GRETA. A BELIEVABLE EMBODIED CONVERSATIONAL AGENT

I.Poggi, C.Pelachaud, F. de Rosis, V. Carofiglio, B. De Carolis

1. INTELLIGENT BELIEVABLE EMBODIED CONVERSATIONAL AGENTS

A wide research area in the realm of Autonomous Agents is presently devoted to the construction of ECAs, Embodied Conversational Agents (Cassell et al. 2000; Pelachaud & Poggi, 2001). An ECA is a virtual Agent that interacts with a User or another Agent through multimodal communicative behavior. It has a body (a realistic or a cartoon-like one), thanks to which it can produce spoken words, sentences, discourses, dialogues, use voice with appropriate prosody and intonation; it can exhibit the visemes corresponding to the words uttered and make gestures, assume postures, produce facial expression and communicative gaze behavior.

A particular feature that is required from an ECA is that it is also a Believable Agent, that is, one able to express emotion (Bates, 1994) and to exhibit a given personality (Loyall & Bates, 1997). But, according to recent literature (Trappl & Payr, in press; de Rosis et al., in press a), an Agent is made even more believable if it can behave in ways typical of given cultures, and finally, if it has a personal communicative style (Canamero & Aylett, in press; Ruttkay et al., in press). This is, in fact, what makes a human a human. Again, in its being a Conversational Agent, an ECA must be interactive, that is, take User and context into account, in such a way as to tailor its interaction to the particular User and context at hand.

An ECA that fulfils all of these constraints is one where the communicative output, that is, the particular combination of multimodal communicative signals displayed (words, prosody, gesture, face, gaze, body posture and movements) are determined by different aspects: *a.* contents to communicate, *b.* emotions, *c.* personality, *d.* culture, *e.* style, *f.* context and User sensitivity. At each given moment of a communicative interaction, all of these aspects combine with each other to determine what the Agent will say, and how.

In this paper we show how these aspects of an ECA can be modeled in terms of a belief and goal view of human communicative behavior. We then illustrate Greta, an ECA following these principles which is being implemented in the context of the EU project MagiCster (IST-1999-29078).

T.H.E. Editor(s) (ed.), Book title, 1—6.
© yyyy *Kluwer Academic Publishers. Printed in the Netherlands.*

2. WHAT AND HOW WE COMMUNICATE

Before focusing on the structure of the dialogue, let us see how a single move can be represented and simulated. An Agent (the Sender), generates a set of beliefs - about itself or about external objects and events - and has the goal to make another Agent (the Addressee) believe them. To achieve this, the Sender produces a set of communicative signals (for example words, gestures, gaze, head, face and body behavior), taken from its communicative repertoire and temporally ordered in a particular way.

Now, what are the beliefs an Agent may conceive to communicate? What is the structure of its communicative repertoire? What are the rules for the temporal ordering and synchronization of different signals? But also: how does the system go from the input – a set of beliefs – to the output – that particular arrangement of words, voice, hands, body, face signals? In fact, our communicative repertoire is very complex, in order for us to modulate our communication in a very sophisticated way. For example, words have formal and informal variants, connotations, positive or negative, tender or insulting nuances; and not only is the verbal lexicon so rich, but we can also communicate by subtly different intonations, gestures, facial expression, gaze, posture, spatial behavior.

In each move of a dialog, how do we choose the best way to communicate, the combination of verbal and non verbal signals that are most fit to express our communicative goal? How do we activate the goal of using that given word, replacing it with a gesture or using both to communicate our meaning?

According to a goal and belief model of social action (Conte & Castelfranchi, 1995), choice, that is, the decision to pursue a goal instead of another is determined by the relative values attached to the alternative goals. But the value of a goal in its turn stems from the value of its superordinate goals, or from the algebraic sum of the values of two or more of them. So, whoever discovers his car was stolen might shout. But if this happens to someone who is just starting a work where he needs his car, his shout will be sharper or longer and his utterance more aggressive. In other words, resources we specifically use in a given communicative situation (a gesture in the place of a word, a very colloquial term instead of a more formal one) are determined by a number of permanent and contingent factors.

3. PERMANENT AND CONTINGENT FACTORS

As an Agent enters a communicative interaction, having the goal of communicating some meanings, two kinds of factors affect the final aspect of his/her communication: permanent and contingent ones (Table 1). The former are the goals and resources coming from the Senders' biological and cultural endowment, that are always active in them; the latter are the goals activated and the resources provided by the contingent situation in which the Senders communicate.

Long-lasting internal resources are (1) personality, (2) social identity (age, gender, cultural roots) and (3) cognitive traits. Among them, we may distinguish (4) innate and (5) culturally learned features: innate ones may be (6) a higher or lower

capacity of making inferences, different reasoning styles (more abstract, intuitive, or imaginative); or (7) the different aptitudes, partly depending on neurological dispositions, towards musical, mathematical, visual or linguistic skills. Other cognitive capacities are culturally dependent: culture entails beliefs about the environment (8) and about strategies of behavior typical of a population (see Section 8.), but also norms on how to do things, how to behave, what and how to communicate (9), and finally the communication repertoires (verbal and nonverbal) one comes to learn since infancy.

Table 1. Factors affecting the choice of communication resources

contents to communicate ***communicative choice***

<table>
<tr><th colspan="3">PERMANENT</th><th colspan="3">CONTINGENT</th></tr>
<tr><td colspan="3">1 S's personality</td><td rowspan="2">11
Self</td><td colspan="2">14 Physical resources
(motor capacity and energy)</td></tr>
<tr><td colspan="3">2 S's social identity</td><td colspan="2">15 Cognitive resources
(drunk, concentrated)</td></tr>
<tr><td rowspan="7">3
S's
Cognitive
Traits</td><td rowspan="2">4
Innate</td><td>6
Inference capacity</td><td rowspan="7">12
Other</td><td rowspan="3">16
Physical resources</td><td>18
Sensory capacity</td></tr>
<tr><td>7
aptitudes</td><td>19
Media</td></tr>
<tr><td rowspan="5">5
Learn-
ed</td><td>8
Knowledge Base</td><td rowspan="4">17
Cognitive resources</td><td>20
Knowledge Base</td></tr>
<tr><td>9
cultural techniques, norms, values</td><td>21
Inference capacity</td></tr>
<tr><td rowspan="3">10
communicative repertoire</td><td>22
Communicative repertoire</td></tr>
<tr><td colspan="2">23
Personality</td></tr>
<tr><td rowspan="6">13
Situation</td><td rowspan="2">24
Physical setting</td><td>25
Available modalities</td></tr>
<tr><td>26
Presence of referents</td></tr>
<tr><td>27
Social setting</td><td>28
S-A relation
(status, role, affect)</td></tr>
<tr><td colspan="2">29
Type of encounter
(service, affective)</td></tr>
<tr><td colspan="2">30
Relations to others
(in public, in private)</td></tr>
</table>

The combination of personality traits, attitudes and culturally learned behavior habits gives rise to style, an idiosyncratic tendency to behave in some peculiar way that allows others to recognize the Agent's identity. The contingent resources taken into account for the choice of the output communicative behavior are provided by the intrinsic features of context (Poggi & Pelachaud, 2000).

The presence or absence of these resources in oneself (11), the Other (12) or the Situation (13) activate specific goals in Senders as they plan their discourse and then the communicative signals to exploit. If I am very tired (physical resources, 14) I'll tend to speak low and not to make conspicuous gestures; if my interlocutor is a bit deaf (sensory capacity of other, 18), I'll speak loud. If he is a tourist (Other's communicative repertoire, 22) I'll talk slowly; with a student (Knowledge Base, 20) I'll explain things at length; with a not so smart person I'll explain even obvious causal links (inference capacity, 21). And if the interlocutor is a touchy person (personality, 23), I'll be more indirect in my criticism.

Beside the features of oneself and the interlocutor, resources present or absent in the environment trigger different goals about what and how to communicate: in a noisy discotheque (physical setting, 24), I'll use gestures instead of words (available modality, 25), while I will not do this on the phone (absence of referents, 26). But more than the physical, the social setting (27) is relevant: we use more polite words with a high status Addressee (S-A relation, 28) or in formal situations (type of encounter, 29); less colloquial words in public than in private (relation to others, 30).

How do we choose, then, how to communicate in a given situation? Presence or lack of individual resources lead us to choose one signal or combination of signals instead of another. Our hypothesis is that every combination of an Agent's multimodal communicative behavior (that particular word, uttered by a particular intonation, while making that gesture, that gaze, facial expression, and posture) is the result of the final choice of the communicative goal to pursue. This goal is selected among the Agent's different goals, determined in their turn by contingent events (content to communicate, felt emotions, context and interlocutor) and by long-lasting features (the Agent's culture, personality and style).

In Sections 4. through 9 we overview how the above aspects (meanings to communicate, emotion, personality, culture and style) can be viewed according to a goal and belief model of human communicative interaction, and how they may be represented in an ECA. In Section 10. we describe the architecture of the ECA "Greta", while showing how some of the above principles can be applied to construct a Believable Embodied Conversational Agent.

4. MEANINGS TO CONVEY

Let us first overview the beliefs that may form the content of a communicative act. Three classes of meanings can be distinguished (Poggi, 2002 b): Information on the **World**, Information on the **Speaker's Identity** and Information on the **Speaker's Mind**.

Information on the World. As we communicate, we provide information on concrete or abstract events, their actors and objects and the time and space relations

among them. Such information is provided mainly through words, but also by gestures or gaze. To mention the referents of our discourse, we may point at them by deictic gestures or gaze; to refer to some properties of objects we may use iconic or symbolic gestures and even gaze (as when we squeeze eyes to mean 'small' or 'difficult').

Information on the Speaker's Identity. Physiognomic traits of our face, eyes, lips, the acoustic features of our voice and often our posture provide information on our sex, age, socio-cultural roots, and personality. And, of course, our words can inform on how we want to present ourselves.

Information on the Speaker's Mind. While mentioning events of the external world, we also communicate why we want to talk of those events, what we think and feel about them, how we plan to talk of them. We provide information on the **beliefs** we are mentioning, our own **goals** concerning how to talk about them and the **emotions** we feel while talking (Poggi 2002 a).

About **beliefs**, we inform about:

1. *degree of certainty*: words like 'perhaps', 'certainly'; conditional or subjunctive verb modes; but also frowning, which means 'I am serious in stating this'; opening hands, which means 'This is self-evident';
2. *metacognitive* information: that is, the source of mentioned beliefs, whether they come from memory, inference or communication (we look up when trying to make inferences, snap fingers while trying to remember,...)

We inform about the following **goals**:

1. *performative* of the sentence (by performative verbs, intonation, facial expression);
2. *topic-comment* distinction (by batons, eyebrow raising, voice intensity or pitch);
3. *rhetorical relations*: class-example (saying first, second, third...; counting on fingers) topic shift (expressed through posture shift);
4. *turn-taking and backchannel*: raise hand for asking turn; nod to tell the Interlocutor we are following what she says.

Finally, we inform on the **emotions** we feel while talking (by affective words, gestures, intonation, facial expression, gaze and posture).

5. COMMUNICATIVE REPERTOIRE

Let us now see what is the structure of the communicative repertoire (Table 1, n.10): the set of innate and learned features and behaviors that we can use as signals in order to convey the meanings specified above. The Human Agent is endowed with different "mode-specific communicative systems", that is, sets of rules to link meanings to signals that work in different modalities.

Several parts of the Human Agent's body produce communicative signals: so head, face (with eyes and mouth), hands, trunk and legs can be viewed each as the depository of a communicative system. Communicative systems may be of at least

two kinds: “codified” and “creative”. In a “codified” system, or “lexicon”, the same signal-meaning link is shared and coded in the Agents’ memory. Not only words or symbolic gestures but also gaze, facial expression, posture form lexicons, where each signal represented in an Embodied Agent’s mind corresponds to a specific meaning. In a “creative” system, instead (for instance the system of iconic gestures), what is coded in memory is only a small set of inference rules about how to create a new signal starting from a given meaning, or about how to retrieve a meaning from a given signal (Magno Caldognetto & Poggi, 1995).

In our hypothesis, in all communicative systems, meanings may be represented in terms of ‘mental images’ or logical propositions. For example:

Let `Ai` denote the Sender and `Aj` the Addressee, `a` an action and `g` a goal that may be achieved by means of `a`. The performative of a ‘Peremptory order’ may be represented as follows:

1. Goal Ai (Do Aj a)
2. (Goal Ai g) $\wedge$ Bel Ai (Achieve a g))
3. Goal Ai (Bel Aj (Power-on Ai Aj a))
4. If (Not (Do Aj a)) then (Feel Ai Angry)

while the performative of an ‘Imploration’ may be represented as follows:

1. Goal Ai (Do Aj a)
2. (Goal Ai g) $\wedge$ Bel Ai (Achieve a g))
3. Goal Ai (Bel Aj (Power-on Aj Ai a))
4. If (Not (Do Aj a)) then (Feel Ai Sad)

In both ordering and imploring, the Sender wants the Addressee to do some action. However, the two performatives differ for the power relationship between the two interlocutors (`Ai` has power over `Aj` in the former, and the reverse in the latter), and for the potential emotion (anger vs. sadness) in case `Aj` does not perform the requested action (Poggi & Pelachaud, 2000).

A compositional representation can be adopted not only on the side of the meaning, but also for the signals: a signal is represented as a combination of behavioral parameters, each with its number of values. Gestures are decomposed into hand-shape, arm and wrist position, type of movement; gaze into direction of eyes, eyelid aperture, movements of the eyebrows, and so on (Poggi, 2002 b). This compositional representation on the two sides enables achieving a high flexibility in the correspondence between signals and meanings. It opens, as well, the possibility that a combination is not always conveyed by the same signal, but may be expressed through different parameters of signals taken from different communication systems (Pelachaud & Poggi, in press).

The overall communicative repertoire, in fact, manages the combination of signals taken from different communicative systems, the distribution of meanings across them, and their temporal ordering and synchronization.

6. EMOTIONS

Before showing why and how emotions are an important determinant of an Agent's communication, let us define them according to a goal and belief model of action and social interaction (Conte & Castelfranchi, 1995).

6.1. Emotions and goals

Actions in our life are often part of a plan aiming at some goal. Take for example an action of Oetzi, the 5000 B.C. pre-historic man of Similaun, who chooses a stone apt to sharpen well and makes a lance for chasing the wild-pig successfully. The actions of looking for a good stone and to sharpen it are just means for the complex action of chasing. But also chasing is aimed at feeding himself and his group, which in turn aims at the goal of survival. The goals of our everyday plans are not ends in themselves. They all aim at more general goals of biological import that are common to all humans, like the biological goals of survival and reproduction and some subgoals of them, physical well-being, safety, loving and being loved, self-realization, image and self-image. These are *terminal* goals, that are ends in themselves and ones to which we assign the highest weights. So much that, if two of them are incompatible (as for instance freedom vs. life itself), giving up one of them is a heavy renunciation. With respect to terminal goals, the goals of our everyday life are *instrumental* goals, in that they directly or indirectly serve our terminal goals. For instance, chasing the wild-pig with a sharpened stone is instrumental to survival: if the lance is not sharp enough and does not hit the wild-pig to death, the wild-pig might aggress and kill Oetzi. Instrumental goals are more or less important to us, depending on the strength of their link with terminal goals. At the extent to which an Instrumental Goal is likely to be the only possible means to reach a Terminal Goal, that Instrumental Goal receives a high weight, just because it inherits its weight from the Terminal Goal it serves.

Emotions are a biological device aimed at monitoring the state of reaching or threatening our most important goals, be they Terminal or Instrumental (see, for instance, Carbonell, 1980; Oatley & Johnson-Laird, 1987). Anytime something happens (or the Agent believes it happens) that is likely to produce the achieving or threatening of a highly weighted goal, the biological device of emotion is triggered. From the agent's interpretation of the situation, a complex subjective state originates, generally of a short duration and with different degrees of intensity. This state includes physiological, expressive and motivational aspects. If Oetzi throws his lance but sees it has not run into the wildpig's heart, fear is triggered since his goal of survival is challenged: physiological reactions are activated (blood flowing away from face to limbs) some of which may show in the perceivable state of his body (pale face, tremors); and the specific goal of escaping, that might serve the terminal goal of survival, is activated.

There is a strong relationship, then, between goals and emotions: goals both cause emotions and are caused by emotions. They cause emotions since, if an important goal is achieved or threatened, an emotion is triggered: emotions are therefore a feedback device that monitors the reaching or threatening of our high-

weighted goals. At the same time, emotions activate goals and plans that are functional to re-establishing or preserving the well being of the individual, challenged by the events that produced the emotions. So, fear triggers flight, anger triggers aggression, guilt triggers the goal of helping the harmed person or of escaping sanction (Castelfranchi, 2000).

6.2. Emotion triggering vs. emotion display

Emotions may be implied in communication in at least two ways.

1. they may be the very reason that triggers communication: we activate the goal of communicating just because we want to express our emotion;
2. they may intervene during our communication, as a reaction to what our interlocutor is saying, or to some thought suddenly coming to our mind, either related to the ongoing dialogue or not.

In both cases, the triggering of emotion does not necessarily imply that the Agent displays it. There are many reasons why we may refrain from expressing our emotion, and the final (aware or non-aware) decision of displaying it may depend on a number of factors (Prendinger & Ishizuka, 2001; De Carolis et al., 2001). Some of them concern the very nature of the emotion felt (emotional nature), others the interaction of several contextual (scenario) factors.

1. Emotional nature
    a. *Intensity* (a more intense emotion might be more likely displayed);
    b. *Valence* (it is not the same to display negative or positive emotions);
    c. *Social evaluation* (some emotions, like envy or shame, are subject to social sanction: then it is more difficult to express them);
    d. *Addressee* (it is different to express an emotion to the one who caused it or to a third person).
2. Scenario Factors
    a. Agent's *Display motive* (displaying or not depends on whether you do it to be helped, consoled, or if you want to demonstrate or teach something);
    b. Agent's *personality* (an impulsive person is generally more keen to displaying than a reflexive or a shy one);
    c. *Interlocutor*'s features (displaying depends on the other's personality, empathy, intelligence...).
    d. Agent - Interlocutor *Role relationship* (whether he has power over you or you over him);
    e. Agent - Interlocutor *Personal Relationship* (you might not display your being worried to someone you love, if you want to protect him);
    f. Type of *social interaction* (being in public makes a difference for emotion display).

7. PERSONALITY

Personality is also linked to goals: it can be viewed in terms of weights people attribute to terminal goals (Carbonell, 1980; Poggi and Pelachaud, 2000). For example, a sociable person gives more importance than others to knowing and staying with other people. A selfish person, when the goals of physical safety and others' care are conflicting, chooses to pursue the former, while an altruist pursues the latter. A proud person attributes a high value to self image and autonomy, while a dependent person cares others' image more than self-image.

Since both emotions and personality have to do with the relative importance of goals, there is also some link between emotion and personality. Some personality traits may be viewed in terms of the general 'propensity to feel emotions' (Plutchik, 1980; Poggi and Pelachaud, 1998). Picard (2000) calls 'temperament' this subset of personality traits, while other authors relate them directly to one of the factors in the 'Big-Five' model: for instance, neuroticism (Mc Crae & John, 1992). These traits imply, in a sense, a lower threshold in emotion feeling (Ortony et al., 1988). For instance, a 'shy' person is keener to feel 'shame', especially in front of unknown people. A 'proud' person, who attributes a high weight to his goals of self-esteem and autonomy, will feel particular pride (will be proud of himself) every time one of these goals is achieved. And, conversely, every time they are threatened (if, say, he is obliged to ask for help), the person will feel the opposite emotion, shame. Thus, a personality trait (proud) is related to attaching a higher weight to a particular goal (self-esteem, autonomy); and, since that goal is particularly important to that kind of person, the person will feel the corresponding emotions (pride or shame) with a higher frequency or intensity.

8. CULTURE

Culture may also be viewed in terms of different weights on goals. Both a Somali shepherd and an Italian housewife have the goal of feeding themselves and their family; but the sub-goal chosen to pursue this goal may be for the former to search the bush, for the latter to go shopping in a store.

8.1. Goals and culture

Humans pursue their goals by using their internal and external resources. External resources are the objective conditions holding in the environment (presence of food, characteristics of the territory, climate conditions and so forth); internal resources are the human's action capacities (physical strength, body agility, manual skill) and beliefs.

But, in different environments the physical conditions, hence the most easily available resources, and consequently also the actions to get them, are different. In the land it will be easier to get food by rearing sheep or cows, while on the coast

fishing will be the most direct way to feed. So, in the land the most useful beliefs to store and process and the most necessary actions to learn will concern sheep and cows rather than fish or shrimps, and people will more easily become shepherds than fishermen. Any population, given the environment in which it lives, comes to accumulate a set of beliefs on the instrumental goals that most easily and economically serve the biological terminal goals in that environment. In other words, an instrumental goal becomes more or less important in a culture depending on the strength and necessity of its link to a terminal goal. At the extent to which that instrumental Goal is (or is very likely to be) the only possible means to reach a terminal Goal, given the external conditions available in that culture, that instrumental Goal will receive a higher weight than other possible ones. Thus, the instrumental goal chosen becomes a strategy of survival typical of that culture; and culture, overall, may be defined as a set of beliefs on the typical techniques to pursue goals.

Of course, the techniques chosen are determined by beliefs on the environment: for example, until a population does not know the mechanisms of plant reproduction, the technique of agriculture cannot be chosen as an instrumental goal to survival. Therefore, culture entails beliefs about the external world. And since language is both produced by beliefs and a vehicle of them, culture typically shows up in language. Language is made of the beliefs of a population, but is also a way to organize them, a set of rules on how to conceptualize and categorize information. Consequently, it implies, again, a set of settled typical communicative techniques, that is of settled instrumental goals stating how to convey information.

In addition to the beliefs about the best techniques to pursue terminal goals, culture entails also values and norms. Values are evaluative beliefs about what is good and then has to be pursued as a goal (Miceli & Castelfranchi, 1989). But since particular ways to behave may be good or bad according to the environment, again due to what are the most useful techniques of survival, different populations in different environments may hold different values. For example, in the environments where individualistic behavior has proved to be convenient, individualistic values will develop; on the contrary, in environments where collectivistic behavior is more fit, values centered on the family or the group will hold.

Norms are obligations that rule the relationships among people in a group (Conte & Castelfranchi, 1995). Again, in a culture more centered on interdependency, a highly weighted goal, and then a norm that holds, may prescribe to be very cooperative with each other, even when this implies intruding in the other person's affair. On the contrary, in a culture more centered on the individual's autonomy, the goal of keeping one's privacy will be more weighted, and a norm will hold of not intruding in others' affairs and of contrasting others' intrusions.

Now, both values and norms generate goals in people (the goal to pursue that value or to respect that norm); if they are thwarted, they provoke emotions. Not living up to one's values may induce shame, while violating norms may cause a sense of guilt. Therefore, if two populations have different values and norms, they will also feel these emotions as a consequence of different events.

To summarize our definition, we may say that culture is a set of beliefs shared by a population. These beliefs regard the environment in which the population lives and

the best techniques (the most highly weighted instrumental goals) to reach the biological terminal goals in that environment, given the means-end relations that hold in the given physical conditions and the set of beliefs accumulated. Culture also includes beliefs about how to gather and organize beliefs themselves and about the norms and values that are functional to techniques of goal achievement that best fit the surrounding environment.

According to this definition, one may try to figure out how the way people communicate changes with cultural differences, by trying to distinguish what is universal (biological) and what is culturally determined in the different aspects of communication. These differences may then be simulated in a Believable Embodied Agent, in words or discourse planning, in gesture, gaze, facial expression, body posture and proxemic behavior.

8.2. Semantic rules versus norms of use in communication systems

To see what is culturally variable and what is universal in communication systems, we need to distinguish between two kinds of rules: semantic rules and norms of use. *Semantic rules* concern the correspondence between signals and meanings in that system. For example:

if you want to communicate the meaning "I greet you", say "Hello".

Lexical and syntactic rules of this kind compose the communication system of a verbal or sign language. Also gesture and gaze systems contain rules of this type:

if you want to communicate the meaning "I greet you", raise your eyebrows, or
if you want to communicate the meaning "I greet you", wave your hand.

Norms of use, instead, do not state how some meaning has to be conveyed, but if some meaning can, should or should not be conveyed in a given situation. An example of such a rule is the following:

if you meet a person you know, apply the rule for the meaning "I greet you", or
if you meet an unknown person, do not apply the rule for the meaning "I greet you".

Thus, a communication system tells us, on one hand, what we should do if we want to convey a specific meaning; on the other hand, if it is prescribed, accepted or forbidden to convey a such meaning. The latter are the norms of use of a communication system, the 'norms of appropriateness' traditionally studied by sociolinguists.

Now, in some communication systems (say, a spoken language) both semantic rules and norms of use vary across cultures. For other systems (typically, the facial expression of primary emotions) semantic rules might be everywhere the same (a grimace of anger is performed in the same way in all cultures), while the difference across cultures is in the norms of use (in Japan expressing anger is much more sanctioned than in the USA).

We can distinguish two kinds of norms of use:

1. *global rules*: in some cultures gestures are sanctioned as a more primitive or socially low-level way of communicating; so, they occur seldom and tend to be less varied and less conspicuous.

2. *local rules*, regarding:
   a. **meanings.** Take language interdiction: using a particular signal may be forbidden just because the meaning it conveys is tabooed. Obscene words may be more tabooed in a traditional society than in one with fewer inhibitions. Therefore, for obscene words cultural variation holds not only about how a given meaning is conveyed (semantic rule) but also about whether or when that meaning may be conveyed (norm of use). In communication systems different from the verbal one, cultural variation seems to hold not so much about semantic rules as about norms of use. In gaze, a loving gaze, a seductive gaze or a gaze looking down at someone are probably performed in the same way and have the same meaning all over the world. But if, according to cultural norms, looking down at someone is considered very impolite, or clearly showing love to someone is seen as incorrect or obscene, in those cultures those types of gaze will be tabooed and will not be easily performed.
   b. **signals**. Some norms of use impinge on signals themselves, their physical forms and the actions to perform them. For instance, in Ghana a taboo holds on the use of the left hand, that is considered impure; hence, gestures are more easily performed with the right than with the left hand (Kita and Essegbey, 2001).

8.3. Culture-sensitive words, sentences and discourses

Even if, at a very deep level, the syntax of all languages might be universal (as argued by the Chomskian Universal Grammar approach), specific words and syntactic rules are very different from one language to another. Again, some mechanisms like iconicity hold in all languages, but cultural variation across languages holds at various levels. These variations concern strategies of discourse planning, importance given to politeness or rhetoric and rules defining what to speak about, how much to mention the self in communication, whether to convey new information or just make reference to shared knowledge, and so on.

8.4. Culture-sensitive vs. universal gestures

The issue of universally shared vs. culture specific signals is particularly tricky in gestures, because different types of gestures exist. Among codified gestures, some are culturally codified: for example, the gesture for ‘OK’, or Churchill’s gesture for ‘Victory’). Others may be biologically codified: for example, those which are a

ritualization of physiological movements, like the gesture of raising fists up to show elation. If we want to simulate gestures of the former kind in a culture-sensitive Agent, they will have to be varied from a culture to another. Creative gestures like the iconics, instead, might all be generated through the same set of inference rules (supposedly universal), whatever the culture the Agent comes from. Of course, at the extent to which the referent represented is typical of a culture or an action is performed in a way typical of it, then also a creative gesture may be culturally dependent.

8.5. Culture-sensitive vs. universal gaze and facial expression

Facial expression and gaze are more likely to be universally shared than gesture. They can communicate information on the world (we point at things with chin or gaze, squeeze eyes to say that something is little or difficult), information on the Speaker's beliefs, goals and emotions (we raise eyes while we remember or make inferences; we frown to communicate anger, concentration, or an order) and information on the Speaker's identity (our face and gaze provide information on sex, age, ethnicity, personality, sometimes even social class).

Let us focus our analysis on face and gaze expression of emotions, to discuss whether the feeling of emotions is universal. Affective lexicons do differ across cultures (Russell et al., 1989); however, the so-called *basic emotions* (happiness, sadness, anger, fear, surprise and disgust) are felt in all cultures, and everywhere they trigger an innate universal neural program for facial expression (Ekman, 1982). This does not necessarily imply that people in different cultures always show their emotions in the same way in the same situation. Two cultural factors may intervene in this case, to produce wide differences in emotion expression. First of all, an emotion is triggered by the cognitive categorization of a situation on the part of the subject. So, a situation that in a culture (because of its beliefs, norms and values) is categorized as a cause of sadness, in another culture (with different beliefs, norms and values) might be categorized as a cause of happiness. For example, the death of a beloved person in a catholic group or the death of a martyr in the Islamic culture may be greeted as a cause of joy. Secondly, the filtering of emotion display to decide whether and how the felt emotion should be expressed, includes factors like cognitive and personality traits of the Agent and of the Interlocutor, their relationship and the situation, but also the cultural norms about the expression or non expression of given emotions.

9. STYLE

Style is an internal feature of an Agent that affects its choice of communicative resources. We define style as the idiosyncratic stable tendency of a specific agent to exhibit specific communicative or non-communicative behaviors. Communicative style is then the tendency one has to choose some signals, or arrangements or aspects of signals, instead of others. Agent X uses formal words also in private informal situations; Y's discourse is always thoroughly explained; Z makes ample

gestures; W avoids the Addressees's gaze. In its being a stable tendency to communicative choice, style is the ultimate result of the combination and interaction of the permanent goals that rule an Agent's behavior, namely the goals stemming from its personality, attitudes, and culture. A particular combination of these goals, at the moment of the communicative output, affects the choice of a particular signal or set of signals, or simply the value of some parameters in the production of signals. High level goals implied in a specific personality get instantiated into lower level goals. For example, in an introverted person, the general goal of not being too visible may result in a goal of avoiding too wide gestures. A tendency to visual thought may lead to use metaphors in discourse. The habit of explaining to students may induce to be always didactic.

Style can manifest itself both in the choice of whole signals (I may prefer a metaphorical to a literal word) and in the choice of a particular way to produce a signal. For example, people may differ in their style as to various aspects of gesture use. Differences may concern:

1. **whether to use a gesture or not**, which depends on at least three factors:
   a. *width of repertoire*: we cannot convey a meaning by a gesture if our culture or personal gestural competence does not include a gesture for that meaning;
   b. *threshold*: the degree of formality, as well as other factors like need for redundancy, may provide a threshold for the use of a gesture;
   c. *redundancy*: I may use both a word and the corresponding gesture if I want to stress some meaning or if I want to be very clear.
2. **ways of performing gestures.** While some parameters and sub-parameters of nonverbal signals like gestures are subject to lexical and sociolinguistic variation, some sub-parameters of movement (tense or relaxed (**tension**), wide or narrow (**amplitude**), smooth or angular (**manner**) might be more affected by style. That is, they are directly determined by the current goal set of the Sender and then by his/her emotional or physiological state.

In conclusion, each signal chosen and the combination of particular values in each parameter or sub-parameter of this signal is determined by the combination of goals that are activated at a certain moment in the Sender, all stemming possibly by the goal set described above.

10. GRETA'S EMOTIONS, PERSONALITY AND COMMUNICATION

In this Section, we show how some of the principles introduced above have been applied in the construction of an Embodied Agent: Greta. Specifically, the device of emotion triggering and its influence on the Agent's behavior have been already implemented, along with some aspects of its personality, communicative planning, voice and face communication. How culture and style affect the final output of the Agent's communication and, in particular, some aspects of gestural communication are still to be thoroughly implemented in Greta or other Agents, although some steps

have been moved in this direction (de Rosis et al., in press a; Ruttkay, et al., in press; Hartmann et al., 2002).

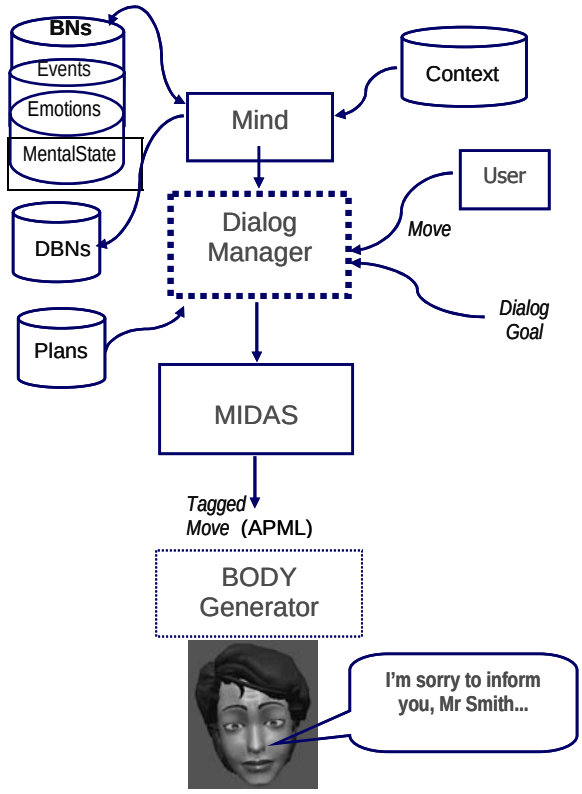
10.1. The architecture of Greta

In the context of the EU project MagiCster, we developed an Embodied Conversational Agent that appropriately combines verbal and nonverbal signals when delivering information, to establish a natural communication with the User. Our Agent achieves a quite rich expressiveness in conversation, by displaying various types of information typically transmitted in human-human dialogs (syntactic, dialogic, meta-cognitive, performative, deictic, adjectival, and rhetorical relation). It is embodied in a 3D talking head whose name is 'Greta'; it has a personality and a social role, as well as the capability of expressing emotions, consistently with the context in which the conversation takes place and with its own goals.

The type of conversations we simulate at present are information-giving dialogs, where the main function of Greta is to provide some kind of information to the User, in a given domain. In our 'mixed-initiative' system, the User can ask Greta some questions; this opens a question-answering sub-dialog after which, if needed, Greta revises her discourse plan according to the User request.

The architecture of our Conversational Agent includes the following main components (Figure 1): 1) the Agent's 'Emotional Mind', 2) a Dialog Manager, 3) a Plan Enricher and 4) a Generator of the Agent's Body.

Figure 1. The Architecture of our Conversational Agent.



The system is domain-independent. A particular application domain is therefore selected as a first step of interaction: a conversational goal in that domain is passed to the Dialog Manager and the dialog may start. From this goal, an overall discourse plan is produced for the Agent, by retrieving an appropriate 'recipe' from a plan library (de Rosis et al, 2002). This plan represents how the Agent will try to achieve the specified communicative goal during the conversation. The way the dialog goes on is a function not only of this plan but also of the User Moves and of the Social Context. This context includes variables concerning the topic of the dialog, the personality traits of the two interlocutors and their 'attitudes' towards the dialog itself. For instance, the Agent may be defined as more or less 'empathic' towards the User, more or less easily affected emotionally, more or less prone to feel specific emotions with a high intensity (for more details, see Carofiglio et al., in press).

Once an Agent's move has been generated, the User replies with a move in his or her turn. The Dialog Manager then asks the emotion simulation module (Mind) whether a particular affective state of the Agent is activated as a consequence of this move and with which intensity. The Agent's moves are enriched by the plan enrichment module (MIDAS) that adds tags indicating the meanings to be displayed through voice and facial expressions. This enriched move is then passed to the Body Generator, that interprets and renders it by producing a corresponding expressive behavior. We now describe in some more detail the above mentioned modules.

1. *Mind* is responsible for updating the Agent's mental state by deciding whether a particular affective state should be activated and with which intensity, and whether the felt emotion should be displayed and how, according to the context variables. Mind is based on a dynamic belief network (DBN), that combines a belief network (BN) representing the agent's mental state at time T with a network representing its mental state at time T+1 and a network that monitors the triggering of emotions in the interval (T, T+1). Three kinds of nodes can be found in the Agent's mental state: 'belief' nodes, 'goal nodes' and 'goal-achievement' nodes. A weight is associated with goal-achievement nodes, as a function of the agent's personality. The belief network at time (T+1) is generated according to the network at time (T) and to the events occurred in the interval (T,T+1). These events are modeled, as well, by belief networks.

2. the *Dialogue manager* (DM) is built on the top of TRINDI toolkit (http://www.ling.gu.se/projekt/trindi/). It controls the dialogue flow by iterating the following steps:
    a. after a 'dialog goal' has been specified, an appropriate discourse plan is selected from the library of plan recipes and the first move is generated according to the first step of this plan. The 'dialog goal' becomes the main topic of the conversation;
    b. at the end of this first move, the initiative is passed to the User, that may ask questions to the agent about any subject concerning the main topic under discussion;

c. the User move is translated into a symbolic communicative act (through a simplified interpretation process) and is passed to the DM;
d. the DM decides "what to say next" by selecting the sub-plans to execute. At the same time, the information state of Trindi is updated;
e. the DM goes on cycling over steps 2 to 4, until the user leaves the conversation.

3. The *Plan Enrichment* module, called MIDAS, has the role of translating the symbolic representation of a dialog move into an Agent's behavior specification. A dialogue move may be very simple (i.e. greet, ask) or it may correspond to a small discourse plan (for instance: describe an object with its properties). In both cases MIDAS generates a tagged output expressed according to a particular XML specification which is interpretable by the Body of any Animated Agent (Affective Presentation Markup Language: APML. See De Carolis et al., 2002 for details). APML is used as a middle-ware level, to overcome integration problems between the mind and body components and, at the same time, to allow independence and modularity between them. The move is then represented as an APML string, in which the verbal part of the dialog act is enriched with the tags that are needed by the speech and body generation components of the Agent to produce the required expressions. In particular: rhetorical relations, deictic or adjectival information, certainty values, metacognitive or turn-taking expressions.

4. The *Body Generator* module interprets the APML-tagged dialog move and decides which signal to convey on which channel for each communicative act. We defined a lexicon of facial expressions and gaze: that is, a set of (meaning, signal) pairs. Each meaning in the taxonomy specified in Section 4 (certainty, metacognitive, comment etc.) corresponds to a particular configuration of parameters of gaze or facial expressions that Greta is able to exhibit. For example, among the same class "certainty", the meaning "certain" corresponds to a "small frown" while the meaning "uncertain" to "raise eyebrow". The Body we use is a combination of a 3D face model compliant with the MPEG-4 standard (Pelachaud et al., 2001) and of the speech synthetizer Festival (Black et al., 2002). The facial model is capable of performing the face and gaze expressions foreseen for our conversational agent. Each signal may be expressed as a set of facial parameters. We have developed a language to describe facial expressions easily (De Carolis et al 2002). The text of each dialogue move with its tags is given as input to the animation module and to Festival, which provides the duration of the phonemes and a wav file (an audio file). Phonemes are the smallest temporal unit considered. Knowing the phoneme duration enables us to retrieve the exact duration of any expression as defined by the tags in the dialogue move, thus ensuring synchrony between speech and other visual

activities.Tags get instantiated by their corresponding signals which are in turn transformed into facial parameters' values. When several tags occur in the same text span, if their corresponding signals act on the same facial regions with different values, a conflict may arise. For instance: the agent might have to perform a frown and a raised eyebrow on the eyebrow region; or a head nod and a head shake might have to be performed, on the head direction. In such cases, a conflict solver module takes as input the co-occurring meanings and produces as output a complex expression made of a mix of signals from different meanings (Pelachaud & Poggi, in press).

5. The various components (Mind, MIDAS, Greta and the DM itself) are connected through a *Graphical Interface* which controls activation, termination and information exchange for the various processes involved in the dialogue management.

10.2. An Example

To give an idea of how the system works, we show a simple example of the output; in this dialog, Greta represents a doctor who explains a drug prescription to a patient (the User). After an initial 'greeting', the Agent informs the patient about her disease. Then, the patient can ask any question concerning the main topic of the conversation. For instance, she may ask for clarifications about the disease and its severity, or even about the related therapy to undergo. Greta answers the User question by selecting the adequate plan steps and by showing the appropriate expression that is appropriate to the social context. Figure 2 shows a few annotated moves of the example dialogue.

Figure 2. An example of dialog annotated with APML.

Greta:

<APML><turn-allocation type="take">
<performative type="inform" affect="sorry-for" certainty="certain**">**
I'm sorry to tell you that you have been diagnosed as suffering from what we call angina pectoris,</performative>
<belief-relation type="eoa"> **which**
<performative type="inform" certainty="certain"> **appears to be**
<adjectival type="small">**mild.** </adjectival> </performative> </belief-relation>
</turn-allocation></APML>

User: What is angina pectoris?

Greta:

<APML><turn-allocation type="take">
<performative type="inform" certainty="certain">
<belief-relation type="gen-spec">**This is** </belief-relation>
a spasm of <deictic obj="chest"> **chest**</deictic></perfomative>
<belief-relation type="cause-effect"> **resulting from**

<performative type="inform" certainty="certain">
overexertion when heart is diseased.</performative></belief-relation>
</turn-allocation></APML>

User: Which is the therapy?

Greta:

<APML><performative type="suggest">
<belief-relation type="cause-effect"> **To solve**
<deictic obj="patient">**your**</deictic> **problem,** </belief-relation>
you should take two drugs.</performative>
<performative type="inform">
<topic-comment type="comment"> **The first one** </topic-comment>
is Aspirin <belief-relation type="sequence">**and**</belief-relation>
<topic-comment type="comment"> **the second one** </topic-comment>
is Atenolol </performative></APML>

… and so on.

Annotations are translated into facial expressions and voice intonations. Figure 3 shows some examples of these expressions (neutral, sorry-for and surprise).

Figure 3. Three expressions of Greta: neutral, sorry-for and surprise

11. CONCLUSIONS

In this paper, we have overviewed some factors determining the communicative behavior of a Human Agent, that should be simulated in a Believable Embodied Agent to make it 'natural' (or 'believable'). In particular, we have considered contents to communicate, emotions, personality, culture, style and contextual variables. Then, we have briefly described the architecture of a conversational agent embodied in a 3D face, Greta, that tries to achieve a believable behavior while interacting with the User.

Our Agent follows the seminal work of other a number of research groups involved in this domain (to mention only some of them: Arafa et al, 1999;

Cassell et al, 1999; Rickel and Johnson, 1999; André et al, 2000; Lester et al, 2000). It builds upon research on how to model emotion activation and regulation (Frjida and Swagerman, 1987; Ortony et al, 1988; Elliott, 1992; Castelfranchi, 2000). The main characteristic of our Conversational Agent is in its ability to dialog with the User in any application domain whose knowledge has been represented with the appropriate formalism that is required by TRINDI and by MIND. Another peculiar feature is in the Mind-Body independence: although the examples we showed in this contribution are all about Greta, the APML tagging of dialog moves enables us to render them with any animated agent (2D or cartoons, for instance) according to the needs of the application domain and of the particular category of users to which the system is addressed. In the future, we plan to study how the variables considered (in particular, affective factors) influence not only the external behavior of the Agent but also the inner part of it: typically, its reasoning style, its decision making ability and the argumentation process.

REFERENCES

André, E.., Rist, T., Van Mulken, S., Klesen, M., & Baldes, S. (2000). The automated design of believable dialogues for animated presentation teams. In J.Cassell, J.Sullinvan, S. Prevost & E.Churchill (eds.). Embodied Conversational Agents. Cambridge, MA: MIT press.

Arafa, J., Charlton, P., Mamdani, A., & Fehin, P. (1999). Designing and building Personal Service Assistants with personality. In S. Prevost & E. Churchill (Eds.). Proceedings of the Workshop on Embodied Conversational Characters, Tahoe City.

Bates, J. (1994). The role of emotion in believable agents. Communications of the ACM, 37, 7, 122-125.

Black, A.W., Taylor, P., Caley, R. & Clark, R. (2002) Festival.. http://www.cstr.ed.ac.uk/projects/festival/

Carbonell, J.G. (1980). Towards a process model of human personality traits. Artificial Intelligence, 15, 49-74.

Carofiglio, V., de Rosis, F. & Grassano, R. (in press). Dynamic models of mixed emotion activation. In L Canamero & R.Aylett (eds.): Animating Expressive Characters for Social Interactions. Amsterdam: John Benjamins.

Cassell, J., Bickmore, J., Billinghurst, M., Campbell, L., Chang, K., Vilhjalmsson, H. & Yan, H. (1999). Embodiment in conversational interfaces: Rea. In Proceedings of CHI'99, Pittsburgh, PA, pp 520-527.

Cassell, J., Sullivan, J., Prevost, S. & Churchill, E. (eds.) (2000). Embodied Conversational Agents. Cambridge, MA: MIT press.

Castelfranchi, C. (2000). Affective Appraisal Versus Cognitive Evaluation in Social Emotions and Interactions. In A. Paiva (Ed.). Affective Interactions. Springer LNAI 1814, Berlin. pp. 76-106.

Conte, R. & Castelfranchi, C. (1995). Cognitive and Social Action. London : University College.

De Carolis, B., Carofiglio, V. & Pelachaud C. (2002). From Discourse Plans to Believable Behaviour Generation. Proceedings of the International National Language Generation Conference. New York.

De Carolis, B., Pelachaud, C., Poggi, I. & de Rosis, F. (2001). Behavior planning for a reflexive agent. In IJCAI'01. Seattle, August 2001.

de Rosis, F., Pelachaud, C. & Poggi, I. (in press a). Transcultural believability in embodied agents: a matter of consistent adaptation. In R.Trappl & S.Payr (eds.) Agent Culture: Designing virtual characters for a multi-cultural world. Dordrecht: Kluwer Academic Publishers.

de Rosis, F., Pelachaud, C., Poggi, I., Carofiglio, V. & De Carolis, B. (in press b) From Greta's mind to her face: modeling the dynamics of affective states in a Conversational Embodied Agent. International Journal of Human-Computer Studies.

Ekman, P. (1982). Emotion in the human face. Cambridge: Cambridge University Press.

Elliott, C. (1992). An Affective Reasoner: A process model of emotions in a multiagent system. Technical Report No. 32 of The Institute for the Learning Sciences, Northwestern University.

Frjida, N.H. & Swagerman, J. (1987). Can computers feel? Theory and design of an emotional system. Cognition and Emotion. 1(3), pp. 235-257.

Hartmann, B. Mancini, M. & Pelachaud, C. (2002). Formational parameters and adaptive prototype instantiation for MPEG-4 compliant gesture synthesis. Proceedings of Computer Animation '02, Geneva: IEEE Computer Society Press.

Kita, S. & Essegbey, J. (2001): Pointing left in Ghana: How a taboo on the use of the left hand influences gestural practice. Gesture 1, 1, 72-95.

Lester, J.C., Stuart, S.G., Callaway, C.B., Voerman, J.L., & Fitzgerald, P.J. (2000). Deictic and emotive communication in animated pedagogical agents. In J.Cassell, J. Sullivan, S. Prevost, & E. Churchill (eds.). Embodied Conversational Agents. Cambridge, MA: MIT press.

Loyall, A.B., & Bates, J. (1997). Personality-rich believable agents that use language. Proceedings of Autonomous Agents 1997, pp. 106-113.

Magno Caldognetto, E. & Poggi, I. (1995) Creative Iconic Gestures: some Evidence from Aphasics. In R.Simone (ed.) Iconicity in Language. Amsterdam: John Benjamins.

Mc Crae, R. & John, O.P. (1992). An introduction to the Five-Factor model and its applications. Journal of Personality. 60, pp. 175-215.

Miceli, M. & Castelfranchi, C. (1989). A cognitive approach to values. Journal for the Theory of Social Behavior, 19, 169-193.

Ortony, A., Clore, G.L., & Collins, A. (1988). The Cognitive Structure of Emotions. Cambridge University Press.

Oatley, K. & Johnson-Laird, P.N. (1987). Towards a cognitive theory of emotions. Cognition and Emotion, 13, 29-50.

Pearl, J.(2000). Probabilistic reasoning in intelligent systems: Networks of plausible inference. Morgan Kaufman Publishers.

Pelachaud, C., Carofiglio, V., De Carolis, B., de Rosis, F. & Poggi, I. (2002). Embodied Contextual Agent in Information Delivering Application. In Proceeding of the First International Joint Conference on Autonomous Agents & Multi-Agent Systems. Bologna, Italy.

Pelachaud, C. & Poggi, I. (eds.) (2001). Multimodal Communication and Context in Embodied Agents. Workshop at the 5th International Conference on Autonomous Agents, Montreal.

Pelachaud, C. & Poggi, I. (in press). Subtleties of Facial Expressions in Embodied Agents. Journal of Visualization and Computer Animation.

Pelachaud, C. (2002) Visual Text-to-Speech. In I.S. Pandzic, R.Forchheimer (eds.) MPEG-4 Facial Animation - The standard, implementations and applications. New York: John Wiley & Sons.

Picard, R.W. (1997). Affective Computing. Cambridge, MA: MIT Press.

Plutchick, R. (1994). The Psychology and Biology of Emotions. New York: Harper & Collins.

Poggi, I. & Pelachaud, C. (2000). Performative Facial expression in Animated Faces. In J.Cassell, J.Sullinvan, S. Prevost & E.Churchill (eds.). Embodied Conversational Agents. Cambridge, MA: MIT press.

Poggi, I. (2002a). Mind Markers. In M.Rector, I.Poggi & N.Trigo (eds.) Gestures. Meaning and Use. Oporto: Universidad Fernando Pessoa Press.

Poggi, I. (2002b). Towards the Lexicon and Alphabet of Gesture, Gaze and Touch. www.semioticon.com/virtual

Prendinger, H. & Ishizuka, M. (2001). Smart brains in simplistic bodies. In C.Pelachaud & I.Poggi (eds.) Multimodal Communication and Context in Embodied Agents. Workshop at the 5th International Conference on Autonomous Agents, Montreal.

Rickel, J. & Johnson, W.L. (1999). Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. Applied Artificial Intelligence, 13, 343—382.

Russell, J.A., Levicka, M. & Niit, T. (1989). A cross-cultural study of circumplex model of affect. Journal of Personality and Social Psychology, 57, 848-856.

Ruttkay, Z. Pelachaud, C., Poggi, I. & Noot, H. (in press) Excercises of Style for Virtual Humans. In L.Canamero & R.Aylett (eds.) Animating Expressive Characters for Social Interactions. Amsterdam: John Benjamins.

Taylor, I.C., Mc Inness, F.R., Love, S., Foster, J C, & Jack, M.A. (1999). Providing animated characters with designated personality profiles. In S. Prevost & E. Churchill (eds.) Proceedings of the Workshop on Embodied Conversational Characters. Tahoe City.

Trappl, R. & Payr, S. (eds.) (in press). Agent Culture: Designing virtual characters for a multi-cultural world. Dordrecht: Kluwer Academic Publishers.