by an utterance of the speaker. Being able to detect it automatically is an important challenge. As we have seen in Section 3, conversational acts are joint actions where actions of speakers are complemented by actions of listeners. The recipients of communicative behaviours will typically display certain behaviours that provide feedback to the speakers. From this it should follow that if we can recognize this feedback behaviour, we may use this to identify a potential addressee of the speaker action. Clearly, someone who provides feedback, felt addressed in some way or another. Similarly, certain actions of speakers may provide us with information on whom he is addressing as well. It has been pointed out (for instance by [Goodwin, 1981]) that speakers may indicate whom they are addressing by looking at them at certain points in the utterance. The combination of feedback and gaze cues also leads one to an hypothesis about their co-occurrence. Could it be a regular feature of conversation that feedback of listeners occurs at positions where the speakers gaze at them? In those cases, listeners know that they are being addressed and that speakers are attending to them and can perceive the feedback. Information about statistics such as these[3] can also be used in the design of our conversational agents. In particular, we have been looking at the implementation of agents that can provide appropriate feedback and this kind of information would help in the timing of the feedback ([Heylen, 2007]).

The AMI corpus consists of more than 100 hours of video and audio recordings of four person meetings. We already mentioned the annotation layers with which the data was enrichted. Several of these are relevant to answer this question. The hand-coded dialogue acts contain several labels for feedback acts and other relevant information: on the relations between different utterances and on addressing. Information on the focus of attention, to whom or what somebody is looking, is also present in the AMI corpus. The corpus and the annotations thus appear to be ideal to answer our questions. However, for several reasons the solution was not as straightforward as may appear.

The AMI dialogue act annotation manual distinguishes three types of feedbacks: Backchannel, Assess and Comment-about-understanding. The Backchannel class largely conforms to Yngve's notion of backchannel and is used for the functions of contact. Assess is used for the attitudinal reactions, where the speaker expresses his stance towards what is said, either acceptance or rejection. Comments about understanding are used for explicit signals of understanding or non-understanding.

In addition to dialogue acts the coding scheme specifies that certain relations between dialogue acts should be annotated. Relations are annotated between two dialogue acts (a later source act and an earlier target act) or between a dialogue act (the source of the relation) and some other action, in which case the target is not specified. In the AMI scheme, relations are a more general concept than the adjacency pairs from the Conversational Analysis literature,

[3] See also [Heylen, 2006] and [Poppe et al., 2007].

like question-answer. Relation have one of four types: positive, negative, partial and uncertain, indicating that the source expresses a positive, negative, partially positive or uncertain stance of the speaker towards the contents of the target of the related pair. For example: a "yes"-answer to a question is an inform act that is the source of a positive relation with the question act, which is the target of the relation. A dialogue act that assesses some action that is not a dialogue act, will be coded as the source of a relation that has no (dialogue act as) target.

Since Backchannels were assumed to be always in response to what the main speaker at that point is saying, annotators did not enter them in a relation with another dialogue act, assuming that this could be detected automatically. But in order to check the relation between the gaze target of the speaker and the one who gives the feedback, we need to find for each occurrence of a backchannel act the related dialogue act that the backchannelor responds to, as well as the speaker of this related dialogue act. As the AMI dialogue act annotation does not contain the annotation of the relation between backchannel acts and this dialogue act (or turn), we have to define a new challenge: define an algorithm that decides which utterance a back-channel is related to.

We implemented and tested a procedure for doing this. We have validated our method for finding the related dialogue act (measured by recall and precision) and selected some parameter values for the time between act and backchannel act that gave us the best performance. The results can be found in the following table.

	correct	incorrect	uncertain	total
found	77	3	3	83
not unique	44		17	61
total	121	3	20	144

Fig. 1. Results of the method for finding the related dialogue act that a backchannel responds to.

For 83 out of the total of 144 backchannel events the procedure reported to have found a unique related dialogue act. Of these 77 was the correct one, 3 were incorrect and in 3 cases the answer was questionable. In these cases it was actually not clear what the related utterance is. Of the 61 cases in which the procedure reported that no unique related act was found, there were 44 cases in which there was a unique related dialogue act but the algorithm failed to identify it. In 17 cases it was unclear also from manual inspection to identify a related dialogue that the backchannelor responded to. If we leave out these uncertain instances we end up with 124 cases, and 77 correct and 3 incorrect answers, hence the method has a recall of 77/124 (62%) and a precision of 77/80 (96%).

When we checked the outcome of the procedure the major causes for not finding a *unique* dialogue act were the following.

1. There is simultaneous speech of multiple speakers. This occurs in animated discussions, where speakers sometimes express their ideas in cooperation. The backchannelor is responding to the idea expressed not to one speaker.
2. A particular situation arises when a speaker pauses then continues and right after the continuation the backchannelor is reacting on the part before the short pause. The method will not find the correct act because it seems to respond to the continuation and not the previous part. This is a serious case because continuer signals often occur in the middle of speaker turns where short pauses or segment boundaries occur.

It is also interesting to look at the case where the relation between back-channel and the related utterance was also unclear in the manual annotation. The following cases were found where the situation was indeed unclear to identify a related speaker and act. These are cases

1. with simultaneous cooperative talk
2. with the absence of a dialogue act (in particular "backchannels" such as "Okay" were used mostly as a closing signal)
3. where back-channels appeared to be instances of self talk ("Mmm", "yeah") and not directed to a particular other contribution.

The analysis we performed shows how initial assumptions that inform the annotation of data might need revision.

Next we looked at the relation between speaker gaze and the occurrence of feedback. For each of the 13 meetings we computed for each pair of participants (X, Y) the length of time that X looks at Y while X is speaking and we computed the length of time X looks at Y when X performs a dialogue act and Y responds with a backchannel act. Analysis of these pairs of values shows that in a situation where someone performs a backchannel the speaker looks significantly more at the backchannelor than the speaker looks at the same person in general when the speaker is performing a dialogue act ($t = 8.66$, $df = 101$, $p < 0.0001$). The mean values are 0.33 and 0.16. This confirmed our hypothesis.

This example shows a number of issues related to the methodology that we and many others follow in human computing research. We collect and describe data for our data-driven methods, building on theories of human communication. The theories and insights that lead to the development of an annotation scheme may not be full-proof. The cracks in the theory or the cases that are not covered can be detected by the cycle of research that is exemplified here. In this way, engineering is based on knowledge but also the basis of knowledge. In our case, it lead us further to rethink the theories of participation and action in conversations that formed the basis for our initial annotation framework. However, this does not mean that we have now reached the ultimate theory of communication. This leads us to the conclusion.

6 Human Computing

One can view human computing technology as the current state of the historical development of technology, that aims at simulating human interaction as an aspect of human behavior as such. It shows the boundaries of the scientific methodology, inherited from mathematics, physics and technology, and the way we conceptualize nature and human behavior, according to the principles of this world view. What is central in this development is a principle conflict between on the one hand the natural openness of natural language and human behavior, in which the human mind freely assigns through his practice and in communication with the world and others new means of signifying, and new processes of control. Technology itself is the phenomenon where this creativity shows and this means that technological development will always be essentially an incomplete objectification of human creative, sense giving, behavior. We have pointed out some of the consequences that this has for the design of human computing systems. We have put emphasis on the role of the *contract* between user and technology, its essential role for the working of technology in practice. The contract is the issue that has to be evaluated, if we evaluate our technical systems in practice, because it forms the interface between the qualitative measures of the user and the quantitative measures that function in the design of the system.

The identity of the "virtual human" that technology brings forward, i.e. the ambient intelligence that is our technological communication partner, is essentially a mathematical identity, either in the form of a statistical model or of an analytical model. When we construct this technology we have the obligation to "show" that it satisfies the contract that is implicit in its design. The reliability of the data-analysis is essential for the contract being possible and for the ecological validity of the experiments that underly the models that are implemented. The contract between users and system designer make up the context for the services that the system offers the user in using the technology in his practice.

The great challenge for human computing is in further clarification of the central concepts that play a role in the embedding of technology in interaction with man; concepts like service, what technology affords, and context, and related notions as context-awareness.

Acknowledgements This work is supported by the European IST Programme Project FP6-0027787. This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein.

References

[Bavelas and Chovil, 1997] Bavelas, J. and Chovil, N. (1997). Faces in dialogue. In Russell, J. and Fernandez-Dols, J.-M., editors, *The psychology of facial expression*, pages 334–346. Cambridge University Press, Cambridge.

[Buytendijk, 1964] Buytendijk, F. (1964). *Algemene theorie der menselijke houding en beweging.* Het Spectrum, eight printing 1976, (in Dutch).

[Chomsky, 1965] Chomsky, N. (1965). *Aspects of a Theory of Syntax*. MIT Press, Cambridge Massachussets.

[Clark, 1996] Clark, H. (1996). *Using Language*. Cambridge University Press, Cambridge.

[Clark and Schaefer, 1989] Clark, H. and Schaefer, E. (1989). Contributing to discourse. *Cognitive Science*, 13:259294.

[Duncan and Niederehe, 1974] Duncan, S. and Niederehe, G. (1974). On signalling that its your turn to speak. *Journal of Experimental Social Psychology*, 10:234–47.

[Goffman, 1976] Goffman, E. (1976). Replies and responses. *Language in Society*, 5(3):2257–313.

[Goodwin, 1981] Goodwin, C. (1981). *Conversational Organization: Interaction between Speakers and Hearers*. Academic Press, New York.

[Heylen, 2006] Heylen, D. (2006). Head gestures, gaze and the principles of conversational structure. *International Journal of Humanoid Robotics*, 3(3):241–267.

[Heylen, 2007] Heylen, D. (2007). Multimodal backchannel generation for conversational agents. In van der Sluis, I., Theune, M., Reiter, E., and Krahmer, E., editors, *Workshop on Multimodal Output Generation*, pages 81–91, University of Twente. CTIT.

[Heylen et al., 2005] Heylen, D., Ghijsen, M., Nijholt, A., and Akker op den, H. (2005). Facial signs of affect during tutoring sessions. In Tao, J., Tan, T., and Picard, R. W., editors, *Affective Computing and Intelligent Interaction - First International Conference, ACII 2005*, pages 24–31. Springer-Verlag. ISBN=3-540-29621-2.

[Heylen et al., 2004] Heylen, D., Vissers, M., Akker op den, H., and Nijholt, A. (2004). Affective feedback in a tutoring system for procedural tasks. In André, E., Dybkjr, L., Minker, W., and Heisterkamp, P., editors, *ISCA Workshop on Affective Dialogue Systems, Kloster Irsee, Germany*, pages 244–252, Berlin Heidelberg New York. Springer-Verlag. ISBN=3-540-22143-3.

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

[Kraut et al., 2003] Kraut, R., Fussell, S., and Siegel, J. (2003). Visual information as a conversational resource in collaborative physical tasks. *Human-Computer Interaction*, 18:13–49.

[McCowan et al., 2005] McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., M.Kronenthal, Lathoud, G., Lincoln, M., Lisowska, A., Post, W., Reidsma, D., and Wellner, P. (2005). The ami meeting corpus. In *Measuring Behaviour, Proceedings of 5th International Conference on Methods and Techniques in Behavioral Research*.

[Montague, 1970] Montague, R. (1970). English as a formal language. In Visentini, B., editor, *Linguaggi nella società e nella tecnica*, pages 189–224. Edizioni di Comunità, Milan.

[Nakano et al., 2003] Nakano, Y. I., Reinstein, G., Stocky, T., and Cassell, J. (2003). Towards a model of face-to-face grounding. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 553–561. ACL.

[Pantic et al., 2007] Pantic, M., Pentland, A., Nijholt, A., and Huang, T. (2007). Machine understanding of human behavior. In *Proceedings AI for Human Computing (AI4HC'07). Workshop at IJCAI 2007*, pages 13–24, Hyderabad, India.

[Poppe et al., 2007] Poppe, R., Rienks, R., and Heylen, D. (2007). Accuracy of head orientation perception in triadic situations: Experiment in a virtual environment. *Perception, to appear*.

[Schegloff, 1982] Schegloff, E. A. (1982). Discourse as interactional achievement: Some uses of "uh huh" and other things that come between sentences. In Tannen, D., editor, *Analyzing discourse, text, and talk*, pages 71–93. Georgetown University Press, Washington, DC.

[Traum and Rickel, 2002] Traum, D. and Rickel, J. (2002). Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS 2002)*, pages 766–773.