

User Modeling And Adaptation In Health Promotion Dialogs With An Animated Character

Fiorella de Rosis, Nicole Novielli, Valeria Carofiglio, Addolorata Cavalluzzi and Berardina De Carolis

Department of Informatics, University of Bari

<http://www.di.uniba.it/intint/>

{derosis, novielli, carofiglio, cavalluzzi, [decarolis](mailto:decarolis@di.uniba.it)}@di.uniba.it

Abstract

In this paper, we describe our experience with the design and implementation of an embodied conversational agent (ECA) that converses with users in order to change their dietary behavior. Our intent is to develop a system that dynamically models the agent and the user and adapts the agent's counseling dialog accordingly. Towards this end, we discuss our efforts to automatically determine the user's dietary behavior stage of change and attitude towards the agent on the basis of unconstrained typed text dialog, first with another person and then with an ECA controlled by an experimenter in a Wizard of Oz study. We describe how the results of these studies have been incorporated into an algorithm that combines the results from simple parsing rules together with contextual features using a Bayesian network in order to determine user stage and attitude automatically.

1. Introduction

Conversational systems can be employed in support of health care in limited domains given their potential for low-cost and wide accessibility. Such systems may be especially efficacious for patients and consumers who are used to using computer systems and the web to obtain health information. Conversational systems that counsel users on dietary behavior represent an especially promising application in this area. Fruit and vegetable consumption alone plays a protective role in a large number of cancers, and is associated with reduced risk for heart disease, stroke, and hypertension, yet only a small percentage of adults meet the government guidelines for daily fruit and vegetable consumption (Li, et al, 2000).

In behavior change counseling, expert counselors must be finely attuned to their patients' emotional state, including their attitude towards the counselor, and adapt their dialog accordingly. This *adaptivity* is one of the features which gives new conversational systems the potential to become, if not competitive, at least supportive of encounters with human therapists, when compared with previous seminal experiences like e.g. PARRY or GURU (Colby, 1981; Colby et al, 1990)¹.

Embodied Conversational Agents (ECAs) are a new metaphor of human-computer interaction which aims at providing the users with the illusion of cooperating with a human partner rather than just 'using a tool': their application to health promotion dialogs might therefore be of benefit to increase the usability of these systems (Cassell et al, 2000). The more ECAs become 'believable', with all the shades this term has

¹ A reflection on these experiences may be found in (Colby, 1999), with an interesting discussion of early works on computer conversations in the same Volume (Wilks, 1999).

acquired after the initial definition in (Bates, 1994), the more users can be expected to show some sign of ‘social relationship’ with them: in addition to understanding the user problems, agents should be equipped to perceive these signs and to respond appropriately.

In this paper, we describe our experience with the design and implementation of an ECA that converses with users in order to change their dietary behavior. Two theories guided us in designing this system: the Transtheoretical Model of health behavior change and theories on the role of affect in persuasion. Prochaska and Di Clemente’s Transtheoretical Model of health behavior change outlines a series of stages that people naturally go through when changing their health behavior, as well as the subset of behavior change techniques that are especially effective in each stage (Prochaska et al, 1992). In addition, several persuasion theories state that the emotional state of the persuadee—including their attitude towards the persuader—should be considered as a significant factor in selecting and applying persuasion strategies (Wegman, 1988; Sillince and Minors, 1991; Walton, 1992; Miceli et al, in press).

The conversational system we wanted to implement therefore had to be endowed with the following features:

- *dynamic user modeling*, to recognize the users state and revise this image during interaction;
- *dynamic agent modeling*, to represent the agent’s emotional reaction to what the user says;
- *double adaptation* of the dialog, to the user and the agent.

In the dialog system we were ultimately designing, agent utterances will be spoken by an ECA. However, to reproduce the typical interaction mode of information systems on the web, user utterances would be typed. Thus, a prerequisite for our system was the development of methods to determine users’ mental and affective state on the basis of typed text utterances within the context of a diet counseling dialog.

As many researchers have pointed out, affect recognition requires integration of cues from multiple modalities in order to achieve an acceptable level of accuracy: see (Picard, 1997 and 2002) and the website of HUMAINE, the European Human-Machine Interaction Network of Excellence². Uncertainty is an unavoidable factor of any emotion recognition process. To address this problem, we represent the user model with an appropriate formalism (Bayesian networks) in which knowledge about the context in which the user utterance was made can be integrated to clarify recognition and to increase the predictive value of the model.

The following are the main design, implementation and revision steps we followed in our research:

1. *definition of a theoretical background* for our research (Section 2 of the paper);
2. *development of a first prototype* (Section 3) after analysis of a corpus of ‘natural’ dialogs to identify signs of ‘stage of change’ and the emotional state of the user, and adapt the dialog strategies accordingly;
3. *informal evaluation of this prototype*, (Section 4) with reflection on the difference between affective problems related to interacting with a conversational character vs. a therapist;

² <http://emotion-research.net>

4. *collection of a corpus of dialogs with an ECA* by means of a Wizard of Oz study (Section 5);
5. *design of a dynamic user modeling component* (Section 6) to integrate the dialog system with the ability to deal with ‘social relationship’ factors.

We will discuss the interest and limits of our work in Section 7 by reflecting, in particular, on the role that embodied agents might play in this application domain.

2. Motivation

Health promotion is considered a promising application domain and, at the same time, a fertile testing arena for computer-simulated dialogs in general and for socially intelligent animated characters in particular. Successful experiences in this application domain, together with a body of established theories, were the background on which we built our experience.

2.1 Previous experiences

As far as we know, Daphne (Grasso et al, 2000) was one of the first systems in which the problem of promoting health nutrition was considered. Although the project was aimed at developing a theory of ‘informal argumentation’ rather than implementing a conversational agent in the domain, the system adapted its argumentation strategy to the respondent’s preferences rather than attempting to solve conflicts due to differences in opinions between the two participants to the dialog. More focused on ECAs, Silverman and colleagues (2001) combined a generic simulation package with animated pedagogical agents to promote health behavior shifts in heart attacks. Bickmore employed Laura in FitTrack, by endowing this character with the ability to establish a *working alliance*, an essential prerequisite of any successful therapist-client relationship. Trust and empathy were considered, in particular, as key factors of the bond relationship between the two participants to the dialog, to be achieved by means of adequate dialog strategies (Bickmore, et al, 2005). By extending their previous experience with Carmen’s Bright IDEAS (Marsella et al, 2003), DESIA integrated an embodied pedagogical agent into a psychosocial intervention deployed on a handheld computer, to help mothers of pediatric cancer patients to cope with problems encountered in care giving (Johnson et al, 2004). As in the case of FitTrack, this system was designed to work on long-term interaction. It adapts the choice of presentation modes to the context of use and keeps track of information about past encounters to tailor its explanations. The common denominator of these applications is in considering forms of social intelligence like encouragement, display of empathy, promotion of ‘positive face’ and developing of a rapport as key factors in motivating the users in their application domain.

2.2 Theoretical Background

a. affective state

Affective factors may include long-term personality traits or shorter-term states ranging from ‘affect dispositions’, ‘attitudes’ (liking, loving, hating,...), ‘interpersonal stances’ (distant, cold, warm,...), ‘moods’ (cheerful, irritable, depressed,...) or ‘real emotions’ (K. Scherer, Glossary in the HUMAINE website). Emotions have been defined, in particular, according to at least two views: as points in a two-dimensional space of valence and arousal (according to the ‘Circumplex Model of Emotions’: Plutchik and Conte, 1996) or as a set of ‘basic emotions’ (Ekman, 1992), possibly classified according to their activation factors (Ortony, Clore and Collins, 1988). Other authors proposed a categorization into ‘individual’ emotions, referring to self (fear, hope, joy etc) and ‘social emotions’, which originate from relationships with others: sympathy, antipathy, tenderness, sense of friendship etc (Poggi, 2004). The second category is tightly related to Scherer’s concept of interpersonal stance. Both individual and social emotions are expected to occur in client-therapist dialogs, with a relative frequency that depends on the client’s problems discussed.

b. transtheoretical model of change

Interventions to promote a health-behavior change often adopt the Transtheoretical Model of Change (Prochaska et al, 1992). According to this theory, health promotion plans should be adapted to the degree of advancement of the process followed by subjects in changing their beliefs and attitudes, that is their ‘stage of change’. Expert therapists apply this theory by selecting intervention techniques that are known to be particularly effective for the stage of change their client is currently in. Some dialog systems try to simulate this situation: in the most famous of them, which is due to Bickmore (Laura), dialogs alternate between a listening and a persuading phase. The encounters of the client with the artificial agent are repeated at fixed time intervals by adapting, every time, to the evolved stage of the subject.

The Stage of Change (SoC) Model identifies five main steps in changing behavior :

- in the *pre-contemplation* stage, subjects believe that their behavior is acceptable and do not want to change it;
- in the *contemplation* stage, they doubt that their behavior is acceptable, seriously consider the opportunity of changing it but do not want to commit to do it soon;
- in the *preparation* stage, they believe that their behavior should be changed and intend to do it soon;
- in the *action* stage, they are following a plan to change their behavior (for some months);
- in the *maintenance* stage, they are maintaining the change for more than 6 months.

These definitions suggest how stages may be recognized from a set of ‘signs’ which are related to the mental state of the subjects:

- *belief that their behavior is ‘right’ or ‘wrong’, value attributed to the ‘right’ behavior and knowledge of reasons influencing the adoption of a problem behavior;*
- *intention to change their own behavior if wrong;*
- *belief that (internal and external) conditions exist to change this behavior;*
- *knowledge of an acceptable plan which enables achieving this intention;*
- *level of perseverance in following the plan.*

The actions which the therapist may apply at every ‘stage of change’ to promote a correct behavior respond to the following *goals*: (i) to recognize the situation, (ii) to inform and encourage about the evaluation processes rather than enforcing persuasion, (iii) to influence intentions, (iv) to check abilities, (v) to suggest plans and (vi) to offer support during plan execution. The theory may therefore be adopted as a powerful source of knowledge to build cognitive models of the users’ attitude towards the problem behavior and to decide how to tailor advice-giving to their stage of change.

As suggested in (Velicer et al,1998), stage of change and emotional state are strongly interrelated. For instance, in the *pre-contemplation* stage subjects may be *demoralized* about their ability to change if they have tried to do it and failed; in the *maintenance* stage, people are increasingly more *confident* that they can progress in this change, and so on. Recognizing some aspects of the emotional state (valence and intensity, for instance) may therefore contribute to inferring the stage of change and the inverse; at the same time, it may influence the choice of an appropriate persuasion strategy. On the other hand, recognizing the attitude of users towards the virtual therapist enables adapting other aspects of the dialog: level of familiarity of the style employed, introduction of small talk and similarities.

Accurate measurement of the stage of change requires acquiring some information on the clients’ behavior and their mental state: accurate collection of this information is a critical step for applying the model successfully. In health services, the stage of change is usually estimated with a questionnaire which is supplied to the patient at the beginning of interaction (e.g., LaForge, et al, 1994). Applying this procedure to a conversation with an Embodied Agent risks negatively influencing the attitude of the user. In addition, if the counseling dialog is successful, the mental state of the user may change during the interaction and the system will have to dynamically adapt to this situation. Our solution to this problem is to dynamically infer users’ stage of change, emotional state and attitude towards the agent on the basis of their dialog behavior during the interaction.

3. Our first prototype of dialog simulator

Dialog simulation systems are usually built after careful analysis of a corpus of naturally occurring dialogs in order to emulate the behavior humans show when communicating among themselves. In health promotion, the ideal source of corpora is that of transcripts of conversations between patients and specialists: these, however, are generally not available for research due to privacy concerns. Instead, we focused our initial analyses on published transcripts available in specialized journals or books dealing with various aspects of health promotion in a variety of behavioral domains, including smoking and alcohol abuse cessation and dietary change.

3.1 Method

a. The corpus

Our corpus included five dialogs which were published in specialized books and an email dialog with a computer scientist who played the role of a dietician (Grasso et al, 2000). The cases were patients with

problems of varying severity and at various stages of change. This corpus was quite heterogeneous, due to the variety of sources employed. In particular, the email dialog was midway between human-human and human-computer conversations. The subject believed that he was interacting with a human dietician but conducted the interaction via a computer-mediated communication channel. He included some emoticons in his text and some closing sentences which were more representative of a written communication than an oral communication style.

b. Corpus labelling

We extracted from the six dialogs the moves which included potential signs of either emotional state or stage of change: overall, 78 moves were selected for analysis of the emotional state and 115 for stage of change. We then defined a markup language with which to label the selected moves and asked ten raters to label them with this language.

Table 1: Markup language for emotion and stage of change

Sign		Definition	Values	Examples (with shared interpretation)
Emotional State	Valence	whether the state is perceived as pleasant to the individual in that state.	Positive / Negative / Unknown	Well, it does help to talk with someone. (Positive) I feel I'm being lectured rather than listened to! (Negative)
	Intensity	whether the manifestation of the state is strong	High / Low / Unknown	When the doctor told me I could never work again, I was very depressed. (high) Well, it does help to talk to someone. (low)
Believes Behavior Wrong	Behavior	Whether the subject is aware that his/her behavior is "wrong"	Yes/No/Maybe	Yes, I know it's bad for me. (Yes) Well, I'm not really sure if it's a problem at all. (No)
Intends to Change		Whether the subject wants to change behavior	Yes/No/Maybe	Well, I want to do something. I don't want to just let this go on (Yes). Well, it's like this: I'd like to give up, but it is just too much for me at the moment. (No)
Knows About a Plan		Whether the subject knows about a plan to follow in order to change behavior or, in general, if he/she has enough information about how to change	Yes/No/Maybe	I think I should cut down on caffeine (Yes)
Accepts plan		Whether the subject accepts the plan proposed by the therapist	Yes/No/Maybe	That seems best (Yes)
Is following a plan		Whether the subject is already following a plan.	Yes/No/Maybe	I changed my diet in 1992 (Yes).

The emotional tags were valence, intensity and emotion name; SoC tags were the mental state components which we called 'signs' in Section 2. Table 1 shows the definitions of these tags with some examples of labeled sentences for each of them: these examples show that multiple labeling was requested. For instance, the sentence: *"Well, it does help to talk with someone"* shows a positive emotional valence of low intensity. Therefore, when tagging a move for the emotional state, raters were requested to indicate the value of valence and intensity and (if recognizable!) the emotion name. This apparent redundancy was motivated by our belief that emotions are difficult to recognize, while valence and intensity are recognized more easily. The names and descriptions of emotions were drawn from the OCC classification (Ortony et al, 1988) with a very few additions: demoralization, frustration, disappointment, irritation. In defining the emotional tags, we

adopted effect-type descriptions (Cowie, 2000) which refer to the effect that emotional characteristics of speech have on the listener. Raters were asked to label the sentences according to what the language (style, syntax, lexicon, etc.) suggested to them. For all labels (both emotional and related to the stage of change), we introduced a categorical scale which enabled the raters to distinguish between three grades of the feature of interest; this has proven to be preferable for subtle phenomena, such as emotions or mental state components (Craggs and McGee Woods, 2004).

c. Measures of agreement among raters

Various measures of agreement among raters in labeling corpora have been proposed. In a paper aimed at describing the role of agreement measures, Craggs and McGee Wood (2004) discuss the agreement statistics classification of Di Eugenio and Glass (2004) by examining their advantages and limits. The class of ‘percentage agreement’ statistics (which measures the proportion of agreement among raters) has the advantage of not suffering from unequal distribution of the labels used by raters. However, it excludes any notion of the level of agreement one could expect to achieve by chance, without which any deviation from perfect agreement is not interpretable. Chance-corrected measures compute agreement by considering both the observed values and those we could expect by chance; these measures may or may not assume an equal distribution of categories between coders. The most common measures applied in computational linguistics belong to the latter category (Kappa and Alpha, by Krippendorff). In particular:

$$\text{Kappa} = (p(a) - p(e)) / 1 - p(e)$$

where $p(a)$ denotes the observed agreement rate and $p(e)$ the expected rate.

Some authors (again, Di Eugenio and Glass, 2004) propose to use the two categories of measures as a means to judge agreement from several viewpoints; others believe that this may be seen as a lack of confidence in each of them. In our opinion, if agreement measures are applied in order to assess the difficulty to recognize a given feature from text, combining percentage agreement statistics with chance corrected measures is worthwhile. The first measure is immediately interpretable, while the second measure enables comparing features with different frequencies in the corpus (we will see some examples in the results). For this reason, we combined the kappa statistics with the following percentage agreement measures:

- *full agreement rate*, as the proportion of moves in which the raters ‘fully agree’ on the labeling of the move: for instance, two raters are said to fully agree on labeling the emotion valence of a sentence if they both label it as ‘positive’, ‘negative’ or ‘unknown’, and in the case of its intensity, if they both label it as ‘high’, ‘low’ or ‘unknown’;
- *weak agreement rate*, as the proportion of moves in which the raters ‘weakly agree’ in the labeling of the move. A ‘weak’ agreement on a tag is defined as a case in which the values assigned to the tag by the two raters are not the same but are not opposite either: for example, ‘positive’ vs. ‘neutral’ or ‘negative’ vs. ‘neutral’ valence;

- *disagreement rate* as the proportion of moves in which the raters ‘strongly disagree’ in the labeling of the move: for example, ‘negative’ vs. ‘positive’ valence, or ‘high’ vs. ‘low’ intensity.

d. Move classification

To summarize the results of tagging by our ten raters, we classified the moves into three broad categories:

- a) moves with a *shared interpretation*, for which there was a rate of full agreement in more than 60% of raters,
- b) moves with a *likely interpretation*, for which the rate of full agreement among raters ranged from 40 to 60% and
- c) moves with a *questionable interpretation*, for which the full agreement was less than 40 %.

3.2 Results

Table 2 summarizes the agreement among raters for the first three signs in table 1: valence, intensity and believes behavior wrong. Emotion names varied considerably among the raters; for instance, some moves were interpreted, by different raters, as indicative of ‘reproach’, ‘anger’, ‘frustration’ or ‘disappointment’; while others as indicative of ‘disliking’ or ‘reproach’; for this reasons, we do not include emotion names in this analysis.

Every dialog included moves which were classified as belonging to the pre-contemplation, contemplation or preparation stages, indicating that the subject’s stage of change varied during the dialog, probably due to the interaction with the counselor. We only show results regarding the ‘believes behavior wrong’ tag, as very few moves denoting the other components were found in the corpus; these sentences received a high full agreement rate.

Table 2: Agreement rate and frequency for emotion and stage of change tags

	Frequency (in the subset of labelled moves)	Full agreement	Weak agreement	Disagreement	Kappa
Valence	83%	.55	.02	.43	.27
Intensity	65%	.45	.20	.34	.17
Believes behavior wrong	17%	.80	.04	.11	.26

The Table shows that valence was identified a bit more easily than intensity, as the full agreement rate is higher for this tag (.55 vs. .45). The majority of moves with ‘shared’ interpretation were those labeled with a ‘negative’ valence, while the majority of moves with ‘likely’ interpretation were labeled with an ‘unknown’ valence. This may due to different reasons: first, as we said, five dialogs referred to situations involving serious behavior problems of alcohol abuse or smoking; these were likely to induce negative emotions. In the email dialog regarding dietary behavior, the problems discussed were less severe: however, in this dialog the subject showed a negative emotion (*irritation*) due to the overly-persuasive behavior of the counselor. Moves

tagged as indicative of a ‘high’ emotional intensity were more frequent than those tagged with a ‘low’ value, possibly because they were easier to recognize or for reasons similar to those mentioned for valence: health-related dialogs deal with problems which deeply involve subjects since they involve discussion of their health, family, job, etc. Therefore, it seems reasonable that becoming aware of a situation which is (potentially or actually) negative generates a negative emotional reaction of more or less high intensity, according to how serious the discussed problem is.

3.3 Prototype design

a. user modeling

The results of the corpus analysis described above guided our design of the user modeling module of our dialog simulation prototype. This module parses user utterances and propagates the results of this parsing, as ‘observed variables’, in a dynamic Bayesian network which infers the stage of change and the emotional valence of the user with some level of uncertainty. In Section 6 we will describe with more detail the method behind user model building and updating. In another paper (Carofiglio et al, in press), we describe how the emotional impact of the user move on the agent may be simulated, again with some uncertainty. In the next subsection, we will describe how the two models are employed to adapt the dialog plans and style.

b. dialog adaptation

As we anticipated in Section 2, the Transtheoretical Model of Change suggests specific plans to apply at every ‘stage of change’ to help the subjects in changing their behavior. For instance, in the pre-contemplation stage the *plan* includes the following steps: i) validate lack of readiness, ii) clarify, decision is yours, iii) encourage evaluation of pros and cons of the behavior change and iv) identify and promote new positive outcome expectations.

To apply this model, our agent requires some information about the user: it may employ uncertain default information in the first dialog steps, provided that this approximate picture of the user is subsequently refined so as to also refine the advice provided. Therefore our dialog simulator needs on one hand a knowledge updating system which deals with uncertainty in knowledge about the user and, on the other hand, a description of the current situation and the dialog history on which to base its planning activity. This description of the problem orients the choice of the dialog management system towards an *information state* model. This model was developed as part of the TRINDI EC Project to enable implementing flexible dialog simulation systems with a plan-based approach (Traum and Larsson, 2003). The information state (IS) is a blackboard on which data needed to develop the dialog are represented with a logical formalism and are revised dynamically by means of *IS update rules*. In our case, the IS structure includes a model of the agent and a model of the user with two main components:

- *permanent characteristics* (in the ‘STABLE’ part) which do not change in the course of the dialog: for instance, ‘name’, ‘age’, ‘personality’, background;

- *transitory characteristics* (in the 'UNSTABLE' part) which are revised during the dialog: the user's affective state and her stage of change.

A set of *updating rules* is used to update the 'unstable' part of the user model following the interpretation of every user dialog move.

The agent's goal can be achieved by means of one or more *plans*, each characterized by a set of *applicability conditions* on the user's mental state: a set of *select rules* establishes the next plan and move to perform. A plan includes a *main part* (the essential of the plan) and a *secondary part* with optional details. It may be linked to another plan by a *causal link*, to represent that the execution of at least its main part should precede the execution of that plan. Introduction of precedence relationships and distinction between the main and secondary parts was our solution to the problem of uncertainty in the knowledge of the user state. The system provides the user with the opportunity to react to its moves, by fragmenting its plans into short moves and by interleaving suggestions with questions. The user may answer these questions or input new comments; both kinds of moves enable the system to reason on her state and on the emotional effect produced by its move. Since interpretation of user moves is highly error prone, dialog plans are selected by default but may be revised as new information becomes available. When the user model is revised, an on-going plan may be interrupted to be substituted with a new one. In this case, priority and causal relationships insure that a 'believable' dialog flow is maintained after plans are changed. This distinction between the main and secondary parts ensures that a goal can be achieved with plans at different levels of detail, depending on the situation and the dialog history.

c. An example

Let us describe the dynamics of interpretation and planning, in an example dialog which was generated with our prototype: we will show how the system's image of the user and the dialog plans evolve. The example starts after the user ('U') declared that she feels her eating behavior is not as she would like it to be:

```
=====
...
  User's stage of change: Contemplation (p=0.26)
  Agent plan: Clarify: decision is yours
S3: I will try to help you, but cannot make decisions for you.
U3: I would like to change but it's so difficult!
  Move interpretation: (Say U 'change is difficult')
  Agent emotion: light sorry-for
  User's stage of change: Contemplation (p=0.24)
  Valence: negative
  Agent plan: Display empathy; Encourage the evaluation of the pros and cons of behavior change
S4: I know, a diet is difficult to follow, particularly if you eat out because of work. But the effects may be very positive!
U4: I know! I definitely intend to change my habits.
  Move interpretation: (Say U 'intends to change')
  Agent emotion: light hope
  User's stage of change: Preparation (p=0.24)
  Valence: positive
  Agent plan: Display satisfaction; Encourage initial steps
```

S5: Good! Why, for instance, don't you try to avoid going to restaurants for your daily lunch?

...

=====

The user is initially assumed to be in the 'contemplation' stage, with a probability above the threshold (that we fixed as .20)³. The agent adopts the plan 'Clarify, decision is yours' which matches the goal activated by this state. After move U3, the agent takes on the emotion 'sorry-for' which activates a plan of 'empathy display': "*I know, a diet is difficult to follow...*". The inferred 'stage of change' does not differ from the previous one (its probability only slightly decreases) and the inferred valence is negative. Hence, a plan to 'encourage evaluation of pros and cons of behavior change' is carried out. At move 4, the user manifests an intention to change her habits and a positive valence. The user model is revised (the most likely 'stage of change' is now 'preparation'), the agent takes on the emotion of 'hope' and a plan to 'Display satisfaction and encourage initial step' is applied.

The *agent move* is produced by an animated agent using a standardized API that consists of an APML string (De Carolis et al, 2003) interpreted by a 'wrapper' to the particular animated agent in use (de Rosis et al, 2003) . This API allows several different animated agents to be used, including GRETA (Pelachaud and Bilvi, 2003), MS-Agent or Haptek⁴. A *graphical interface* enables the users to interact with the system and is responsible for scheduling the various functions and activating the related modules. More details about the dialog manager may be found in (de Rosis et al, 2003) and in (Cavalluzzi et al, 2004).

4. Is interacting with an ECA the same as interacting with a human counselor?

This is a provocative question which we asked ourselves after an internal evaluation of our first prototype. Although, as we said in Section 1, the idea behind conversational agents is to endow the system with the ability to emulate human behavior, we felt that we could not assume that the subjects' behavior when interacting with our character would be the same as would be adopted when interacting with a human counselor. Therefore, we decided that the knowledge we had acquired from analysis of the natural corpus had to be integrated with some theoretical background and empirical data concerning the nature of human-agent interaction. In this second step of our research, we were looking for answers to questions such as: *What kind of relationship do users establish with an ECA when discussing their health-related problems? How can the nature of this relationship be recognized? How do users expect the ECA to respond to their manifestations of social relationship?*

A number of evaluation studies have been published, which describe how users see ECAs and how their vision is influenced by variations in the agent characteristics: see (Nass et al, 2000; Berry et al, 2005; Ruttkay and Pelachaud, 2004) for a survey of more recent results. The majority of these studies only involved agent monologs and therefore do not contribute much to the understanding of the exact nature of

³ to avoid unstable changing of plans, if the probability of more than one stage is larger than the threshold and these values are similar, the last move's stage is preferred.

⁴ <http://www.haptek.com>

the human-ECA relationship. In the famous *media equation*, the Stanford group formulated the hypothesis that social science theories may be applied in this domain (Nass et al, 2000). Recently, however, the need to specify the applicability conditions of this hypothesis was raised. Some studies demonstrated that there are some significant differences between human-human and human-agent interactions. In some situations, people tend to use ‘computer talk’ when speaking to a computer agent (Batliner, 2003), manifested by behavior such as shortened move length and adaptation of their speech level and tone to the agent’s speech characteristics (Oviatt and Adams, 2000; Darves and Oviatt, 2002; Coulston et al, 2002). These changes in style may be due to a simple style-induction effect or may be an index of lack of trust in the computer’s ability to recognize and interpret the user input. But they might also be explained in terms of a more complex theory of the social relationship humans tend to establish with technology in general, and with agents in particular.

Several psychosocial theories have been considered by computer scientists engaged in research about ECAs. In designing the interaction attitude of REA, Cassell and Bickmore (2003) referred to Svennevig’s model of ‘interpersonal relations in conversations’. They identified the dimensions of social relationship (familiarity, power, solidarity and affect) and extended these dimensions with the concept of trust. Other authors included politeness, deception and irony among the ingredients of social relationship. Terms like ‘empathy’ or ‘friendship’ have also been employed to denote key aspects of this relationship (Paiva et al, 2004). To Poggi (2004), *empathy* is the ability to identify with and understand another’s situation, feelings and motives: it requires listening skills and emotional intelligence and may occur even in absence of any emotional expression by the interlocutor. To Vaknin⁵, this concept goes beyond pure emotion transmission as “*The empathor empathizes not only with the empathee’s emotions but also with his physical state and other parameters of existence*”.

We will use the term *social attitude* to denote the kind of relationship users establish with our ECA. We denote, with this term, ‘*the process of entering into a warm social relationship with someone else, of being somehow involved in her goals and feelings*’. Although this is a shorter-term kind of relation than friendship, it shares with this concept the characteristics of intimacy, affection and mutual assistance; it is influenced by interpersonal attraction but also by rewards, which should outweigh costs such as irritation or disappointment. In advice-giving dialogs, rewards are affected by what subjects expect to receive from the interaction: information and, in some cases, entertainment. Therefore, the social attitude of users towards the agent is probably affected by their degree of satisfaction with the information received and by how pleasant they found the interaction. A positive social attitude may be displayed through an increase of *intimacy* and *common ground* over the course of the conversation, a decrease of *interpersonal distance*, the use of *non explicit ways* of achieving conversational goals and the display of *expertise* (Cassell and Bickmore, 2003). Humorous acts may also be taken as an offer of sympathy: “When the participants are in the mood for jokes, joke telling occurs naturally and there is some meta-level cooperation” (Nijholt 2004). These were therefore the signs of social attitude we expected to find in human-ECA dialogs.

⁵ <http://samvak.tripod.com/empathy.html>

5. A corpus of Wizard-Of-Oz dialogs

To have some insight into the kind of relationships users might establish with our advice-giving ECA, as well as how this relationship may be recognized, we performed a Wizard of Oz (WOZ) study. Subjects interacted with the ECA to discuss their eating habits and receive information and suggestions about problems in their eating behavior that were possibly discovered during the conversation. As in all WOZ studies, subjects believed that an automated system was generating the ECA's answers, while in actuality these were selected by a human confederate ("wizard") from a set of precompiled moves (Dahlback et al, 1993). This method is usually considered to produce results that are plausible simulations of real dialogs between an automated, virtual dietary expert and a client.

5.1 Method

a. the corpus

We employed an experimental setup which enabled us to perform studies under diverse conditions by varying the agent's physical aspect, its expressivity, the dialog moves, the evaluation questionnaire and other factors (Cavalluzzi et al, 2005). Data of various kinds may be collected with this tool: the subjects may be asked to evaluate the agent's behavior with a questionnaire and dialogs may be recorded to be later analyzed using quantitative and qualitative methods. The tool was employed in an 'iterative design' mode, with data collection steps ending with the design of the next one: at every iteration, the moves that the agent could employ were revised according to the problems found in the previous iteration.

Subjects involved in the study completed a pre-test questionnaire which was aimed at assessing their level of knowledge, habits and interest for healthy eating, in addition to their cultural background. To insure the uniformity of the experimental conditions throughout the whole study, we established some rules the wizard was requested to follow. After every subject move, the wizard selected her next move so as to respect a well-defined dialog plan and at the same time to insure the internal coherence in every dialog. This was achieved by a careful preliminary training of the wizard and by employing the same wizard with all the subjects. We employed an ECA that used an animated humanoid head built with Haptex's toolkit, with a rather realistic and pleasant aspect (figure 1) and with two kinds of voices: a mechanical and not very natural one produced with the Microsoft Speech API, and a much more natural voice produced with Loquendo⁶. The three clickable icons on the right side of the agent enabled the subject to evaluate whether the agent move was 'good', 'bad' or 'unclear' without interrupting the natural course of the dialog. At the same time, subjects could respond to the agent by typing any text in the textfield at the bottom of the window.

At the end of the experiment, a final questionnaire was displayed on the same computer monitor on which the agent had been displayed, to collect the subject's evaluation of several features of the message and the agent, each with a Likert scale from 1 to 6: how *credible*, *plausible*, *clear*, *useful* and *persuasive* was the

⁶ <http://www.loquendo.com>

message and how *sincere*, *likable*, *natural*, *intelligent* and *competent* was the agent. Dialogs were stored in a log at the end of the interaction for subsequent analysis.

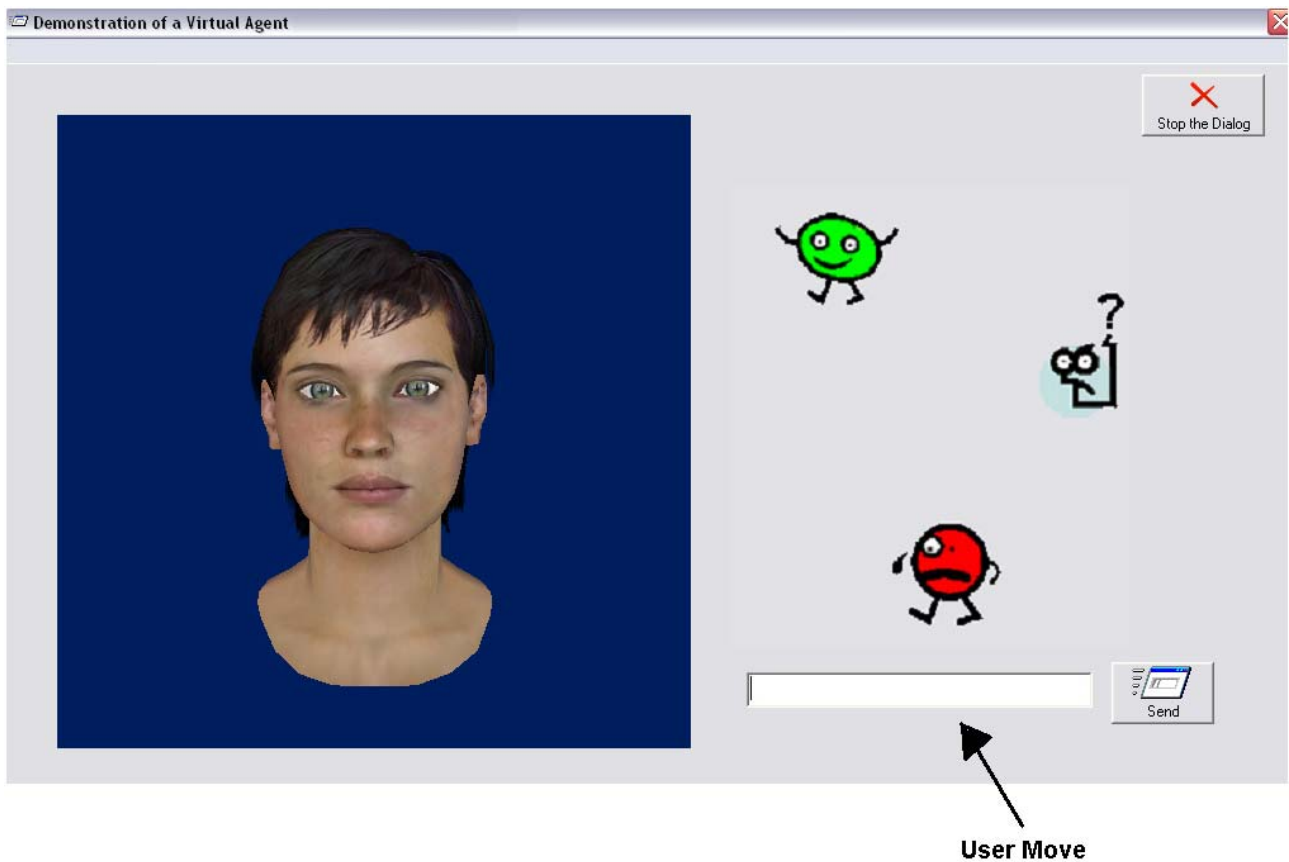


Figure 1: the character employed in our Wizard of Oz studies

b. method of analysis

We defined two measures of the subject's attitude during the dialog:

- *level of involvement*, as a function of the dialog duration (in number of adjacency pairs⁷) and the average length of the user moves (in number of characters), and
- *degree of initiative*, as a function of the percentage of questions raised by the subject over all the dialog moves.

These measures were integrated with a set of 'signs' that we defined after comparing the signs mentioned in the literature (see end of Section 4) with a preliminary analysis of the subject moves. These signs were aimed at estimating the social attitude of the subjects towards the agent and at assessing its relationship with their subjective evaluation of the agent (from the questionnaire), their level of involvement and their degree of initiative.

⁷ An *adjacency pair* is a couple of adjacent wizard-subject moves in the dialog.

c. corpus labelling

We extracted two subsets from the corpus of 712 moves:

- 237 moves to label for ‘social attitude’ with the language described in Table 3: this table shows the language features which we considered as signs of this attitude. For each, we provide an example of an adjacency pair which is translated from Italian: some pairs belong to several classes;
- 143 moves to label for signs of ‘stage of change’ with the same criteria that were employed in the first study (see Table 2). However, emotion intensity was excluded from this language because we did not notice any case of ‘strong’ emotions in the corpus.

Table 3: Markup language and agreement among raters for signs of social attitude towards the agent

Sign with definition	Values	Example
<i>Friendly self-introduction</i> The subjects introduce themselves with a friendly attitude (e.g. by giving her name or by explaining the reasons why they are participating in the dialog)	Yes/No	Oz: Hi. My name is Valentina. I'm here to suggest how to improve your diet. S: Hi, my name is Isa and I'm curious to get some information about healthy eating
<i>Familiar style</i> Whether the subject employs a current language, dialectal forms, proverbs etc	Yes/No	Oz: Are you attracted by sweets? S: I'm crazy about them.
<i>Talks about self</i> Whether the subject provides more personal information about self than requested by the agent	Yes/No	Oz: Do you like sweets? Do you ever stop in front of the display window of a beautiful bakery? S: Very much! I'm greedy!
<i>Personal questions about or suggestions to the agent.</i> Whether the subjects tries to know something about the agent preferences, lifestyle etc, or to give suggestions rather than receiving them	Yes/No	Oz: What did you eat at lunch? S: Meet-stuffed peppers. How about you?
<i>Humor and irony</i> Whether the subjects make any kind of verbal joke in their move	Yes/No	Oz: I know we risk entering into private issues. But did you ever ask yourself what are the reasons for your eating habits? S: Unbridled life, with a light aversion towards healthy food.
<i>Comments on the dialog</i> Whether the subjects comments the agent behavior in the dialog: comments may be about the agent experience, its degree of domain knowledge, the length of its moves etc.	Positive	Oz: I'm sorry, I'm not much of an expert in this domain. S: OK: but try to get more informed, right?
	Negative	Oz: Good bye. S: What are you doing? You leave me this way? You are rude!!
<i>Friendly farewell</i> This may consist in using a friendly farewell form or in asking to carry-on the dialog.	Yes/No	Oz: Goodbye. It was really pleasant to interact with you. Come back when you wish. S: But I would like to chat a bit more with you.

As we wanted to apply the results of this analysis to an automatic recognition of these features from a user move given as input to a parser, individual user moves were presented to raters in random order. In addition, since we wanted to develop a recognition method that incorporated some discourse context, we had raters analyze adjacency pairs, including a user move and the agent move that preceded it. Due to the large number of moves in the two sets, only three independent raters were recruited in this case. As in the first study, two measures of inter-rater agreement were computed: a percentage agreement, in which *we considered the label of a move as ‘agreed’ when at least two raters gave it the same value*, and a chance-corrected Kappa statistics (Carletta et al, 1996).

5. 2. Results

We performed six WOZ tests, with five subjects in each with a total of 712 moves in the 30 collected dialogs. As we said, these tests were considered as steps of an iterative design of our ECA: therefore, in designing each step we considered the results of the previous ones to discover the main limits of the ECA and revise its behavior. After the first three tests, we were able to stabilize the agent moves and behavior. We also included in the study subjects with different backgrounds (humanities and computer science), to evaluate the role played by this factor. The pre-test questionnaire enabled us to verify that the six groups of subjects were comparable in their level of knowledge, habits and interest in healthy eating. They belonged to the same age group (23 to 26 years) and the groups were gender balanced.

There was a high variability among the subjects for the two measures of level of involvement and also for the subjective evaluation of the message and the agent. The dialog duration ranged from 9 to 60 moves and increased only slightly on average when the number of moves among which the wizard could choose her answers was increased. The average duration was of 22.4 moves when the wizard could select among 58 moves and 25.5 when available moves were increased to 78. The average length of moves ranged from 29 to 95 characters. The average rating of the five message features (credibility, plausibility etc) ranged from 1.5 to 5 and the rating of the agent features (sincerity, likeability etc) ranged from 1.75 to 4.7. A multiple regression analysis showed that the message rating was correlated negatively with the dialog duration, the average length of the moves and the percentage of user questions in a dialog. We cannot say whether this rating was correlated with any emotional feeling, because the subjects showed very few ‘individual’ low intensity emotions (disappointment, satisfaction, etc.) compared to our first study. This was probably due to the difference in problems declared by the two groups of subjects: serious cases of problem behavior in the natural dialogs, vs. problems of unhealthy dieting in the WOZ studies. We may, however, conclude that *the subjective evaluation of the message does not seem to be a good indicator of the level of involvement of the subject in the dialog as one might expect.*

The dialogs included several signs of ‘social’ emotions (sympathy, appreciation or irritation, disappointment) and social attitude. The percentage of moves with these signs was positively associated with the ratings in the initial questionnaire and the subject involvement, while it was negatively correlated with their level of initiative. The subjects’ background was the factor which most highly influenced their behavior: computer scientists conducted dialogs with fewer and shorter moves, a larger proportion of questions and a lower proportion of social moves when compared to subjects with a background in humanities. More detailed data on this quantitative analysis of our corpus may be found in (de Rosis et al, 2005) while, in this paper, we will focus on their qualitative aspect.

Table 4 shows that there was a good agreement rate and Kappa for ‘friendly self-introduction’, ‘talks about self’ and ‘friendly farewell’ and a reasonably good value for ‘personal questions about the agent’. ‘Irony’ and ‘comments’ had a high rate but a low Kappa while ‘familiar style’ had a very low rate and Kappa. As we said in section 3, the Kappa statistics is a function of both the observed and the expected chance of agreement; this last measure depends, in its turn, on the distribution of values taken by the variables. Given

an observed rate of agreement, the variables whose values were not uniformly distributed produced a lower Kappa value: this is the reason why signs with low frequency (like ‘irony’ and ‘comments’) have a low Kappa level even if the agreement among raters is good. The value of kappa for the ‘comment’ sign is unique because this was a three-valued variable (positive, negative or no comment).

Table 4: Agreement among raters for signs of social attitude towards the agent

Sign	Values	Frequency (subset of labelled moves)	Frequency (whole corpus)	Agreement rate	Kappa
<i>Friendly-self introduction</i>	Yes/No	5%	2%	.98	.87
<i>Familiar style</i>	Yes/No	85%	28%	.33	.16
<i>Talks about self</i>	Yes/No	57%	19%	.73	.64
<i>Personal questions about or suggestion to the agent</i>	Yes/No	38%	12%	.70	.56
<i>Humor and irony</i>	Yes/No	7%	2%	.84	.36
<i>Comments on the dialog</i>	Positive	13%	4%	.82	.42
	Negative	16%	5%	.86	
<i>Friendly farewell</i>	Yes/No	11%	4%	.93	.65

As shown in Table 5, agreement was quite good also for the two signs of stage of change. Although these data cannot be compared with the results of the first study (in Table 2) because the number of raters is not the same (ten in the first study, and three in the second), we can say that, in this second study, raters tended to ‘weakly agree’ on tagging ‘belief behavior wrong’ more frequently than in the first study: recognizing this sign therefore seems to be easier when more serious problems are discussed. As in the natural corpus, subjects involved in WOZ studies were mostly in the pre-contemplation, contemplation and preparation stages of change and some of them apparently changed stage during the dialog. This is not surprising, considering that the conversation dealt with simple and common dietary rules that are not particularly difficult to accept.

Table 5: Agreement among raters for emotion valence and stage of change, in the subset of selected moves

	Frequency (subset of labelled moves)	Frequency (whole corpus)	Full agreement	Weak agreement	Disagreement	Kappa
Believes behavior wrong	48%	10%	.65	.26	.09	.46
Intends to change	28%	6%	.76	.21	.02	.54

6. Building a model of the user

The function we assign to our user model is to infer how the mental state of the user evolves during the dialog, in relation to his/her ‘stable’ characteristics and to the dialog history. The mental state components which are relevant in health promotion dialogs are stage of change and social attitude towards the agent, while emotional valence is important only in case of serious eating-related problems. Since we would like to assess these continuously throughout a dialog without asking the user to fill out a questionnaire after every

dialog move, we consider these ‘hidden’ variables whose values are to be inferred. We take our ‘observable’ measures to be the user’s stable characteristics, the context in which the move was entered (previous agent’s move), the length of the user move and its linguistic features as recognized by parsing. Intermediate variables are the signs of mental state: belief behavior wrong, intention to change and the features listed in Table 3 for social attitude.

The model is dynamic: it is initialized by assigning a value to the stable characteristics and (at every step of the dialog) to the context and the move characteristics, and produces a revised set of values for the hidden variables after every user move. The user model was built according to the philosophy described in Section 2: very simple parsing of the moves with integration of results in a Bayesian network which handles uncertainty in the relationships among the various features. We will describe the two components separately in the next two paragraphs.

a. move parsing

Language resources employed by humans to express their affective state are lexical (‘emotional words’), syntactic (e.g., emphatic construction) and morphological (e.g., terms of endearment or contempt) (Poggi and Magno Caldognetto, 2003, Storm and Storm, 1987, Batliner et al, 2003; Pennebaker et al, 2003).

Several researchers have investigated the assessment of affective state based on analysis of written language: Poggi and Magno Caldognetto (2003) worked on the Italian emotional language; Gill and Oberlander (2002) worked on recognizing personality traits in emails; Carberry et al (2002) studied the recognition of doubt; Guinn and Hubal (2003) proposed a semantic grammar enriched with emotional or attitudinal tags to recognize features like politeness, urgency, and satisfaction.

Other researchers have investigated affect assessment from spoken language, combining prosodic information with language features. Lee et al (2002) focused on recognizing ‘negative’ and ‘non negative’ emotions and found that combining acoustic information with salient keywords in utterances improved the recognition performance. Ang et al (2002) aimed at detecting frustration and annoyance in telephone-based dialogs after annotating a large corpus by five raters: they found that the labeling of emotions and speaking style was a ‘inherently difficult task’ and that emotion characteristics varied enormously from person to person and from context to context; therefore, although the language analysis method they applied was refined, they could only discriminate the ‘neutral’ case from the ‘annoyed and frustrated’ category. Litman et al (2003) combined acoustic-prosodic features with word analysis to recognize the valence of emotions in spoken tutoring dialogs. In working with WOZ data, Batliner et al (2003) demonstrated that the combination of prosodic with linguistic and conversational data yielded better results than the use of prosody only, for recognizing ‘troubles in communication’, that is, the beginning of emotionally critical phases in the dialog; this last study considered features, like ‘repetitions’, which require examining the dialog history rather than individual moves.

Table 6: Recognition criteria and predictive capacity of the parser

Signs	Criteria	Sensitivity	Specificity	Proportion of correctly classified cases
Friendly self-introduction	Expressions of greetings ('ciao', 'hello',...) or of self-presentation ('my name is...')	0.91	0.98	0.97
Familiar style	Agent name ('Valentina'), interjections ('!', 'Hurrah',...), friendly lexicon ('papa', 'mummy', 'greedy', 'chat', 'my passion', 'dear', ...), dialectal expressions ('cute', 'espressino',...), diminutive or expressive forms ('little sweet', 'fatty', ...)	0.36	0.96	0.79
Talks about self	Personal pronouns ('I', 'my', 'to me',...), auxiliary verbs ('I have', 'I am',...), expressions of knowledge ('I know', 'I believe',...), of attitude ('I try', 'I think', 'I tend to', 'I care of',...), domain verbs ('I eat', 'I drink',... all at the first person	0.81	0.79	0.80
Question about the agent	Similar to the previous one, but with the second pronoun	0.80	0.92	0.91
Positive comments	Expressions of agreement ('OK', 'right', 'good', 'true',...), of attitude ('I agree', 'I trust',...), of opinion about the agent ('That's kind of you',...)	0.56	0.94	0.92
Negative comments	Objections ('no', 'but',...) negative evaluations about the agent ('you are rude', 'you don't know', 'you don't understand') or about the message received ('this is too much', 'too little', ...)	0.16	0.98	0.93
Friendly farewell	Expressions of farewell ('bye', 'see you soon', ...), of thanking and wishes ('thanks', ...)	0.83	0.97	0.96
Believes behavior wrong	Declaration of own 'wrong' behaviors ('I don't believe I take all the substances I need' or 'I can't follow a more correct dietary habit',...); description of own shape as being 'not ideal' ('I'm overweight')	0.40	0.97	0.92
Intends to change	Weak or strong manifestations of desires to change ('I would like to', 'I should', 'I must',... 'change my dietary habits',... or similarities). Requests of suggestion ('What should I do?')	0.51	0.97	0.95

We defined our parsing criteria based on a preliminary analysis of the moves in which the raters agreed (fully or weakly) and based them on recognition of short word sequences (one, two or three keywords). The criteria applied for each sign of social attitude (in Table 6) combined the knowledge about the sign semantics with the analysis of word salience in the corpus: a word was considered to be 'salient' for a category if it appeared more often in the category than in other parts of the corpus (Lee et al, 2002). The predictive capacity of the parser was evaluated from confusion matrices in terms of sensitivity (true positives TP / total positive cases, also named 'recall'), specificity (true negatives TN / total number of negative cases, which is equal to $1 - \text{'fallout'}$) and proportion of correctly classified cases ($TP + TN / \text{total number of cases}$). The table shows that the specificity of parsing was high for all signs (ranging from 79 % to 98 %) while sensitivity was low for some of them; negative comments were the most difficult to recognize (sensitivity = 16 %), followed by a familiar style and positive comments. These values are, of course, a consequence of parsing criteria: in defining these criteria, we had to decide whether we preferred a high sensitivity or a high specificity (Manning and Schutzer, 1999). We felt that, in the case of social attitude, the consequences of false positives were less harmful for system effectiveness than those of false negatives: for the agent, false positives imply interpreting the user move as 'friendly' and answering in a friendly way even if this were not the case; false negatives imply, on the contrary, a 'cold' answer to some user attempts to establish a 'warm' relationship with the agent. In our view, the second risk should be avoided while the first is more acceptable. Therefore we did our best to design the parser so as to maximize the sensitivity even at the expense of a

lower specificity. As column 3 in table 6 shows, we succeeded in our attempt for some signs but not for all of them. In particular, recognizing negative comments, familiar style, beliefs and intentions requires far more complex parsing methods than those we applied in this study.

b. integration of signs in a Bayesian network

As the relationships among the user features can only be estimated, either subjectively or objectively, with some level of uncertainty, we decided to treat uncertainty probabilistically and to represent the user model with a *dynamic Bayesian network*. Bayesian networks (BNs) are probabilistic models that may be based on expert knowledge, empirical data or a combination of both. A BN consists in a network of assumed causal relationships between random variables and a set of conditional probability tables that relate every variable to its assumed causal variables (Pearl, 1988). In their dynamic version, BNs replicate at defined time intervals by establishing links with the previous layers for some of the nodes; monitored events occurring in every time interval produce new evidence to be propagated in the subsequent layer (Nicholson and Brady, 1994). When dynamic belief networks are applied to represent evolving user models in dialogs, the component at time t_i represents the system's image of the user at the i -th step of the dialog and the monitored event is the user move in the interval (t_i, t_{i+1}) . This formalism was proposed some years ago as a method to model affective human-machine interaction. Ball (2003) employed them to represent the relationships among emotion and personality components and their observable effects. Carofiglio et al (in press) modeled mixed emotion activation with dynamic models; they proved how these cognitive models may be employed in communication processes to represent, at the same time, prospective reasoning on possible consequences of some communicative act being planned and reasoning on the possible causes of a communication received (Carofiglio and de Rosis, 2005). Conati and Maclaren (2005) built and refined a probabilistic affective model to represent activation of emotional states as a consequence of goal satisfaction or threatening; the model was built on the famous Ortony, Clore and Collin's categorization of emotions (Ortony et al, 1988) and its validation was based on contrasting results with the subject's 'feeling'.

We employed the dataset of our Wizard of Oz studies to train the static component of the model with the K2 learning algorithm (Cooper and Herskovitz, 1992) which is provided by Bayesware⁸. This algorithm reduces the search space of possible BN structures by allowing developers to specify an ordering of the variables from which the network should be built. It is appropriate in our case (and more in general, we claim, in learning user models) because it enables distinguishing 'trigger' variables (which are placed at the top level of the network) from the variables which describe the resulting behavior of the user (which are placed at the lowest level, as leaf nodes). Links therefore describe the causal relationships among stable characteristics of the users, their behavior and the dialog dynamics via intermediate nodes.

Table 7 describes the variables in our model, with the labels employed to denote them:

⁸ <http://www.bayesware.com>

- a. *stable user characteristics*: background, in humanities or computer science ⁹;
- b. *context*: category of the previous agent move and of the current user move;
- c. *monitored variables*: social attitude towards the ECA and stage of change;
- d. *'hidden' subject characteristics*: the signs with which the user characteristics manifest themselves: believes behavior wrong and intends to change for the stage of change. Friendly self-introduction, familiar style, talks about self, personal questions to the agent, comments on the dialog and friendly farewell for the social attitude;
- e. *'observable' linguistic features* in the subject move produced by parsing.

K2, as well as other current BN learning algorithms, tries to find the model that fits data best by maximizing the log likelihood (MLL), but does not care about the use of the resulting model and its predictive ability. We tested several kinds of models by automatically learning both their structure and their parameters and by introducing a few links between some nodes to avoid problems in evidence propagation due to d-separation properties. The resulting structure is shown in figure 2. This figure does not include irony, that we finally excluded from the model because of the difficulty of defining parsing rules able to recognize it.

Table 7: Variables included in the model

Variable category	Variable name	Label
Stable user characteristics	Background	Back
	Gender	Gend
Context	Type of last Agent move	Ctext
	Type of user move	Mtype
Monitored variables	User attitude towards the agent	Satt
	Stage of change	SoC
Signs of social attitude	Familiar style	Fstyl
	Friendly self-introduction	Fsint
	Talks about self	Perin
	Questions about agent	Qagt
	Friendly farewell	F-Fw
	Comments	Comm
Signs of stage of change	Believes behavior wrong	Bbw
	Intends to change	Itc
Results of parsing	Cues of familiar style	Pfstyl
	Cues of friendly self introduction	Pfsint
	Cues of talks about self	Pperin
	Cues of questions to the agent	Pqagt
	Cues of friendly farewell	Pffw
	Cues of comments	Pcomm
	Cues of belief that behavior is wrong	Pbbw
	Cues of intention to change	Pitc

⁹ As this node is settled at the beginning of the interaction and does not change during the simulation, we omitted it from figure 2

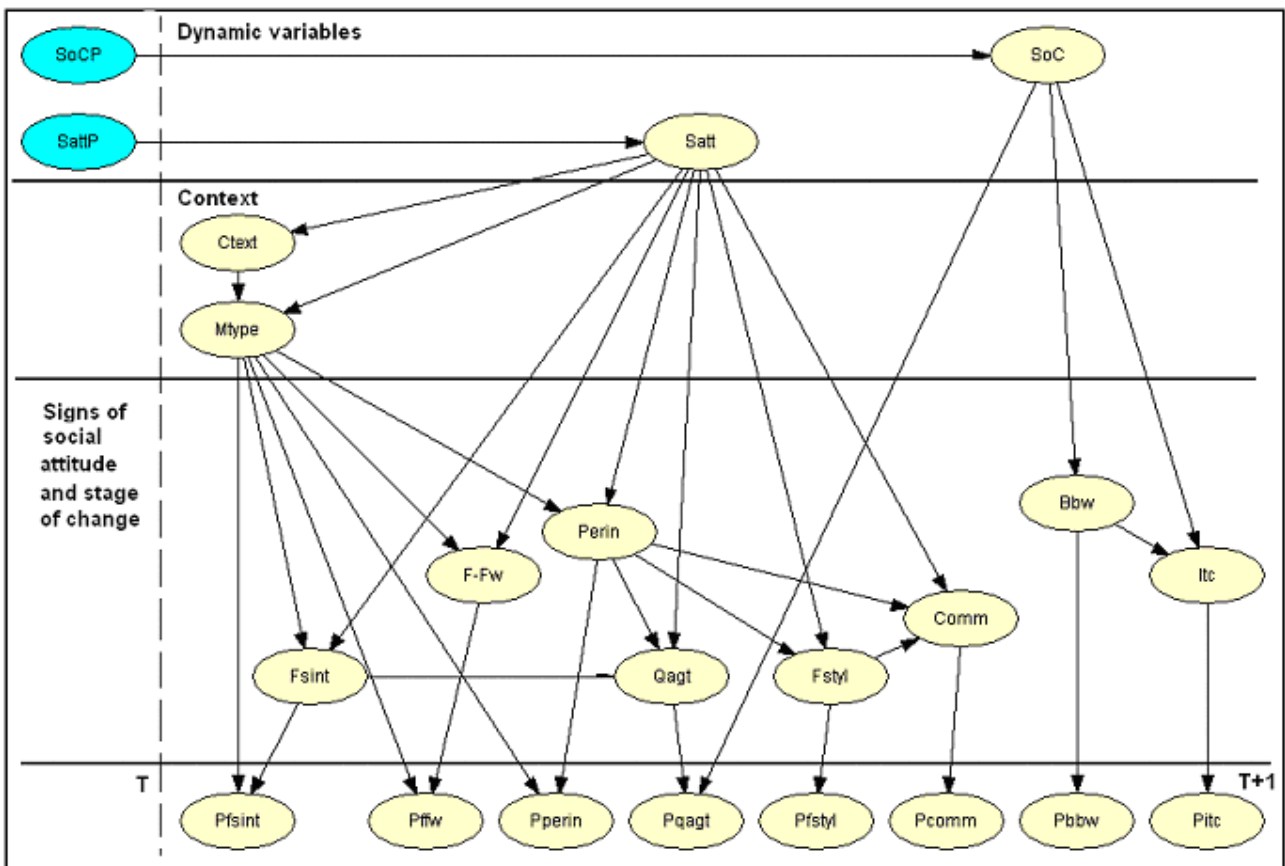


Figure 2: Structure of the dynamic user model.

Apart from the obvious associations between the agent's and the user's move types (self introduction, farewell, question-answering or question-suggestion couplings), this model shows that a background in humanities (an 'observable' variable in our case) implies a higher level of social attitude in general, and of irony and friendly self introduction in particular. When the level of social attitude is low, self presentation and farewell are sometimes omitted, and few comments are made after the agent's answers or suggestions. The use of familiar style is associated with irony and non-neutral valence. Self disclosure is frequently included in comments or answers to the agent's questions and is common in the contemplation and preparation stages of change, Personal questions about the agent are more frequent when the emotional valence is not neutral and are associated with a familiar style of move.

While the 'static' component of our model was learned from the dataset by considering every user move individually, the parameters associated with the two dynamic nodes were learnt from pairs of adjacent moves. This enabled us to estimate the probability that a move displaying a positive or a neutral 'social attitude' comes immediately after a move of the same or a different type (and the same for the stage of change). Probability tables associated with links between the two monitored variables in adjacent time slices t_i, t_{i+1} are such that, in absence of any sign of social attitude or stage of change in the user move occurring between t_i and t_{i+1} , the distributions of the two variables tend to smooth. This produces a *decay* of the probability of the value that was found as the most likely at time t_i .

The model runs as follows:

- *at the beginning of interaction*, it is initialized by propagating evidence about stable user characteristics; the probabilities of the other variables take prior values which correspond to this category of users;
- *after every user move*, linguistic features of the move are analyzed with the parser; results are introduced and propagated in the network together with evidence about the previous agent's move;
- *the new probabilities of the social attitude signs* are read and contribute to formulating the next agent move; the new probabilities of *stage of change and social and emotional attitude* (which evolve more slowly during the dialog) are employed to revise the high-level planning of the agent behavior.

c. model validation

Our model needs to be validated by contrasting its results with some external assessment of the two monitored variables. This external reference might consist in the observation of the user behavior (facial expression, gesturing etc), in some form of self-report (questionnaire-based user data) or in an independent, subjective evaluation by experts. Observation of the user behavior would probably be the preferable method. As far as the stage of change is concerned, standard self-report questionnaires might be employed as well. The problem is a bit more complex for social attitude, for which behavior observation methods and standard self-report questionnaires do not exist: in addition, for this kind of data (and, more in general, for short-term affective features) the reliability of self-report is being debated (see, e.g., Picard and Daily, 2005 and Höök et al, 2005). This kind of data might not be collected, in any case, after every agent move without influencing the user behavior: questionnaires might only be administered at the beginning and the end of interaction, to assess whether and how the user mind changed as a consequence of the system use.

For these reasons we decided, as a first step, to validate our model by comparing the model results after every user move with an independent expert evaluation. We evaluated the model ability to represent the dynamic aspects of the user with test on some of the dialogs included in the corpus, by comparing its results with the subjective estimates of the three raters (again, with a majority agreement criterion). The predictive value of the model was estimated in terms of the *percentage of correctly classified cases of social attitude* and was equal to .75, while we did not get enough data to attempt a similar evaluation for the stage of change.

The following is an excerpt of one of the test dialogs¹⁰, in which we show how the system's belief about dynamic features of a subject with a background in computer science evolves:

=====

TIME 0

P(SoC = Pre-contemplation) = .88

P(Satt = friendly) = .37

¹⁰ Translated from Italian

TIME 1

A: Hi, my name is Valentina. I'm here to help you understand how you can improve your diet.

U: *Hi, my name is Carla.*

Ctext = self-introduction *the agent did a 'self-introduction' type of move*
Pfsint = Y *the parser recognizes signs of a friendly self introduction in the user move*
P(Satt = friendly) = .99 *this move is showing a friendly social attitude*

TIME 2

P(Satt = friendly) = .80 *decay*

A: What are your eating habits?

U: *I always try to eat in a healthy way*

Ctext = question
Pperin = Y *the parser recognizes signs of 'talking about self'*
P(Satt = friendly) = .89 *the probability of social attitude increases again*

TIME 3

P(Satt = friendly) = .68 *decay*

A: What did you eat for breakfast?

U: *Milk and biscuits. How about you?*

Ctext = question
Pqagt = Y *the parser recognizes signs of 'personal questions to the agent'*
P(Satt = friendly) = .85 *the probability of social attitude increases again*

TIME 4

P(Satt = friendly) = .66 *decay*

A: Maybe you forget I'm only an artificial agent: I can't eat, and therefore I don't follow any particular diet.

U: *Sorry!*

Ctext = answer
Pfstyl = Y *the parser recognizes signs of 'familiar style'*
P(Satt = friendly) = .91 *the probability of social attitude increases again*

... some 'neutral' moves...

TIME 16

P(Satt = friendly) = .60 *the probability of social attitude increases again*

A: Sweets should be avoided, or limited to special occasions.

U: *But I can't resist a beautiful dessert!! What should I do?*

Ctext = suggestion
Pfstyl = Y *the parser recognizes signs of 'familiar style' and of 'believes behavior wrong'*
P(Satt = friendly) = .77 *the probability of social attitude increases again*
P(SoC = Contemplation) = .49 *the subject might now be in the contemplation stage*

... ..

=====

7. Conclusions

In this paper, we discussed the role of adaptive models in dialog systems for health promotion. We focused our attention on users' mental state: their stage of change (behavior, beliefs, intentions, knowledge of plans etc) and the relation between stage of change and affective characteristics, both individual and social. In our Wizard of Oz studies, responding appropriately to subjects' attempts to establish a social relationship with the agent proved to be a key factor to engage subjects' attention and increase their level of involvement in the dialog: these effects provide an important platform upon which more sophisticated persuasion strategies

may be implemented. In addition to providing an insight into the user-agent relationship which is established in the health promotion domain, the collected dialogs served to define recognition criteria for the mental state of the users by means of linguistic analysis of their moves. To improve the parser performance, we employed dynamic Bayesian networks: these models enabled us to consider, in the recognition process, the stable characteristics of the user and the dialog history, resulting in increased accuracy of user modeling.

We based our work on two kinds of corpora. The first comprised natural conversations between a human dietician and a client, and the second consisted of dialogs with an artificial agent. The two situations were very different from one another with respect to the severity of problems tackled and the characteristics of subjects involved.

As for the methods employed, we found that the signs of stage of change in our corpus were infrequent and were not difficult to recognize by our raters; whereas the signs of social attitude were frequent and easy to recognize in some cases, and quite uncommon and difficult to recognize in others.

In constructing our user model, we did not rely completely on automatic learning algorithms for building the Bayesian networks, but integrated this statistical method with the theoretical knowledge of relationships among variables in the model. Defining an order of search for variables and carefully revising the learned links and parameters were key steps of this interactive procedure. This corresponds to the prevailing attitude in the model learning community (see, e.g. Sebastiani et al, 2000), where it is suggested that models learnt from a database should be evaluated not only according to their accuracy in explaining data, but also according to their theoretical plausibility. However, models induced from a sample of data can be over trained, and may not generalize well to new situations (Manning and Schütze, 1999). Therefore, learned models should be validated on additional, external data, and this is the next step in our research, with a systematic analysis of the predictive ability of the model and how it is influenced by changes in its parameters.

Compared with the growing body of work in emotion recognition, to the best of our knowledge this is the first attempt to recognize and model stages of change and social attitude in user-ECA interactions. One of the limits of our recognition method is that we did not measure the personality of subjects in our WOZ studies: we did this to avoid reducing their level of cooperation in the experiment by having to respond to a long initial questionnaire. Therefore, it is possible that the stable characteristics which influence the subject behavior in our studies (gender and background) hide or partially overlap with some personality traits. This hypothesis is supported by the similarity between our signs of social attitude and some of the language features which characterize the behavior of ‘introverted’ or ‘extraverted’ subjects in (Gill and Oberlander, 2002). According to that study, extraverts tend to use a ‘relaxed’ and ‘informal’ style, make less use of the first person singular pronouns than introverts and tend to express a ‘positive affect’ more frequently. These linguistic signs are similar to those we included among our signs of ‘social attitude’, so that it is reasonable to wonder whether social attitude is related to the extraversion dimension of personality.

We learned a great deal from our experience of iterative prototyping of a health promotion dialog system. We started our research with the belief that a key requirement of dialog simulation was the recognition of the

emotional state of the users. This is true when the user problems are serious and therefore produce a strong emotional state (as in the case of natural dialogs with a counselor about drinking and smoking). On the contrary, when the subjects involved are younger and their problems are less serious, different kinds of emotions emerge in the interaction: rather than strong ‘individual’ emotions like fear, joy, anxiety, relief etc, softer ‘social’ emotions like sympathy or antipathy, tenderness, contempt, and sense of belonging are expressed (Poggi and Magno Caldognetto, 2003). To increase the effectiveness of advice-giving, the ability to recognize the degree of involvement of the user and to manifest the reciprocity of social attitudes is probably even more important than displaying realistic expressions of emotions in the animated agent’s face. This raises complex problems, like recognizing and responding to humor (Stock and Strapparava, 2003) or deception attempts (de Rosis et al, 2003) or formulating moves with appropriate ‘politeness’ which are all important areas for future research.

Acknowledgements

This work was financed, in part, by HUMAINE, the European Human-Machine Interaction Network on Emotion (EC Contract 507422). Nicole Novielli began to work at this project when she was at the University of Liverpool under the tutorship of Floriana Grasso, in the scope of a SOCRATES agreement with the University of Bari. We thank Giuseppe Clarizio for cooperating in the implementation of the WOZ tool and Gianni Cozzolongo and Irene Mazzotta for cooperating in labelling the corpus. We thank Loquendo for providing us their software in the scope of a scientific cooperation agreement and Mrs Pauline Butts for being so kind to correct our bad English after interpreting our ideas. We cannot conclude this list without thanking warmly Tim Bickmore, who interpreted his role of Guest Editor in a such cooperative way, to spend a lot of time in reviewing and revising our paper.

References

- J.Ang, R.Dhillon, A.Krupsky, E.Shriberg and A.Stolcke. Prosody based automatic detection of annoyance and frustration in human-computer dialog. *Proceedings of ICSLP*, 2002.
- E. Ball. A bayesian heart: Computer recognition and simulation of emotion. In R.Trapp, P.Petta and S. Payr (Eds): *Emotions in Humans and Artifacts*. The MIT Press, 2003.
- J. Bates: The role of emotions in believable agents. *Communications of the ACM*. 37 (7) 1994.
- A. Batliner, K. Fischer, R. Huber, J. Spilker and E. Noth: How to find trouble in communication. *Speech Communication*, 40, 2003.
- D Berry, L Butler and F de Rosis. Evaluating GRETA. The importance of consistency of behaviour in a multimodal animated agent. *International Journal of Human-Computer Studies*, 63, 2005.
- T. Bickmore, Gruber A, Picard R. Establishing the computer-patient working alliance in automated health behavior change interventions. *Patient Educ Couns* 2005;59(1):21-30.

- S Carberry, L.Lambert and L.Schroeder. Towards recognising and conveying an attitude of doubt via natural language. *Applied Artificial Intelligence*, 2002.
- J. Carletta: Assessing agreement on classification tasks. The Kappa statistics. *Computational Linguistics*, 22, 1996.
- V. Carofiglio, F. de Rosis, and R. Grassano. Dynamic models of mixed emotion activation. In L. Canamero and R.Aylett (Eds): *Animating expressive characters for social interactions*. John Benjamins Publ Co, in press.
- V. Carofiglio and F. De Rosis: In favour of cognitive models of emotions. *AISB'05 Workshop on 'Mind-Minding Agents*, 2005.
- J. Cassell and T. Bickmore. Negotiated collusion: modelling social language and its relationship effects in intelligent agents. *User Modelling and User-Adapted Interaction*, 13, 1-2, 2003.
- J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors. *Embodied Conversational Agents*. Cambridge, MA: The MIT Press; 2000.
- A Cavalluzzi, V. Carofiglio and F. de Rosis: Affective Advice-Giving Dialogs. In E.André, L.Dybkaer, W.Minker and P.Heisterkamp (Eds) *Affective Dialogue Systems*. Springer LNAI 3068, 2004.
- A. Cavalluzzi, F.de Rosis, I. Mazzotta and N. Novielli: Modeling the user attitude towards an ECA.UM'05 Workshop on *Adapting the interaction style to affective factors*. Edinburgh, July 2005.
- K.M. Colby: Modeling a paranoid mind. *Behavioral and Brain Sciences*. 4, 1981.
- K.M. Colby, P.M. Colby and R.J. Stoller: Dialogues in natural language with GURU, a psychologic inference engine. *Phylosophical Psychology*, 3, 1990
- K.M.Colby: Human-computer conversation in a cognitive therapy program. In Y. Wilks (Ed): *Machine Conversations*. Kluwer Academic Publishers. 1999
- C.Conati and H.Maclaren: Data-driven refinement of a probabilistic model of user affect. In L.Ardissono, P.Brna and A.Mitrovic (Eds): *User Modeling 2005*. Springer LNAI 3538.
- G.F.Cooper and E.Herskovitz. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* , 9, 1992.
- R. Coulston, S. Oviatt and C. Darves. Amplitude Convergence in Children's Conversational Speech With Animated Personas. In J Hansen and B.Pellom (Eds): *Proceedings of the 7th International Conference on Spoken Language Processing*. 2002.
- R. Cowie: Describing the emotional states expressed in speech. *Speech Emotion*, 2000.
- R.Craggs and M McGee Woods. A two dimensional annotation scheme for emotion in dialogue. *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, 2004.
- N. Dahlback, A. Joensson, and L. Ahrenberg. Wizard of Oz Studies. Why And How. *Proceedings of the Int Workshop on IUI*, 1993.

- S. Darves and S. Oviatt. Adaptation of Users' Spoken Dialogue Patterns in a Conversational Interface. *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP2002)*. J. Hansen and B. Pellom (Eds). 2002.
- B. De Carolis, C. Pelachaud, I. Poggi and M. Steedman. APMML, a Mark-Up Language for believable behavior generation. In H Prendinger and M Ishizuka (Eds): *Life-Like Characters: Tools, Affective Functions and Applications*. Springer, Berlin. 2003.
- F. de Rosis, A.Cavalluzzi, I.Mazzotta and N.Novielli: Can embodied conversational agents induce empathy in users? *AISB Virtual Social Characters Symposium, 2005*.
- F de Rosis, C Pelachaud, I Poggi, V Carofiglio and B De Carolis. From Greta's Mind to her Face: Modeling the Dynamics of Affective States in a Conversational Embodied Agent. *International Journal of Human-Computer Studies*. B.R.Gaines. Vol.59 (1-2), 2003B.Di Eugenio and M. Glass. The Kappa statistics, a second look. *Computational linguistics*, 2004.
- P.Ekman. An argument for basic emotions. *Cognition and Emotion*, 6, 1992.
- A.J.Gill, and J. Oberlander. Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*. 2002.
- F. Grasso, A. Cawsey and R Jones. Dialectical argumentation to solve conflicts in advice giving: a case study in the promotion of helath nutrition. *International Journal of Human-Computer Studies*, 53, 2000.
- N. Green, T. Britt and K. Jirak. Communication of uncertainty in clinical genetics patient health communication systems. *AAAI Fall Symposium on Dialogue Systems for Health Communication*. 2004.
- C.Guinn and R.Hubal: Extracting emotional information from the text of spoken dialog. In C.Conati, E.Hudlicka and C.Lisetti (Eds): Workshop on "Assessing and adapting to user attitudes and affect: why, when and how?". In the scope of User Modeling 2003.
- K.Höök, K. Isbister and J. Laaksholahti: Sensual evaluation instrument. CHI Workshop on "Evaluating affective interfaces: innovative approaches". Portland, April 2005.
- W.L: Johnson, C. LaBore and Y-C Chin. A pedagogical agent for psychological intervention on a handheld computer. *AAAI Fall Symposium on Dialogue Systems For Health Communication*. 2004.
- Laforge R, Greene G, Prochaska J. Psychosocial factors influencing low fruit and vegetable consumption. *Journal of Behavioral Medicine*. 1994;17(4):361-374.
- C.M.Lee, S.S. Narayanan, and R. Pieraccini. Combining acoustic and language information for emotion recognition. *Proceedings of ICSLP*. 2002.
- Li R, Serdula M, Bland S, al e. Trends in fruit and vegetable consumption among adults in 16 US states: Behavioral Risk Factor Surveillance System, 1990-1996. *Am J Public Health*. 2000;90:777-781.

- D.Litman, K.Forbes and S.Silliman: Towards emotion prediction in spoken tutoring dialogues. *Proceedings of HLT/NAACL 2003*.
- C.D. Manning and H. Schutze. *Foundations of statistical natural language processing*. The MIT Press, 1999.
- S.C.Marsella, W.L.Johnson and C.M.Labore. Interactive pedagogical drama for health interventions. In U.Hopper et al (Eds): *Artificial Intelligence in Education. Shaping the future of learning through intelligent technologies*. 2003.
- M. Miceli, F. de Rosis and I. Poggi. Emotional and non emotional persuasion. *Applied Artificial Intelligence*, in press.
- C Nass, K. Isbister and E-J Lee. Truth is beauty: Researching Embodied Conversational Agents. In *J. Cassell, J. Sullivan, S. Prevost and E Churchill: Embodied Conversational Agents*. The MIT Press, 2000.
- A.E. Nicholson, and J.M. Brady.: “Dynamic belief networks for discrete monitoring”. *IEEE Transactions on Systems, Men and Cybernetics*, 24(11), pp. 1593-1610. 1994
- A.Nijholt. Observations on humorous act construction. *Proceedings of EMCSR 2004*.
<http://www.home.cs.utwente.nl/~anijholt/artikelen/emcsr2004.pdf>
- A.Ortony, G.L. Clore G.L. and A.Collins. *The cognitive structure of emotions*. Cambridge University Press (1988).
- S. Oviatt and B. Adams. Designing and Evaluating Conversational Interfaces With Animated Characters. In *J Cassell, J Sullivan, S Prevost and E Churchill: Embodied Conversational Agents*. The MIT Press, 2000.
- A.Paiva, R-Aylett and S.Marsella. *AAMAS04 Workshop on Empathic Agents*.
<http://gaips.inesc.pt/gaips/en/aamas-ea.html>. 2004
- J.A. Pearl: Probabilistic reasoning in intelligent systems: Networks of plausible inference. Morgan Kaufman Publishers, 1988.
- C. Pelachaud and M. Bilvi. Computational model of believable conversational agents. In M.P. Huget (Ed): *Communication in Multiagent Systems: background, current trends and future*. Springer Verlag LNCS 2650. 2003.
- J.Pennebaker, M.Mehl and K.G.Niederhoffer: Psychological aspects of natural language use. Our works, our selves. *Annual Review of Psychology*, 54, 2003.
- R. Picard: *Affective Computing*. The MIT Press, 1997.
- R. Picard. What does it mean for a computer to ‘have’ emotions? In *R. Trappl, P. Petta and S. Pays (Eds): Emotions in humans and artefacts*. A Bradford Book, MIT Press, 2002.
- R.W.Picard and S.B.Daily: Evaluating affective interactions: alternatives to asking what users feel. CHI Workshop on “*Evaluating affective interfaces: innovative approaches*”. Portland, April 2005.

- R. Plutchik and H.R. Conte (Eds). Circumplex models of personality and emotions. APA Books, 1996.
- I. Poggi. Emotions from mind to mind. In A Paiva (Ed): *Proceedings of the Workshop on "Empathic Agents"*. AAMAS 2004.
- I. Poggi and E. Magno-Caldognetto. Il parlato emotivo. Aspetti cognitivi, linguistici e fonetici. *Proceedings of the Conference "Il Parlato italiano"*. D' Auria, Naples, 2003
- J. Prochaska, C. Di Clemente and H.Norcross. In search of how people change: applications to addictive behavior. *American Psychologist*, 47 (1992) 1102-1114
- Z. Ruttkay and C. Pelachaud (Eds). *From brows till trust: evaluating embodied conversational agents*. Kluwer Human-Computer Interaction Series, 7, 2004.
- P. Sebastiani, M. Ramoni and A. Crea. Profiling your customers using bayesian networks. *Proceedings of SIGKDD Explorations*. ACM SIGKDD 2000.
- J. A. Sillince and R. H. Minors: What makes a strong argument? Emotions, highly-placed values and role-playing. *Communication and Cognition*, 1991
- B. Silverman, J. Holmes, S. Kimmel, C. Branäs, D. Ivins, R. Weaver and Y. Chen. Modeling emotion and behavior change: The case of the HEART-SENSE game. *Health Care Management Science*, 4, 3, 2001.
- O. Stock and C. Strapparava. An experiment in automated humorous output production. *Proceedings of Intelligent User Interfaces*, 2003.
- C. Storm and T. Storm: A taxonomic study of the vocabulary of emotions. *Journal of Personality and Social Psychology*, 53, 1987.
- D.R.Traum and S.Larsson (Eds): The information state approach to dialogue management. *Current and new directions in discourse and dialog*. J.Van Kuppevelt and R.Smith (Eds). Kluwer, 2003.
- W.F.Velicer, J.O.Prochaska, J.L.Fava, G.J.Norman and C.A.Redding. Smoking cessation and stress management: applications of the Transtheoretical Model of Change. *Homeostasis*, 38, 1998.
- D. Walton. The place of emotion in argument. The Pennsylvania State University Press, 1992.
- C. Wegman: Emotion and argumentation in expression of opinion. In V Hamilton, G.H.Bower and N.H Frijda (Eds): *Cognitive perspectives on emotion and motivation*. Kluwer, 1988.
- Y. Wilks (Ed): *Machine Conversations*. Kluwer Academic Publishers. 1999.

