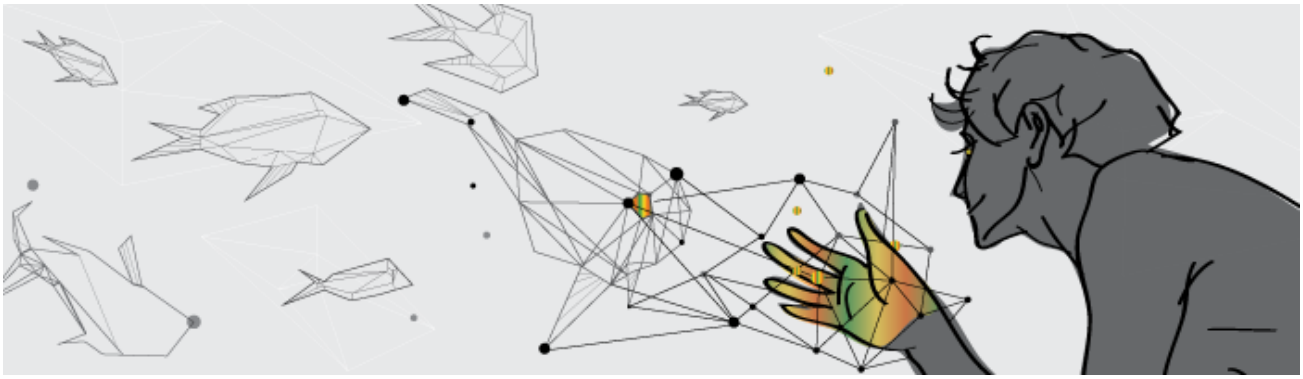
Proceedings of the CASA – Computers as Social Actors workshop 2013

In association with the 13th International Conference on Intelligent Virtual Agents (IVA), Edinburgh, UK



Workshop website: http://casa-workshop2013.blogspot.com/ | Graphics by Xinwei Wang

Foreword

This volume contains the proceedings of the 1st CASA-Computer as Social Actors Workshop, in association with the 13th International Conference on Intelligent Virtual Agents (IVA), held in Edinburgh on August 28th, 2013. The CASAs mission is to bring together researchers from different disciplines and combine their knowledge and expertise contributing in a multidisciplinary way to the advancement of Computers as Social Actors. The CASA Workshop fo-cuses on three main areas of investigation: theory, practice and market.

The scientific field of CASA is highly interdisciplinary, encompassing development of technological components, de-sign methodologies, and the adoption and take up of CASA solutions and services. The main emphasis is to exploit many different human-machine and human-human interaction technologies and methodologies addressing several dif-ferent concrete scenarios identifying key characteristics of social actorship.

Social actorship is a concept that does not have a precise definition in literature. People apply social rules to many as-pects of human-computer interaction independently of whether or not the systems are given explicitly anthropomorphic interfaces. Social actorship refers to systems that present social awareness and intentionality qualities, and possibly some form of embodiment. Humans, when interacting with CASA can be led to feel empathy, and experience a diverse set of emotional reactions. Social actorship can also refer to systems, such as computers, robots and other artefacts, that are able of invoking social responses from its users. Consequently, the social actorship of a system is a combination of different elements that do not depend only on the system itself but also on the context, the presence of, and interaction with other actors. The modulation of these elements contributes to the perception of the system as a social actor.

The CASA Workshop is supported by EIT ICT Labs (www.eitictlabs.eu).

Workshop organizers

Mario Conci TrentoRise, Italy
Virginia Dignum Delft University of Technology, Netherlands
Mathias Funk Eindhoven University of Technology, Netherlands
Dirk Heylen University of Twente, Netherlands

Scientific committee

Workshop papers

User Experience and Social Attribution for an Embodied Spoken Dialog System

Benjamin Weiss and Simon Willkomm

Quality and Usability Lab, TU Berlin, Germany
Benjamin.Weiss@tu-berlin,
home page: http://qu.tu-berlin.de

Abstract. A public information system with an Embodied Conversational Agent is evaluated in a laboratory setting concerning Social Actorship, Social Acceptance, perceived Control, Pragmatic Quality and Hedonic Qualities. Results show a positive experience for Pragmatic Quality and Control, but negative ratings for Social Acceptance. Differentiating these various aspects of User Experience has proven to be fruitful for this summative evaluation, especially considering the potential public situation of interaction.

Keywords: Embodied Conversational Agents, Social Actorship, Spoken Dialog System, User Experience

1 Introduction

Spoken dialog systems (SDS) can provide a natural and intuitive way of interacting due to an interface operated by voice. Embodied conversational agents (ECAs) also use spoken language to interact, but in addition exhibit at least an anthropomorphic interface, for example by visually modeling a human face. From a user point of view, embodiment can result in increased expectations on the capabilities of the ECA, assuming for example social skills and intelligence, which should be reflected in sophisticated (human-like) communication behavior. If such expectations are not met, user experience will be negative.

But the embodiment might also result directly in positive user experience (UX): The multimodal stimulation itself (typically audio-visual for non-robot embodiments) can be positive. It also might increase user attention and thus facilitate interaction with such a system. Additionally, embodiment enables designers to present an attractive interface for more than the acoustic modality. Concerning expectations, assumed social and cognitive capabilities attributed to an ECA will be beneficial when such expectations are not disappointed.

The main objective of this paper is to evaluate UX, social abilities in particular, of an embodied visitor guide.

This virtual visitor guide is a speech operated system with an audio-visual synthesis in the form of a lip-synchronous talking head. Its purpose is to inform visitors in a welcome and demonstration hall about research and development projects. Typical visitors received in this hall are student groups, prospective

students, colleagues from industry on a company outing, professors and managers on a collaboration visit and at last, Berlin citizens and tourists on the annual "long night of science".

By enabling spoken conversation and showing literally a human like face, the virtual guide is supposed to motivate and support interaction and interest and provide an interacting mode which ...

- ... is complementary to visual information on posters,
- activates the visitor (as s/he has to talk to the guide instead of just read the posters available),
- and sends visitors to demonstrators and thus actives visitors to explore.

2 Attribution of social abilities

Researchers from various disciplines have used different approaches on their own definition of UX [1]. Consequently, the aim of defining a standardized definition of UX resulted in "a person's perceptions and responses that result from the use or anticipated use of a product, system or service" [2], which incorporates every aspect of perception and response concerning the usage of an interface.

The focus of User Experience is on any experience of users during interaction with a system. It is not limited to conscious (retrospective) reflections on the usability or usefulness of a given service operated with an interface, but concentrates on sub-conscious affective reactions of the interaction, which can, of course, be asked for in retrospection. This paradigm shift on the last decades aims at understanding the user better, especially event-driven affection ("Wow Effect", frustration) and sometimes confusing decisions concerning, e.g., user acceptance of certain devices based mainly on the big impact of aesthetics or Social Norm [3].

Although dimensions of UX are not fully understood [4], the separation of overall attractiveness (how positive or negative a user rates a device or interface) into one pragmatic (how usable or useful) and two hedonic qualities [5] seem to be quite established. These two hedonic dimensions are Identification – how much can a user identify with a device/interface – and Stimulation – how interesting, exciting is using this device/interface. A questionnaire assessing these dimensions is also already provided.

Still, other dimensions or more concrete aspects of UX are of interest, especially when dealing with embodied spoken interaction and with interaction in public spaces. Social aspects come into play for such interfaces and usage situations, e.g., the attribution of *social actorship* and the experience of *interacting in a social context.*

Whereas the former issue deals with assumed, expected or attributed competences towards the system (e.g., intelligence, intentionality, awareness), the latter issue deals with the user feelings concerning privacy, control, or social acceptance. This view is actually a little different from the definition developed within the EIT RIHA 12124 "Computers as Social Actors" (2012) that subsumes both aspects mentioned under the term "Social Actorship":

Social Actorship is the ability of the system to act in a social context, with an implicit or explicit goal. From the user perspective, Actorship is a characteristic of the system that makes the user perceiving it as a human actor to which s/he can direct their attention and have attention in return (This can be explained by the Mirror Concept: the system that sense something and acts in response). Although, some systems could be seen as just a mediating actor, like mobile phone and ICT in general, that fosters social interaction among people. In this case social Actorship is seen as the ability to influence and support the social life of people.

This definition also takes attribution of social abilities to a system/device/interface and the impact of such a system/device/interface on a user's social situation as two important aspects of UX. Therefore, a questionnaire was used to assess these aspects of UX, based on instruments and definitions available:

Attractiveness (ATT): The overall attractiveness of the system or interface after interaction. The difference to overall quality is the subjective aspect of attractiveness being not limited to pragmatic and general considerations, but including also hedonic subjectively experience aspects. (2 items [5])

Pragmatic Quality (PQ): The usability and usefulness of the system or interface. (4 items [5])

Hedonic Quality–Identity (HQI): The degree this system or interface fits to a user. This aspect is related to Social Acceptance. (2 items [5])

Hedonic Quality–Stimulation (HQS): The degree the interacting is positively stimulating. (2 items [5])

Social Acceptance (SA): User's social acceptance (according to [6]) subsumes how a user feels when interacting with a system regarding to the social situation, e.g. how uncomfortable or embarrassed in the light of potential other people or ones own norm. (5 items [7])

Social Actorship (SH): The degree the system exhibits social capabilities. (5 items [7])

Perceived Control (PC): The degree a user feels in control of the system and knows how to interact with it. (5 items [8])

The questionnaire provided by [8] is actually based on a model of technology acceptance described in [9].

3 Embodied conversational system

This ECA is embodied as a bald male person, based on the Thinking Head system [10]. The German text-to-speech system "OpenMary" [11] was chosen for the acoustic speech output and Sphinx as automatic speech recognition system [12]. The dialog was defined in VoiceXML running with Optimtalk [13]. The system itself is modular and uses events to let the modules communicate with each other.

Fig. 1. Graphical representation of VirtualK.

The chosen visual appearance is a bald male talking head, determined in an informal pre-test with six participants in [14]. This embodiment also exhibits no photographic texture and represents the consensus, as it was considered most pleasant and least irritating (cf. Figure 1).

The ECA gives visual conversational feedback, i.e. a nod signals the processing of a user utterance and if the user is not recognized for 20 video frames, the ECA will close its eyes and stop/pause the conversation.

A webcam is used to detect a user within the interaction sphere of the system, and the ECA will open its eyes and initiate the dialog with general information, and by asking the user about the interest in one of four research fields (video, audio, smartphone apps, or mobile interfaces); however, only one out of two for the experiment conducted. The system provides project-related information either by project name or by suggesting a project based on the preferred topic (audio: music, communication; video: quality, mobile TV; apps: phone control and leisure time; mobile interfaces: security, cross-service). It is able to provide more project related information than the demonstrators and posters.

If a face is not recognized for 20 video frames, the ECA will again close his eyes.

For each project, there are two levels of information (and if available, using a demonstrator is offered): General description and additional information. After each block of information presented, the system asks whether it should proceed or not (see Figure 2 for a simplified scheme of the dialog).

There are actually two versions tested, a typical one and a system with user-centered adaption concerning user recognition after a break, remembering interest for project suggestions, confirmation strategy dependent on no matches

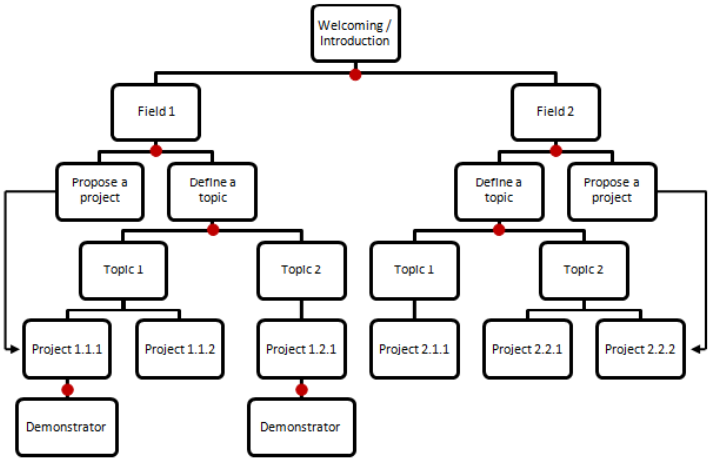


Fig. 2. The simplified dialog structure.

and confirmed false recognitions, and level of detail presented automatically. However, as there are no significant differences in the questionnaire data between both version, there will be no further description presented here.

4 Procedure

The aim of this evaluation is to assess User Experience and social capabilities attributed to the ECA in general and whether adaptive system components increase UX.

A laboratory experiment was chosen for this first evaluation regarding social aspects. The face recognition was set to a maximum by deleting previous users at the start of each experimental trial. For continuous duty, we lack information of the number of visitors a day, but it is expected to "forget" users after about four hours in order to successfully discriminate users. Also, the four research field were split into two categories, each comprising about half of the projects and demonstrators available to avoid boredom when trying out the system repeatedly.

A total of 30 test subjects took part in the experiment, gender balanced (14 female, 16 male), aged between 20 and 43 (average 26.4). All were paid for their contribution.

The initial experimental design was also planned for a comparison of the adaptive and non-adaptive version. Therefore, each user interacted two times with the system, each time with providing two of the fours research fields. The order of both research fields and order of adaptivity was balanced.

All users successively interacted with both versions of the SDS. They were asked to inform themselves about three to four projects and try out at least one

demonstrator. Each individual experimental session took about one hour with roughly 15–20 min. for each interaction.

After each trial the test subjects answered a questionnaire comprising aspects of User Experience on 5-point scales (antonyms for the AttrakDiff [5] and a Likert scale for [7]) to subjectively test for a benefit of the adaptions. The AttrakDiff was also filled out in the beginning after a brief video to assess a user expectations. Also, the perceived ASR quality was assessed on one Likert scale after each interaction.

5 Results and Discussion

There are no differences between the adaptive and non-adaptive version on any of the questionnaire scales assessed, as well as for research field or position of adaptivity ($\alpha = .05$, repeated measures Anova). Therefore, the system is analyzed as one, averaging the rating for the adaptive and non-adaptive version for each user.Consequently, the analysis is concerned whether the ratings on the different scales is positive or negative in comparison to the center of the 5-point scale (see Table 1). The significant results are similar to those with the non averaged ratings (doubled number), anyway.

Table 1. Results for the t-tests on divergence from an average rating.

Subscale	t(df=29)	p-level
ATT	0.05	$p = .959$
PQ	2.70	$p < .05$*
HQI	0.79	$p = .437$
HQS	-1.77	$p < .861$
SH	-1.08	$p = .290$
SA	-2.42	$p < .05$*
PC	4.38	$p < .001$***

For three of the seven scales, there is a significant positive or negative derivation from the center of 3. See Figure 3 for the distribution of ratings (median and quartile). Positively rated are Pragmatic Quality and Perceived Control, whereas Social Acceptance is more negative than the center of the scale. PQ and PC are of course significantly different from SA.

The former two scales represent related constructs, at least based on their descriptions. A strong correlation, actually the highest except for ATT and HQI, strengthens this impression (PQ–PC, ATT–HQI: Pearson's $r = .79$, $p < .001$***).

The other constructs which are assumed to be related, are HQI and SA. But these two do not show similar results and neither the strongest correlation

($r = .63$, $p < .001^{***}$), as both, HQI and SA correlate stronger with ATT ($r = .79$, $p < .001^{***}$ and $r = .74$, $p < .001^{***}$), and SA also with SH ($r = .66$, $p < .001^{***}$).

In summary, the results can be interpret that interacting with this embodied system was quite positive from a usability point of view (PQ, PC), but also quite unpleasant regarding the social situation (SA). The latter scale, however, has to be considered as more important in the frame of this evaluation, as the usage situation is public and the embodiment was explicitly chosen to improve the User Experience. Of course, there is no comparison with a non-embodied version of this SDS, but as a conclusion, this system should be either improved concerning the negative aspects, or even replaced by a different interaction paradigm, e.g. a non-embodied touch-screen.

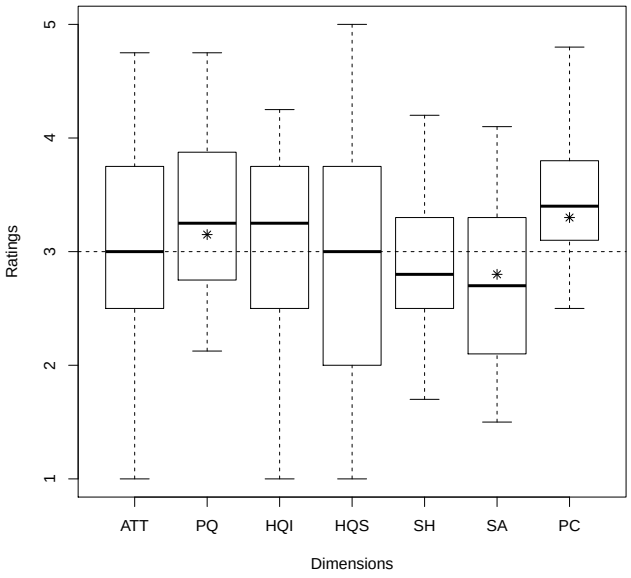


Fig. 3. Questionnaire results for all seven scales. Stars indicate significant divergence from the center (dotted line).

There is only one scale differing for gender: Perceived Control is higher for male users ($F(1, 27) = 4.33$; $p < .0.05$). The related PQ is not significantly different for gender ($F(1, 27) = 1.98$; $p = .17$). As there is no female face tested as well, it cannot be concluded if this result originates from the gender of the ECA or from other sources, e.g., technical affinity. Still, it would be interesting if female users find it especially easy to interact with a male face in this technological domain.

6 Conclusion

The ECA used in a spoken dialog information system was evaluated in a laboratory setting concerning various aspects of User Experience. Results indicate a negative experience concerning Social Acceptance, but a positive experience regarding Pragmatic Quality and Perceived Control. A relationship was found for the last two scales, which are also related in description. The various scales have proven to be useful for summative evaluation of this Embodied Conversational Agent in order to obtain a detailed feedback from users. The two scales with significant negative results have to be taken are more severe than the two positive ones when considering the public interaction situation.

References

1. Hassenzahl, M.: User experience (UX): Towards an experiential perspective on product quality. In: Proc. International Conference of the Association Francophone d'Interaction Homme-Machine, New York, USA, ACM (2008) 11–15
2. ISO 9241-210: Ergonomics of human system interaction – Part 210: Human-centred design for interactive systems (formerly known as 13407). International Organization for Standardization (ISO), Switzerland (2010)
3. Bevan, N.: What is the difference between the purpose of usability and user experience evaluation methods? In: Proc. UXEM'09 Workshop, INTERACT. (2009)
4. Scapin, D., Senach, B., Trousse, B., Pallot, M.: User experience: Buzzword or new paradigm? In: 5th International Conference on Advances in Computer-Human Interactions (ACHI), Valencia. (2012) 336–341
5. Hassenzahl, M., Monk, A.: The inference of perceived usability from beauty. Human-Computer Interaction **25**(3) (2010) 235–260
6. Montero, C., Alexander, J., Marschall, M., Subramanian, S.: Would you do that? – understanding social acceptance of gestual interfaces. In: MobileHCI. (2010)
7. Heerink, M., Kröse, B., Wielinga, B., Evers, V.: Assessing acceptance of assistive social agent technology by older adults: the almere model. International Journal of Social Robotics **2** (2010) 361–375
8. Venkatesh, V.: Determinants of perceives ease of use: Integrating control, intrinsic motivation, and emotion into the Technology Acceptance Model. Information Systems Research **11** (2000) 342–365
9. Venkatesh, V., Morris, M., Davis, G., Davis, F.: User acceptance of information technology: Towards a unified view. MIS Quarterly **27** (2003) 425–478
10. Luerssen, M., Lewis, T.: Head X: Tailorable audiovisual synthesis for ecas. In: Interacting with Intelligent Virtual Characters Workshop (IIVC). (2009)
11. Schröder, M., Trouvain, J.: The German text-to-speech synthesis system mary: A tool for research, development and teaching. International Journal of Speech Technology **6** (2003) 365–377
12. Carnegie Mellon University: Sphinx speech recognition
13. OptimSys s.r.o.: Voice browser
14. Weiss, B., Tönges, R.: Automatic adaption of spoken dialog systems for public and working environments. In: International Conference on Interfaces and Human Computer Interaction (IHCI), Lisbon. (2012) 284–288

The Effect of Variations in Emotional Expressiveness on Social Support

Janneke M. van der Zwaan, Virginia Dignum, and Catholijn M. Jonker

Delft University of Technology

Abstract. There is a growing interest in employing embodied agents to achieve beneficial outcomes for users, such as improving health, or increasing motivation for learning. The goal of our research is to explore how and to what extent embodied agents can provide social support to victims of cyberbullying. To this end, we implemented a proof of concept virtual buddy that uses verbal and nonverbal behavior to comfort users. This paper presents the results of a study into the effect of variations in the virtual buddy's emotional expressiveness (no emotion, verbal emotion only, nonverbal emotion only, or verbal & nonverbal emotion) on user experience, the effectiveness of the support, and perceived social support. The results show that the virtual buddy is successful at conveying support. However, we found no statistically significant differences between conditions.

1 Introduction

Increasingly, embodied agents and robots are being employed to achieve certain effects in users, such as increasing exercise behavior [4], and increasing engagement in a virtual learning system [7]. In order to be able to achieve the beneficial outcomes these companion, coaching and pedagogical agents aim for, they need to behave as social actors. Social actors display and, to some extent, recognize social cues, and show appropriate verbal and nonverbal behavior [12].

The goal of our research is to understand how ECAs can provide social support. Social support refers to communicative attempts to alleviate the emotional distress of another person [5]. We are particularly interested in endowing ECAs with the emotional skills required to comfort users. To this end, we implemented an empathic virtual buddy that uses verbal and nonverbal strategies employed by people to comfort others. In order to be able to provide social support, a context of emotional distress is required. The application domain of the virtual buddy is cyberbullying, that is, bullying through electronic communication devices. Research shows that cyberbullying has a high impact on victims [9], making it a suitable test environment for the virtual buddy. We would like to emphasize that our research is focused on designing supportive interactions between ECAs and users. Our research objective does not include evaluating the buddy's suitability or effectiveness as a tool against cyberbullying.

The goals of the study presented in this paper are 1) to get more insight into how social support can be conveyed by conversational agents, and 2) to

measure the user experience of the virtual buddy proof of concept system. User experience refers to "a person's perceptions and responses that result from the use or anticipated use of a product, system or service" [1]. Poorly designed user interfaces may cause confusion and frustration [3]. These negative emotions may block the positive emotions the virtual buddy aims to evoke. Therefore, we assume that an acceptable level of user experience is required for a user to experience and be able to benefit from the social support communicated by the virtual buddy.

This paper is organized as follows. The next section describes the virtual buddy proof of concept system. In section 3, we explain the online survey used to conduct the study. The results are presented in section 4. In section 5 the results are discussed. Section 6 reviews related work on embodied agents. Finally, in section 7, we present our conclusions.

2 The Virtual Buddy

Figure 1 shows a screen shot of the proof of concept empathic virtual buddy. The user communicates with the buddy by selecting predefined response options. In order to understand, comfort and suggest actions to the user, the virtual buddy combines a conversation and an emotion model. The conversation model specifies the structure and contents of the conversation (see [14] for more details). In the current implementation, the conversation is scripted.

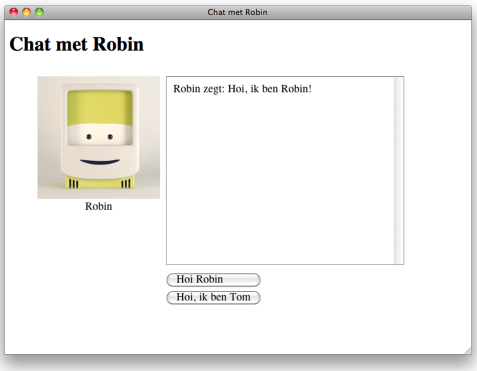


Fig. 1: Screen shot of Robin, the empathic virtual buddy proof of concept system.

The emotion model determines when the virtual buddy expresses sympathy, compliments or encourages the user. It is based on the OCC model of emotions [10]. In OCC, emotions are conceptualized as responses to events, agents, and objects. The OCC model specifies eliciting conditions for all emotion types.

The virtual buddy's emotion model is depicted in figure 3. In the model, response options are interpreted as actions or events. An action or event triggers an OCC emotion type, that is expressed both verbally and nonverbally. In the current implementation, the buddy's emotional state ranges from sad to happy. Figure 3 shows the facial expressions the virtual buddy displays for each emotional state it is capable of expressing (left to right: sadness, medium sadness, neutral, medium happiness, happiness). If a response option triggers a negative emotion, the buddy displays sadness and provides a sympathetic remark, and if a response option triggers a positive emotion, the buddy displays happiness and either provides a sympathetic remark, encourages, or compliments the user. What supportive strategy is used, depends on the response option selected; for example, if a response option refers to a praiseworthy action performed by the user, the buddy compliments the user.

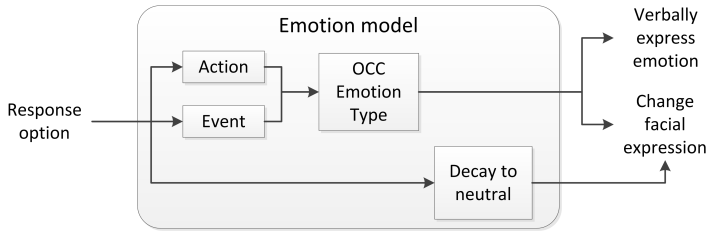


Fig. 2: The virtual buddy's emotion model.

Fig. 3: The virtual buddy's emotional states (left to right: sadness, medium sadness, neutral, medium happiness, happiness).

Not all response options trigger emotions. If a response option does not trigger an emotion, the current emotional state is decayed to neutral (sadness to medium sadness, and medium sadness to neutral). Next, the buddy's facial expression is updated to reflect the current emotional state. When uttering non-emotional messages, the buddy's emotional state also decays to neutral.

In addition to expressing sympathy, encouraging, and complimenting the user, the virtual buddy also gives advice and explains how to execute that advice (teaching).

3 Method

The goal of this study is to explore to what extent verbal and/or nonverbal expression of emotions contributes to the perceived effectiveness of the support provided by the virtual buddy and how these variations in emotional expressiveness affect user perceptions of social support. Additionally, since we assume that an acceptable level of user experience is required to be able benefit from interaction with the virtual buddy, a secondary goal of this study was to measure the user experience of the virtual buddy system.

For the experiment, the virtual buddy was embedded in an online survey. It had four modes of behavior, corresponding to four experimental conditions: 1) the buddy did not express emotions (control condition; No-EM), 2) the buddy expressed emotions by changing its facial expression (nonverbal condition; NV-EM), 3) the buddy expressed emotions verbally (verbal condition; V-EM), and the buddy expressed emotions both verbally and nonverbally (verbal and nonverbal condition; NV&V-EM). The virtual buddy's embodiment was displayed in all conditions. The experiment was set up using a between subjects design; participants were randomly assigned to one of the four conditions.

Before involving the virtual buddy's actual target audience (i.e., children aged 10–14), we decided to perform an experiment with university students. Participants were recruited by e-mail and through social media. The survey was completed by 100 students from different universities in the Netherlands. There were 25 participants in each condition. Of the 100 participants, 32% were female; the average age was 19.5 (SD=2.0).

Interaction with the virtual buddy was based on a fictitious scenario. The scenario tells the story of Tom, a 14-year-old boy that is verbally abused and threatened by a classmate. In the scenario, the buddy is introduced as a computer program that provides support to cyberbullying victims Tom found online. Participants were asked to take Tom's perspective during the interaction.

To capture different aspects of interacting with the virtual buddy and its supportive capacities several measures were included in the survey:

- **User Experience**: User experience was measured by the AttrakDiff 2 questionnaire [6]. AttrakDiff consists of four scales: Pragmatic Quality (PQ), Hedonic Quality-Identity (HQI), Hedonic Quality-Stimulation (HQS), and Attractiveness (ATT). Each scale consists of 7 semantic differentials on a 7-point scale. PQ refers to the utility and usability of products. HQI refers to the identity that is communicated by using certain products. HQS refers to personal development (e.g., development of new skills) triggered by stimulating products. ATT refers to the overall evaluation of the perceived qualities of a product.
- **Effectiveness of the Support**: Participants were asked to indicate on a 9-point scale how they think Tom feels (well-being; 1=feeling bad, 9=feeling good) and how severe they think Tom's problem is (perceived burden of the problem; 1=the problem is not severe, 9=the problem is severe) prior to interacting with the virtual buddy and after the conversation is completed.

– **Social Support**: Users' perception of social support was measured using a questionnaire containing 7 Likert items on a a 7-point scale (1 = completely disagree and 7 = completely agree). The questionnaire is listed in table 1.
– **Open Feedback**: Participants were asked *How can we improve the emotional support provided by Robin?* and *Do you have other suggestions to improve Robin?*

Item	Statement
Support attempt	Robin tried to cheer Tom up
Perceived support	During the conversation, Tom felt supported by Robin
Understood problem	Robin understood Tom's problem
Understood emotions	Robin understood what Tom was feeling
Compassion	Robin was compassionate with Tom
Advice general	Robin's advice is applicable
Advice situation	Robin's advice is applicable in Tom's situation
Persuasion	If I were Tom, I would follow Robin's advice

Table 1: The social support questionnaire (Tom refers to the main character in the scenario; Robin is the virtual buddy).

4 Results

We examined whether the buddy's emotional expressiveness (no emotion, verbal emotion only, nonverbal emotion only, or verbal & nonverbal emotion) affected participants' user experience, the effectiveness of the support, and/or perceived social support.

4.1 User Experience

User experience was measured by the AttrakDiff 2 questionnaire that consists of four scales: Pragmatic Quality (PQ), Hedonic Quality-Identity (HQI), Hedonic Quality-Stimulation (HQS), and Attractiveness (ATT). Figure 4 shows the average scores of PQ, HQI, HQS, and ATT for each condition. The average scores of HQI and HQS are close to 4 (the 'neutral' score); $4.47 < HQI < 4.61$ and $4.15 < HQS < 4.39$. PQ and ATT are slightly higher; $5.16 < PQ < 5.33$, and $4.99 < ATT < 5.25$. We conclude that the user experience provided by the virtual buddy is acceptable and does not hamper the provision of social support.

Oneway between subjects ANOVA was conducted to compare the effects of variations in the virtual buddy's emotional expressiveness on PQ, HQI, HQS, and ATT. There were no statistically significant differences between the four conditions; PQ $F(3, 96) = 0.585$, $p = 0.63$, HQI $F(3, 96) = 0.176$, $p = 0.91$, HQS $F(3, 96) = 0.459$, $p = 0.71$, and ATT $F(3, 96) = 0.708$, $p = 0.55$. These results indicate that the buddy's emotional expressions do not contribute to the user experience.

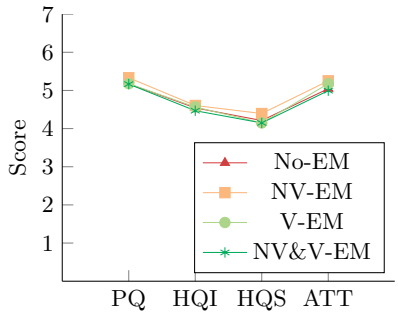


Fig. 4: Average scores for AttrakDiff scales PQ, HQI, HQS, and ATT.

4.2 Effectiveness of the Support

A mixed between-within subjects ANOVA was conducted to assess the impact of four levels of emotional expressiveness of the virtual buddy on participants' scores for well-being and perceived burden of the problem before interacting with the buddy and after interacting with the buddy. The results for well-being and perceived burden of the problem were similar. There were no significant interactions between emotional expressiveness and well-being, or between emotional expressiveness and perceived burden over time; $F(3, 96) = 0.298$, $p = 0.827$ and $F(3, 96) = 0.654$, $p = 0.583$ respectively. However, there were substantial main effects for well-being and perceived burden over time; $F(1, 96) = 344.12$, $p < .0005$ and $F(1, 96) = 24.203$, $p < .0005$, with all four groups reporting an increase in well-being after interacting with the virtual buddy and a decrease in perceived burden of the problem. There were non-significant main effects of the buddys expressiveness, $F(3, 96) = 0.132$, $p = 0.941$ for well-being and $F(3, 96) = 0.372$, $p = 0.774$ for perceived burden. This means there was no difference in effectiveness of increasing well-being or decreasing perceived burden of the problem between the four levels of emotional expressiveness. The results are depicted in figure 5.

4.3 Perceived Social Support

We also examined whether the buddy's emotional expressiveness affected perceived social support. Oneway between subjects ANOVA was conducted to compare the effects of variations in the virtual buddy's emotional expressiveness on the social support ratings. There were no statistically significant differences between the four conditions (Support attempt: $F(3, 96) = 0.431$, $p = 0.731$; Perceived support: $F(3, 96) = 0.433$, $p = 0.730$; Understanding of problem: $F(3, 96) = 0.323$, $p = 0.809$; Understanding of emotions: $F(3, 96) = 0.235$, $p = 0.872$; Compassion: $F(3, 96) = 2.255$, $p = 0.087$; Advice general: $F(3, 96) = 1.294$, $p = 0.281$; Advice situation: $F(3, 96) = 0.231$, $p = 0.874$; Persuasiveness: $F(3, 96) = 1.794$, $p = 0.162$). The results are depicted in figure 6.

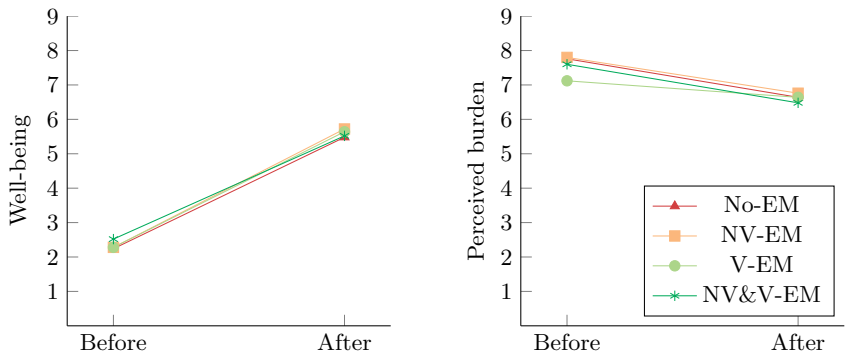


Fig. 5: Well-being and perceived burden of the problem before and after interaction with the virtual buddy.

The average perceived social support scores were generally high, especially for for items referring to information support (Advice general, Advice situation, and Persuasion); 5.6 < average scores < 6.4. In contrast, social support ratings for emotional support (Understood emotions, and Compassion) were lowest; 4.2 < average scores < 5.2. These results raise the question to what extent expressing emotions contributes to or is required for users' perception of social support.

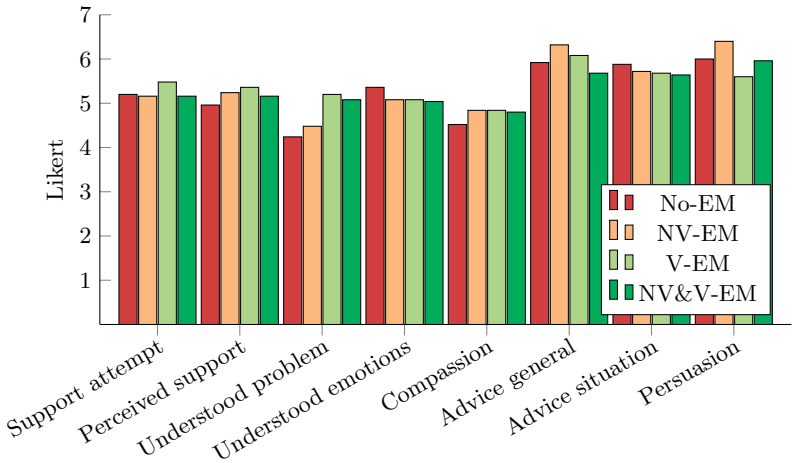


Fig. 6: Average social support ratings.

4.4 Open Feedback

At the end of the survey, participants were invited to suggest improvements for emotional support and other improvements. In total, 93 of the 100 participants provided one or more remarks. Many participants came up with concrete suggestions on how to improve the experience of emotional support. These suggestions are listed in table 2 together with the number of participants from each condition that made them.

Half of the participants in the no emotion condition that left feedback (12 of 25 participants) suggested to add supportive verbal utterances to the conversation. As formulated by one of the participants in the No-EM condition:

> *In addition to suggesting a practical solution, Robin should show compassion and say nice things that may not directly resolve the situation, but give the impression that Robin is sympathetic and cares about the fact that its conversation partner is being bullied.* (P47)

Also, 6 participants in the nonverbal emotion only recognized verbal support was missing and suggested to include supportive remarks. Additionally, 3 of 25 participants in the nonverbal and verbal emotion condition suggested to add more supportive verbal expressions. Remarkably, while many participants in the no emotion condition recognized verbal support was missing, this did not lead to significant differences in perceived social support scores between the different conditions (see figure 6).

	No-EM	NV-EM	V-EM	NV&V-EM
Add verbal expressions	12	6	0	3
Add facial expressions	1	1	5	0
Facial expression mismatch	0	1	0	2
Inappropriate verbal expressions	2	0	4	3
Add other support types	8	8	4	7
Left feedback	24	22	23	24
Total participants	25	25	25	25

Table 2: Participants' suggestions for improving the experience of emotional support.

Five participants in the verbal emotion only condition suggested to have the virtual buddy change its facial expression during the conversation. Three participants, one in the nonverbal emotions condition and two in the verbal & nonverbal emotions condition, noticed emotion mismatches. For example, one participant thought Robin's neutral expression was too cheerful:

> *Robin should look less happy; he was smiling when I told my story. That's rather tactless.* (P68;)

In total, nine participants stated that they felt discouraged by some of the messages conveyed by the virtual buddy. The large number of comments that suggest to increase the virtual buddy's emotional expressiveness indicate that emotional expressiveness is an important factor in the perception of support, even though this is not reflected in the social support scores.

Participants from all conditions suggested other types of support should be added to the conversation, such as explaining why bullies bully, that bullies sometimes randomly select a victim, that Tom is a good person despite what other people say, and that bullying can only be stopped by taking action.

Table 3 lists participants' feedback on the virtual buddy's technical limitations. As these limitations were the same in each condition, we only report the total number of participants that made some remark.

Remark	# participants
Negative about interface design	8
Positive about interface design	2
Negative about appearance of the virtual character	13
More human-like system	13
Typing instead of response options	6
More response options	7
Select multiple response options	11

Table 3: Technical limitations of the proof of concept system identified by participants.

Eight participants expressed dissatisfaction with the design of the interface, while two participants were positive about the design. Thirteen participants criticized the virtual character's appearance; they thought it was too robot-like, and/or static. In addition, thirteen participants suggested to make the system (and not just the virtual character) more human-like.

Another recurring topic in the feedback were the response options. Six participants asked for the possibility to type responses instead of selecting them. Seven participants wanted to more response options to choose from. Finally, eleven participants wanted to be able to select multiple response options instead of just one.

Many participants suggested to improve the experience of emotional support by increasing the virtual buddy's emotional expressiveness. In the verbal and nonverbal emotions condition, the condition in which participants interacted with the most emotionally expressive buddy, there also were participants that suggested to increase the amount of emotional feedback. Additionally, the technical limitations identified by the participants suggest that the system used in the experiment may have been too limited. Even though shortcomings in the virtual buddy's behavior were recognized by many participants, this did not result in lower social support ratings.

5 Discussion

While our study demonstrated that the virtual buddy is able to comfort users, we found no significant differences between the four conditions in user experience, effectiveness of the support, and perceived social support. Additionally, the average perceived social support ratings were relatively high (> 4.24). In this section, we explore explanations for the lack of significant differences between conditions and the high social support ratings.

Nass and Reeves' media equation states that people apply social rules from human-human interaction to computers (and other media) that provide (simple) social cues [13]. Feedback from participants suggest that the social cues provided by the virtual buddy proof of concept system may have been too simple. However, a pilot study with an earlier version of the virtual buddy system demonstrated that children recognize and accept simple social cues like the ones used in the current study [15]. Nevertheless, repeating the experiment with a more advanced emotion model and/or more natural facial expressions may result in statistically significant differences between conditions.

The lack of significant differences between the conditions might also be (partially) explained by the differences between the virtual buddy's behavior in the four conditions; these may have been too small. The buddy's behavior differed in how emotions were expressed. Apart from the control condition in which no emotions were expressed, the amount and valence of the emotions were the same for all conditions (depending on the response options selected by the user). Some participants remarked that the total number of emotions should be increased.

The differences between conditions may also have been too small in the sense that expressing emotions may not be crucial to experience support during the conversation, even though the number of suggestions by participants to increase the virtual buddy's emotional expressiveness indicates that it is an important factor for the perception of support. The virtual buddy uses a variety of strategies to convey support; in addition to expressing emotions, these strategies include the conversation structure, and providing information (advice and teaching). Also, the fact that many participants suggested other ways in which the virtual buddy could provide support to cyberbullying victims indicates that there are more factors that affect the perception of support than 'just' expressing emotions. More research is required to identify these factors, find ways to incorporate them into the conversation, and assess how they affect perceived support.

Even though participants from all conditions were very well able to point out weaknesses in the virtual buddy's behavior, this critical attitude was not reflected in the perceived social support scores. The average scores were relatively high. These high scores could have been caused by socially desirable behavior triggered by the social relevance of cyberbullying as application domain.

6 Related Work

The virtual buddy is an example of an application of embodied agents for creating a particular emotional experience, in our case the experience of social sup-

port. This section briefly reviews related work on embodied agents that trigger emotional responses.

A related project in the bullying domain is FearNot!. FearNot! is an Intelligent Virtual Environment (IVE), where synthetic characters act out bullying scenarios [11]. The goal of the project was to create virtual agents that elicited empathy by displaying believable social and emotional behavior. User tests confirmed that the agents were able to establish empathic relations with users.

Other virtual agents that try to evoke certain emotional responses are pedagogical agents. A study conducted by Arroyo et al. shows that the interacting with a pedagogical agent that provides emotional and motivational support in an Intelligent Tutoring System for mathematics improved affective learning outcomes; users of the pedagogical agents reported less frustration and increased confidence compared to users that did not interact with with an agent [2].

The emotional experience companion agents strive for is engagement. In particular, the goal of companion agents is to keep user engaged for multiple interactions over longer periods of time. Related work on a robotic chess companion for children shows that keeping users engaged over multiple interactions is challenging; participants of the study lost interest in the companion robot over the course of the five weeks they played against the robot [8].

7 Conclusion

The goals of the study presented in this paper were 1) to determine to what extent verbal and/or nonverbal expression of emotions contribute to the effectiveness of social support by an conversational agent, and 2) to verify the user experience of the virtual buddy proof of concept system does not hamper the provision of social support. The results show that the user experience of the virtual buddy is acceptable; and, therefore, does not impede the virtual buddy's potential for providing social support. It was also shown that the social support expressed by the virtual buddy is effective. Additionally, perceived social support was generally high.

We found no significant differences between conditions for user experience, effectiveness of the support, and perceived social support. Therefore, we conclude that emotions expressed verbally and/or nonverbally by the virtual buddy proof of concept system do not contribute to the experience of social support in the context of our cyberbullying scenario. However, the large number of participants suggesting to increase the virtual buddy's emotional expressiveness in order to improve emotional support, indicate that this is an important factor in the perception of support.

The feedback from participants indicated some important limitations of the virtual buddy proof of concept system. We plan to further investigate these limitations and whether social support is conveyed by the virtual buddy in a qualitative evaluation of the system by domain experts and the target audience.

Acknowledgements This work is funded by NWO under the Responsible Innovation (RI) program via the project 'Empowering and Protecting Children and Adolescents Against Cyberbullying'.

References

1. Ergonomics of human-system interaction – part 210: Human-centred design for interactive systems. International Organization for Standardization, 2010.
2. I. Arroyo, B.P. Woolf, D.G. Cooper, W. Burleson, and K. Muldner. The impact of animated pedagogical agents on girls' and boys' emotions, attitudes, behaviors and learning. In *Advanced Learning Technologies (ICALT), 2011 11th IEEE International Conference on*, pages 506–510, 2011.
3. R. Baecker, K. Booth, S. Jovicic, J. McGrenere, and G. Moore. Reducing the gap between what users know and what they need to know. In *Proceedings on the 2000 conference on Universal Usability*, pages 17–23. ACM, 2000.
4. T.W. Bickmore and R.W. Picard. Establishing and maintaining long-term human-computer relationships. *ACM Trans. CHI*, 12(2):293–327, 2005.
5. B.R. Burleson and D.J. Goldsmith. How the comforting process works: Alleviating emotional distress through conversationally induced reappraisals. In P.A. Andersen and L.K. Guerrero, editors, *Handbook of Communication and Emotion: Research, Theory, Applications, and Contexts*, pages 245–280. Academic Press, 1998.
6. M. Hassenzahl, M. Burmester, and F. Koller. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Mensch & Computer 2003: Interaktion in Bewegung*, pages 187–196. B. G. Teubner, 2003.
7. T.-Y. Lee, C.-W. Chang, and G.-D. Chen. Building an interactive caring agent for students in computer-based learning environments. In *Proceedings of the 7th IEEE Int. Conf. on Advanced Learning Technologies, ICALT 2007*, pages 300–304, 2007.
8. I. Leite, C. Martinho, A. Pereira, and A. Paiva. As time goes by: Long-term evaluation of social presence in robotic companions. In *Proceedings of the 18th IEEE International Symposium on Robotics*, pages 669–674, 2009.
9. S. Livingstone, L. Haddon, A. Görzig, and K. Ólafsson. Risks and safety on the internet: the perspective of European children: full findings. http://eprints.lse.ac.uk/33731/, 2011.
10. A. Ortony, G.L. Clore, and A. Collins. *The cognitive structure of emotions*. Cambridge Univ. Press, 1988.
11. A. Paiva, J. Dias, D. Sobral, R. Aylett, S. Woods, L. Hall, and C. Zoll. Learning by feeling: Evoking empathy with synthetic characters. *Applied Artificial Intelligence: An International Journal*, 19(3):235–266, 2005.
12. H. Prendinger and M. Ishizuka. Designing and evaluating animated agents as social actors. *IEICE TRANS. on Information Systems*, E86-D(8):1378–1385, 2003.
13. B. Reeves and C. Nass. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge Univ. Press, 1996.
14. J.M. van der Zwaan, V. Dignum, and C.M. Jonker. A conversation model enabling intelligent agents to give emotional support. In W. Ding, H. Jiang, M. Ali, and M. Li, editors, *Modern Advances in Intelligent Systems and Tools*, volume 431 of *Studies in Computational Intelligence*, pages 47–52. Springer, 2012.
15. J.M. van der Zwaan, E. Geraerts, V. Dignum, and C.M. Jonker. User validation of an empathic virtual buddy against cyberbullying. *Stud Health Technol Inform*, 181:243–7, 2012.

Feel Connected with Social Actors in Public Spaces

Mathias Funk, Duy Le and Jun Hu

Department of Industrial Design, Eindhoven University of Technology
{m.funk, l.duy, j.hu}@tue.nl

Abstract. Public spaces changed in the last couple of years: abundant use of smart phones and other digitally connecting devices draw peoples' attention away from physical neighbors to virtual peer groups. So, will masses of isolated people be the desired future? We think that current technology involving new display technology and, therein, new interaction possibilities hint at a different future vision. Social connectedness and bonding are important aspects of public interaction that are often overlooked, but can be initiated and supported by technology. This paper reports on research investigating how a publicly displayed application can improve social connectedness by acting in a socially accepted way. Blobulous is a novel interactive installation that interacts with participants through projected avatars, which react to the participants' movement and body signals. A functional prototype was implemented and evaluated.

Keywords: System Design Social Connectedness, User Experience, Interactive Displays, System Design, Avatars, Computers as Social Actors

1 Introduction

Public displays in public spaces have been means in addressing multiple people and the same time in aiming at engaging bystanders, people passing by and others for a certain cause. While the use for advertisements, entertainment and promotion is quite far-spread and has been around for decades already, the usual modus operandi of a single display is to engage people as a single person in a dedicated 1:1 message. A second drawback is the limited interaction space for people "using" a public display: messages are mostly unidirectional and there is a little that a person can actually do to be engaged in a richer interaction than simple information broadcast. At the same time, people are currently more and more "distracted" by smart phones, mp3 players and other personal devices in their immediate environment, that are (1) open for bi-directional interaction with direction manipulation, (2) offer personalized content and functionality, and (3) allow users to pursue activities of their interest.

How can public display compete with this? One possibility is to leverage the social situation in the public space, the unique set of people, who potentially follow a shared interest (which might have brought them to the location). While individual, personalized devices tend to isolate people in their dedicated spheres of personal activities and

content, public displays can inspire connectedness and support social bonding between people in the same space.

To achieve this the public installation has to act as a mediating entity, a social actor that addresses multiple people at the same time and increases the level of social connectedness among them. The main challenge is how to realize an interactive installation, which can act in a socially acceptable way and participate in a social multi-user setting. Partly, the motivation of this research is also to investigate whether computers (controlling the public installation) can indeed act as a social actor and improve social connectedness. The concept "social actor", in general ICT uses, was developed into a conceptualization model through a series of empirical studies. There are five dimensions in the conceptualization of a social actor [1]:

- Affiliations: organizational and professional relationships that connect an organization member to industry, national and international networks.
- Environments: regulated practices, associations and locations that define organizational actions.
- Interactions: information resources and media exchange that organization members mobilize as they engage with members of affiliated organizations.
- Identities: representation of the "self" and profiles of organization members as individual and collective entities.
- Temporalities: socially constructed segments of time that elicit and shape the interactions of an individual in response to the expected affiliates.

Social actorship in the context of HCI is an indicator of computers in participating in social activities with humans. These activities focus on social connection and bonding, between humans and computers during socially interactive sessions. It covers from one-one to many-many relations. Social actorship can be represented through elevating levels:

- Attention and awareness
- Information
- Social acceptance
- Social bonding
- Social behavior

Some examples of social interactive sessions are:

- Visiting or passing by a public place,
- Well-being persuasion,
- Supporting elderly,
- Social engaging with children with autism.

In order to apply this definition in designing social actors, design guidelines were derived from the definition of social actorship. Later, the design of the prototype will follow these design guidelines:

1. Users' awareness and attention are considered to be the first level relationship between humans and computers. Factors can be used to attract attention: attractiveness, suddenness, surprise, and confusion.

2. Users need to be provided with information depending on the contexts of use. Information can be presented at concrete or abstract levels. The context is important

for maintaining the connection between computers and humans otherwise they will lose interest in the design.

3. Interactive sessions between users need to be supported by the design in order to gain social acceptance from the users. Users need to understand that the artifacts of the design are connected with the in-context activities.

4. Social bonding between humans and humans, and between humans and computers, should be stimulated after being socially accepted.

5. Stimulating social behaviors can be achieved after going through the above four guidelines.

2 Related Work

Social connectedness also stands out to be a very important psychological feeling that links to personal health and well-being [2].

In the field of HCI, computers are considered to be able to handle social tasks and tend to be treated like humans [3]. There is a growing community around public projection and large-scale installations, and social interaction of their users, which is picked up by user-dedicated devices such as RFID tags and mobile phones [4-6]. In the case of *Blobulous*, an interactive installation to be introduced in the next section, the large (possibly public) projection of abstract avatars is combined with bio-signals, i.e., the heart rate, which other research also consider as a reliable and effective means of communication between people [7, 8]. With the system we explore the possibilities in utilizing related technologies to collect information from wearable objects for social interaction in public spaces [9].

This work is an extended version of a paper for the CHI'13 workshop on Experiencing Interactivity in Public Spaces [10].

3 Blobulous System

Blobulous is a novel interactive installation (see **Fig. 1** for example settings and **Fig. 2** for system overview) that interacts with participants through projected avatars in public spaces, which react to the participants' movement and body signals. Blobulous uses a large projection to show abstract avatars, blobs of dots – therefore the name "Blobulous" – one for each participant and moving around slowly. The movement of the avatars is connected to the participant's movement in the space in front of the projection. The second mapping involved in the installation is from a participant's heart rate to the color of his or her avatar. The mapped colors range from blue (cold, low engagement) to red (warm, high engagement).

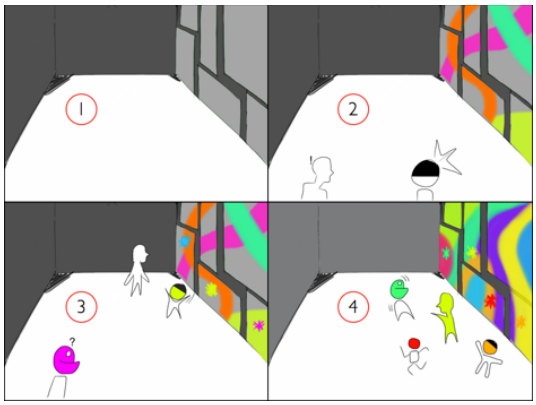


Fig. 1. Example space for using Blobulous

Blobulous is designed considering the proposed definition of social actorship in the context of HCI. Blobulous aims at stimulating physical and mental connections between humans in order to influence social connectedness by using physiological data from users. The concept is designed for all ages as long as they are having activities in a shared space. It is about creating a visualization system shown on a public display, which generates visuals according to the users' biological data and movements. The concept aims to serve three user groups:

- Visitors of exhibitions
- People sharing a public space
- Employees of a company

Biosignals are proposed to integrate in the design of Blobulous to enhance its social actorship. So a physiological model needs to be designed to help improve the feeling of social connectedness between people. Considering a list of design guidelines for social actors, some design assumptions or goals of the prototype were made:

- Blobulous has the ability to draw great attention from people.
- Blobulous raise curiosity to people and trigger discussion or communications, and interaction accordingly.
- Systems with a physiological connection between human and system may improve social connectedness.

With abstract visuals (avatars) and avatars' behaviors, the system aims to improve social connectedness among people in the same space. It provides a bird-eye view of the context to help people be aware of the current social situation unobtrusively.

Heart rate (HR) and heart rate variability (HRV) can indicate people's moods, emotions and activities [8]. At a prototype level, it is feasible to collect HR data not HRV data. Even short-term HRV data analysis requires a five-minute recording in a steady-state physiological condition. In this concept, people have their normal activities in a public space. So it would be very rare that someone would rest or stand still for 5 minutes to provide accurate HRV data to the prototype. Therefore, we use only HR data for the prototype and its evaluation in an explorative study.

An abstract representation is chosen to be the avatars that act as social actors. It is designed to make users believe they are social actors and act accordingly. They mimic users' movements and change colors according to their heart rate. Their shapes and movements depend on people's ways of movement, speeds of movement and heart rate. In other words, they are influenced by the way people move and behave in the current context. Particles' colors, a spectrum from shades of green, red or blue, is mapped with a healthy heart rate range of the target group, from 60 to 150 BPM. Fig. 2 shows an example of the visuals in the upper part, which are unique depending people and context of use.

The final version of the Blobulous system consists of four parts:

(1) Wireless heart rate sensors capture and send heart rate data from users to a central instance. Three of these sensors can be seen at the bottom left of figure 2 on a charging board. The sensors are custom-made 3D-printed enclosures that can be worn on a necklace (see mid bottom of Figure 2), and which provide the housing for an Arduino nano with a Zigbee[1] unit. Through the necklace, a HR sensor is guided towards the ear of the participant. This ensures correct placement as well as certain robustness in case the participant decides to dance or otherwise move rapidly.

(2) A central instance, including a receiver and a visual program, receives data from users' sensors and, after processing this data, it derives avatar behaviors represented as visuals on the projected screen. The instance is realized as a Processing sketch (a program written in Processing, a Java based programming language and environment) that is running in Presentation (full screen) mode.

(3) A projector connected to the central instance will simply project the screen contents on a large display.

(4) A Zigbee network handles the communication between sensors and the central instance.

This system allows for several rapidly moving users and an arbitrarily large display. The setup is also quite independent from the display system, as the interaction between the social actor, Blobulous, and the users will be controlled entirely by their position and HR sensor data. This makes the system quite portable and will hopefully allow for more future evaluations in real-life settings.

[1] http://www.zigbee.org/

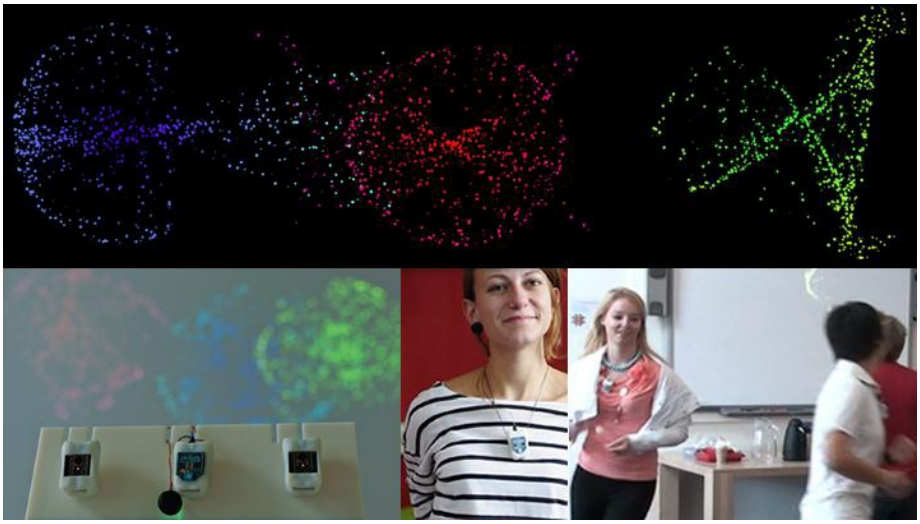


Fig. 2. Overview of system components

4 Evaluation

The objective of evaluating the Blobulous system is to show an improvement of social connectedness among participants and the attractiveness of the installation. In this work the social connectedness will be in the focus. The study was planned to take place in a living lab environment to yield most reliable and realistic results. The results about attractiveness have been evaluated and reported in [11]: 21 participants (14 male, 7 female) in 7 randomly selected groups were asked to experience and interact with the Blobulous system and, later on, they were asked to use the AttrakDiff [12] instrument for rating the system. With the help of pairs of opposite adjectives, they could indicate their perception of the system. The installation was rated as fairly "self-oriented". It provides the user with identification and is generally considered attractive by the participants, although they were aware that they were evaluating a prototypical system. Attractiveness is certainly an important aspect of a system aimed at inducing social connectedness among its users. Although systems are imaginable that operate as a "common enemy" with low attractiveness and thus united its users, this was not an option for this line of research.

Social connectedness is measured by means of a questionnaire that has been derived from Social Connectedness Scale Revised (SCS_R) questionnaire [13]. The two research hypotheses are:

- Hypothesis
    - Blobulous has the effect on an individual of feeling socially connected to others (H1a).
    - Blobulous improves the feeling of social connectedness of people (H1b).

- Null Hypothesis:
  - Blobulous has no effect on an individual of feeling social connectedness to the others (H0a).
  - There is no improvement on the feeling of social connectedness of people from Blobulous (H0b).

4.1 Experiment Setup

In order to evaluate the feeling of social connectedness of people while interacting with each other, it is better to include a group dynamics factor in the evaluation. 21 (14 male, 7 female) participants were recruited online and randomly divided into 7 groups according to their time preference, taking into account the balance of gender, age, and background. So, in most of the groups, participants did not know each other before the experiment. Users' backgrounds were distributed to Industrial Design (7), Electrical Engineering (4), Computer Science (3), Automotive/Logistics (3), Biomedical (2), Architecture (1), and Business (1).

Before coming to the experiment, participants were requested to answer the questionnaire to measure their initial level of social connectedness. During the experiment, this measure is repeated at the end of sessions 1 and 2. In the experiment, participants as a group were asked to perform three sessions: the first two sessions were planned to study social connectedness, the final one is to see how people can interact with Blobulous. Experiments were carried out following the two protocols shown in Table 1 to avoid a direction effect in the evaluation.

Table 1. Evaluation protocols

	Protocol 1	Protocol 2
Session 1	A	B
Session 2	B	A
Session 3	Brainstorm & Demo	

A: Random Blobulous
B: Interactive Blobulous

In both conditions A and B (Tab. 1), participants were asked to watch and explore the visuals projected on the wall (Fig. 3a) while wearing the sensor (Fig. 3b) and then have a short discussion about what they perceive from the visuals. Heart rate data was streaming automatically by the prototype while movement data was manually controlled via an Apple iPad using touchOSC [14] (Wizard of Oz) (Fig. 3c).

Only afterwards, in the demo session, participants were explained details about the functionality of *Blobulous,* and then asked to come up with some ideas and try to demonstrate the ideas together with *Blobulous*. All sessions were recorded for later video analysis. The experiment room was prepared with a large display on the wall,

an interaction space in front of the display, and an experiment control area (depicted at the bottom of Fig. 3).

Fig. 3. Experiment room with a) projection screen, b) heart rate sensors, and c) central control.

4.2 Methodology

A video analysis was proposed to follow up the social connectedness test. The video analysis was to investigate and capture social behaviors that might link to social connectedness but could not be captured by questionnaires. Therefore, the evaluation was carried out in two steps:

Firstly, the Social Connectedness Scale Revised (SCS_R) questionnaire [13] was chosen to evaluate the level of social connectedness of participants in this study. SCS-R consists of 20 items (10 positive and 10 negative). The negatively worded items are reverse scored and summed with the positively worded items to create a scale score with a possible range from 20 to 120. Then, the mean score with a possible range from 1 to 6 is calculated by dividing the total scale score by 20 (or 20 scale items). A higher score on the SCS-R indicates a stronger feeling of social connectedness.

Secondly, the video analysis was carried out to check the feeling of social connectedness in conditions A and B. An observation scheme with behaviors and scores was developed to compare between conditions A and B.

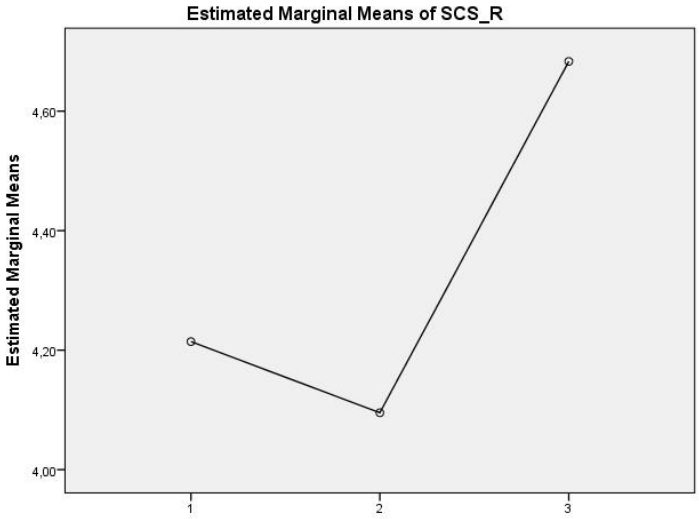


Fig. 4. ANOVA repeated measures (SPSS). 1. Before the experiment; 2. After random Blobulous; 3. After interactive Blobulous

4.3 Results

SCS-R was used to study if there is an improvement or difference in the feeling of social connectedness of participants while interacting with the system. A repeated measures ANOVA with a Greenhouse-Geisser correction determined that the mean SCS_R score differed statistically significantly between different conditions ($F(1.484, 8.107) = 3.791$, $p < 0.046$). Post hoc tests using the Bonferroni correction revealed that there is a slight reduction in the SCS_R score when bringing people from their own setting to a social setting or testing environment ($M = 4.21$ vs. $M = 4.09$, respectively), which was not statistically significant ($p = 1$). However, the SCS_R score had been improved after the interactive session with *Blobulous* ($M = 4.68$), which was statistically significantly different to the random session without *Blobulous* ($p = 0.002$) (see Fig. 4 and Table 2). Therefore, it can be concluded that the *Blobulous* prototype elicits a statistically significant improvement in SCS_R score or the feeling of social connectedness of people but only in certain social contexts.

The internal reliabilities on the SCS_R questionnaire from pre-test, random and interactive condition had been found to be good ($\alpha = 0.936, 0.756, 0.751$, respectively). Strangely, there were slight drops in the alpha values between the testing and pre-test conditions. This can have resulted from the fact that the pre-test participants were at their own places while answering the SCS_R questionnaire, but during the test they were in a controlled room.

Table 2. Pairwise comparison of SCS_R scores between the three conditions (1. *Before* the experiment; 2. After *random* Blobulous; 3. After *interactive* Blobulous).

		Mean Difference	Std. Error	Sig.
before	random	,119	,264	1,000
	interactive	-,469	,249	,223
random	before	-,119	,264	1,000
	interactive	-,588*	,146	,002
interactive	before	,469	,249	,223
	random	,588*	,146	,002

The video analysis consists of seven steps:

1. Conduct qualitative study based on demo videos to categorize users' behaviors while interacting with the system (not shown)
2. Divide the SCS-R questionnaire into groups of behaviors (see Table 3):

Table 3. Predicted behaviors from SCS_R questionnaire

Negative	Positive	General	Context	Behaviors
I catch myself losing a sense of connectedness with society.	I fit in well in new situations.	world	room with Blobulous	Interact with Blobulous
I feel like an outsider.	I feel comfortable in the presence of strangers.			Make sound
I feel disconnected from the world around me.	I am in tune with the world.			
I feel distant from people.	I feel close to people.	people	other participants	Turn head to someone
I don't feel related to most people.	I see people as friendly as approachable.			Staring at someone
I see myself as a loner.	I am able to connect with other people.			Reach to someone
I have little sense of togetherness with my peers.	I am able to relate to my peers.	friend	participants with	Walk to someone
I don't feel I participate with anyone or any group.	I find myself actively involved in people's lives.		same sex	Talk to someone
Even around people I know, I don't feel that I really belong.	I feel understood by the people that I know.		similar occupation	
Even among my friends, there is no sense of brother/sisterhood.	My friends feel like family.		similar education	

3. Combine 1 and 2 to derive an observation scheme for video observation (see Table 4):

Table 4. Combination of observed behaviors and predicted behaviors

Category (react to)	Behavior
Blobulous	Interact with Blobulous
	Make sound
People	Turn head to someone
	Lean toward someone
	Reach to someone
	Staring at someone
	Follow someone
	Mimick someone
	Stand next to someone
	Touch someone
	Walk to someone
	Talk to someone
	Do something with someone
	Tell someone to do something

4. Conduct two pilot observation sessions to revise and finalize the observation scheme (see Table 5):

Table 5. Observation scheme for the level of social connectedness in a social setting

Category	Behavior	Score
Individual	Turns head to someone	1
	Goes and stands next to someone	2
	Touches someone	3
Group	Does something with someone	2
	Talks	2
	Laughs	2
	Moderates an activity	3

5. Observe one random participant (first one on the left) in each video: 5 participants and 10 observation sessions.
6. For participants, a higher score means a higher feeling of social connectedness.
7. Compare random and interactive conditions to see if there is an improvement in the feeling of social connectedness.

The paired t-test is used to check whether the scores derived from observed social connectedness scale (SCS_O) are significantly different from random settings to in-

teractive settings. Normality test was conducted to check the assumption of the t-test that both variables are normally distributed. The results shows that the observed data of individuals is normally distributed ($p = 0.619$ and 0.807, respectively). In the t-test, $t(4) = -4.214$ and $p = 0.014$ (mean = -1.09; standard deviation = 0.58, standard err = 0.26)), which means there is a significant difference between the SCO score of random setting and interactive setting.

Considering results from both statistical tests, a repeated measure ANOVA and a paired t-test, there is a significant difference between the feelings of social connectedness, in general, between the two controlled settings (random and interactive).

5 Conclusions and Future Work

The *Blobulous* prototype was designed to act as a social actor, specifically to improve social connectedness between people. *Blobulous* draws great attention from users due to its colorful appearances and lively movements. It also raises social awareness between people while they are together and informs them about individuals' and the group's condition. With those effects, *Blobulous* makes people talk about it, about each other and sometimes they try to understand *Blobulous* and interact with it. As a system with a physiological connection between humans and computers, *Blobulous* has more impact on social interaction than one without physiological connection: The experiment results showed a significant difference in the level of social connectedness between the two testing conditions (random avatars and interactive, mapped avatars).

Most importantly, the study showed that while *Blobulous* was mediating social activities, peoples' feelings of social connectedness were improved significantly ($P = 0.002$ – one way ANOVA).

The system needs to be further developed with the ability to act independently but not only mimicking to do so, which was a pragmatic design choice in this study.

References

1. Lamb, R. *Alternative paths toward a social actor concept*. in *Proceedings of the Twelfth Americas Conference on Information Systems*. 2006.
2. Ottmann, G., J. Dickson, and P. Wright, *Social Connectedness and Health: A Literature Review.* 2006.
3. Reeves, B. and C.I. Nass, *The media equation: How people treat computers, television, and new media like real people and places*. 1996: Chicago, IL, US: Center for the Study of Language and Information; New York, NY, US: Cambridge University Press.
4. Rogers, Y. and H. Brignull. *Subtle ice-breaking: encouraging socializing and interaction around a large public display*. in *Workshop on Public, Community. and Situated Displays*. 2002.

5. Rukzio, E., S. Wetzstein, and A. Schmidt, *A Framework for Mobile Interactions with the Physical World.* Proceedings of Wireless Personal Multimedia Communication (WPMC'05), 2005.
6. Villar, N., et al., *Interacting with proactive public displays.* Computers & Graphics, 2003. **27**(6): p. 849-857.
7. Cwir, D., et al., *Your heart makes my heart move: Cues of social connectedness cause shared emotions and physiological states among strangers.* Journal of Experimental Social Psychology, 2011. **47**(3): p. 661-664.
8. Slovák, P., J. Janssen, and G. Fitzpatrick. *Understanding heart rate sharing: towards unpacking physiosocial space*. in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*. 2012. ACM.
9. Vlist, B.v.d., et al., *Semantic Connections: Exploring and Manipulating Connections in Smart Spaces*, in *2010 IEEE Symposium on Computers and Communications (ISCC)*. 2010, IEEE: Riccione, Italy. p. 1-4.
10. Le, D., M. Funk, and J. Hu, *Blobulous: Computers As Social Actors*, in *CHI'13 workshop on Experiencing Interactivity in Public Spaces (EIPS)*. 2013: Paris.
11. Hu, J., Le, D., Funk, M., Wang, F., Rauterberg, M., *Attractiveness of an Interactive Public Art Installation*, in *HCI International Conference*. 2013, Springer, Heidelberg: Las Vegas, NV, USA.
12. User Interface Design GmbH, *AttrakDiff Tool to measure the perceived attractiveness of interactive products based on hedonic and pragmatic quality*. 2012: http://www.attrakdiff.de/en/Home/.
13. Lee, R.M., M. Draper, and S. Lee, *Social connectedness, dysfunctional interpersonal behaviors, and psychological distress: Testing a mediator model.* Journal of Counseling Psychology, 2001. **48**(3): p. 310-318.
14. Hexler.net. *TouchOSC*. 2012; available from: http://hexler.net/software/touchosc.

Social Agency in an Interactive Training System

Norbert Reithinger and Ben Hennig

DFKI GmbH, Projektbüro Berlin,
Alt-Moabit 91c, D-10559 Berlin, Germany
norbert.reithinger@dfki.de

Abstract.
Interactive training systems often use avatars to depict an advisor that provides feedback on the exercise. In the framework of the SmartSenior project, which developed technologies for people with age related limitations, we realized an interactive trainer for stroke rehabilitation. The UI contained two avatars, one for the training person itself to provide feedback on her motion, and one for a physiotherapist, who guides the user through the exercises. In the study presented here, we looked especially at the social agency related aspects of this system. We tested the system using the AttrakDiff™ questionnaire and used the results to rate various aspects of social agentship.

Keywords: Interactive Training System, Social Agentship, Multimodal Interaction

1 Introduction

Interactive environments like games or training systems often use avatars that serve as communication partners in the flow of interaction. However, they are hardly explicitly developed with a focus on social agency.

In the context of the German SmartSenior[1] project we jointly developed different technologies to serve people with aged related limitations. With our partners Charité, Fraunhofer FOKUS, Nuromedia, Humotion, and Otto Bock we realized in the context of this project an interactive trainer – trainIT – for stroke rehabilitation [3,4]. During the development, the main emphasis was the clinical effectiveness of the training system. DFKI, the German Research Center for Artificial Intelligence, with its project office Berlin was responsible for the multimodal interface, and especially for the dialogic interaction.

The aspect of social agency of the system was only implicitly addressed at best. However, in the context of the EIT ICT Labs[2] activity "Computers as Social Actors" (CASA), we decided to conduct a separate study, especially looking into the hedonic

[1] http://www.smart-senior.de
[2] http://www.eitictlabs.eu

and pragmatic qualities that are good indicators for the social actorship of the system, independent of the initial target group of the system.

In the second section of this paper we give an overview of the trainIT system that we used for our usability test that is described in detail in section three. We use the AttrakDiff™ [3] questionnaire and website to measure the hedonic and pragmatic qualities of the system. In section four we present the categories for social actorship we agreed upon in the EIT ICT Labs CASA activity and rate our systems on the various dimensions.

2 Description of the system

2.1 System overview

The trainIT interactive training system integrates different sensor systems as well as multimodal input and output devices, controlled by a standard PC. To track the body movement we used in the tests a Kinect-based system, realized by Fraunhofer FOKUS. Additionally a custom-built inertial 3D-body sensor system was developed within the context of the project, which was not part of our test environment.

The body movements for the therapy exercises are mapped by a combination of both sensor types, Kinect and body sensors. The sensor data are analyzed in real-time and mapped to a body model displayed on the display in front of the user. Green, yellow or red lines mark the body's contours and provide immediate feedback for correct or incorrect movements. Additional comments are provided written and acoustically through the user's home TV. Figure 1 shows the basic building blocks of the system.

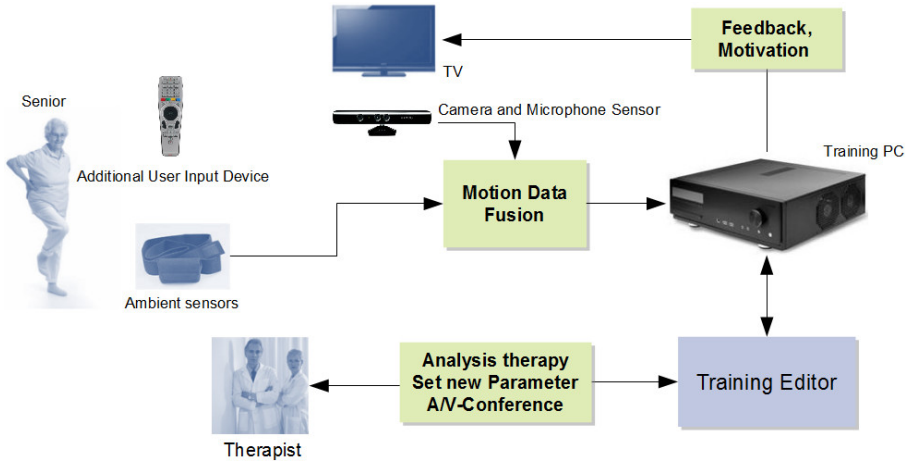


Fig. 1. Overview of the interactive training system

[3] http://www.attrakdiff.de/

Using a therapy editor, the therapist initially configures an individual training plan for the senior. Before starting a training session, the user gets her individual and actualized training plan from the online database, which is updated according to her personal training status. The database is located at Charité in Berlin, the largest geriatric clinic in Germany. After the training session, the training results are transmitted to the electronic health record in the safe and secure server back-end at the clinic. If needed, the system allows the patient also to get into contact with a therapist at Charité via A/V-communication as part of remote monitoring.

The design of the user interface including motivational elements is essential for user acceptance. To create familiarity with the training system in short time, we used an avatar-based approach, realized by Nuromedia. The therapist avatar talks to the user and visualizes reference movements. He or she – depending on the preferences of the user – provides personal interaction. The user avatar provides immediate feedback to her movements, functioning as a sort of mirror for the user. Immediate correctional feedback is provided through the color-coded body-parts (see above) and through comments from the therapist avatar.

The GUI is controlled by the "Interaction Manager" for user interaction and by the sensor engine for the animation of the user avatar (see [2,5,6] for some of the used technologies and approaches).

2.2 Related systems

For physical therapy, many projects exist to increase physical activity and to support motivational factors.

Within the project "GestureTek Health"[4] different gesture-control technologies exist for disability, hospital, mental health and educational sectors. For a virtual reality physical therapy, "GestureTek Health" developed a system called IREX™ (Interactive Rehabilitation and Exercise System). The system involves the user in a virtual game environment, where they are doing clinician prescribed therapeutic exercises. However it does not support a multimodal user interface.

The physical therapy system "Physiofun Balance Training" from Kaasa Health[5] is based on the Nintendo® Wii™ system. It uses the Wii console with a Wii Balance Board and a TV. A similar approach for a therapeutic balance test using comparable sensors is described by Dong et al. [1].

Ongoing projects for physical activities in rehabilitation are, e.g., PAMAP (Physical Activity Monitoring for Aging People)[6] and MyRehab[7], to name just a few. Our most

[4] http://www.gesturetekhealth.com
[5] http://www.kaasahealth.com
[6] http://www.pamap.org

recent search in German and EU project databases resulted in about 15 recently funded projects in that area. These systems usually analyze exercises and provide data for remote monitoring to be evaluated by a medical supervisor. They help patients to perform their rehabilitation and monitor their level of activity. Other projects[8] like Silvergame, age@home, KinectoTherapy, FoSIBLE, Eldergames, or Motivation also address the rehabilitation space. However all systems do not support a multimodal user interface, like ours does.

2.3 Scenario and Examples

To provide an insight in the interaction with the system, we will describe a short walk-through of the "One leg standing" training exercise for stroke patients.

The user starts the training system and is greeted by her virtual therapist. Then she is asked if she feels good or bad. The microphone is activated by the system, and she can reply, e.g., with "I'm fine"[9] or "I feel bad". As an alternative to speech, she can also use the remote control: Button 1 for "I'm fine" or button 2 for "I feel bad". The alternatives are presented on the screen clearly to address every available modality. In case the user feels bad, she is asked in the next step if she wants to be connected with her therapist. If the user wishes, a video call is initiated by the system. Otherwise, the system ends the training session. If the user feels OK, the exercise selection starts, which only shows the exercises that were previously selected by the therapist for the patient.

As an example, we describe briefly the therapeutic exercise "one leg standing" to improve the balance. All exercises, including the important posture parameters, were developed with physiotherapists. In that exercise the goal is to get the user to stand stable with a correct body posture on one leg. Here, the upper part of the body, the arms and the free leg should be kept stable. It starts in the upright standing. To stay in balance, the arms should be kept laterally with a small distance to the body. The next step is to pull up one knee, so that the angle between thigh and hip is 90 degrees. That position is to be held stable between 1 and 20 seconds, depending on the user's state of health. Right afterwards, she should repeat the procedure with the other leg. The evaluation during the exercise measures the upright posture without balance movements of arms, free leg or body, and the angle between thigh and hip. The described motion flow is used to specify the recognition, analysis and evaluation of therapeutic movements.

---

[7] http://www.first.fraunhofer.de/home/projekte/myrehab

[8] http://www.silvergame.eu/, http://www.joanneum.at/index.php?id=4243&L=0, http://www.kinectotherapy.in, http://fosible.eu/, http://www.eldergames.org/, http://www.motivotion.org/site/

[9] The German interactions are translated.

If the user has selected an exercise, it is explained, if desired. When an exercise is started, a start counter counts down, so that the user can prepare herself for the exercise. Then she follows the prescribed motion that is also visualized by the therapist's avatar (see fig.2, left). If the system detects a wrong move or a bad body posture, the user is immediately notified. We use different techniques simultaneously, voice announcement, acoustic signals and graphical feedback. When such an error occurs, the region with a bad posture is colored depending on the error level. The first error level is colored yellow.

For example, if the bearing of the upper part of the body is not correct, and the user leans back slightly, she immediately gets the friendly feedback not to lean back too far. If a critical error is detected, for example, if the user is almost falling down, the therapist gets a message to inform him about the critical event.

(a) One leg standing (b) Pass a river

Fig. 2. Example of a therapeutic exercise and the corresponding game

After an exercise, the user gets a break and then repeats the exercise. The therapist sets the break timing and repetitions in the therapeutic editor. In the end, the user receives an evaluation, which shows whether she has improved, or not.

Afterwards he has the opportunity to make another exercise or game (see fig.2, right). Motivation and retention to the training is of utmost importance. In addition to the training exercises we developed a game for each therapeutic exercise that takes up the theme of the therapeutic goal but has a more playful content. The following exercises including games are currently defined:

- Weight shift back and forth – Drive a motorboat
- Weight shift lateral standing – Slalom in standing
- One leg standing – Pass a river
- Weight shift lateral sitting – Slalom sitting

If no further exercises are scheduled, the therapist's avatar will initiate a dialog to terminate the session and the system shuts down.

3 Testing the system

3.1 Introduction and setup

As a result of discussions in the CASA goup, we decided to use the AttrakDiff™ questionnaire and website to measure the hedonic and pragmatic qualities of the system, using the results to infer the social attractiveness and agency of the system. To be perfectly clear about it: In this study we did not look into the effectiveness of the system wrt. stroke rehabilitation nor in the acceptance of the system in the initially targeted user group of the system! Our main goal was to measure the hedonic and pragmatic qualities of the system with persons from various backgrounds, thus gaining first insights in the overall user acceptance of our system. We used this opportunity also to get insights in the technical stability of the system, which worked without flaws during the tests.

AttrakDiff™ is an instrument for measuring the attractiveness of interactive products.[10] With the help of pairs of opposite adjectives, users (or potential users) can indicate their perception of the product. These adjective-pairs make a collation of the test dimensions possible. The following product dimensions are tested:

- Pragmatic Quality (**PQ**): Describes the usability of a product and indicates how successfully users are in achieving their goals using the product.
- Hedonic quality - Stimulation (**HQ-S**): People have an inherent need to develop and move forward. This dimension indicates to what extent the product can support those needs in terms of novel, interesting, and stimulating functions, contents, and interaction- and presentation-styles.
- Hedonic Quality - Identity (**HQ-I**): Indicates to what extent the product allows the user to identify with it.
- Attractiveness (**ATT**): Describes a global value of the product based on the quality perception.

Hedonic and pragmatic qualities are independent of one another, and contribute equally to the rating of attractiveness.

3.2 Participants and task

For our test we recruited 19 users from the Berlin region, either internally from the DFKI office in Berlin or externally. None of the subjects participated in the development of the system. The interaction sessions were either run at DFKI or at the homes of the users. The distribution wrt. age, gender and education is as follows:

[10] In this section we use and/or paraphrase graphical results and texts from the English report the website generates without special quoting. The evaluation was done in German.

Age	20 to 39	16
	41 to 60	2
	over 60	1
Gender	Male	13
	Female	6
Education	Lower Secondary Education	2
	Higher Secondary Education	2
	University	15

The tasks each participant had to fulfill was to perform one exercise and one game. All users were able to successfully perform their task.

3.3 Results and interpretation

In the portfolio-presentation, see fig. 3, the values of hedonic quality are represented on the vertical axis (bottom = low value). The horizontal axis represents the value of the pragmatic quality (i.e. left = a low value). The medium value of the dimensions are depicted with P and the confidence rectangle as . The confidence rectangle presents the users agreement in their evaluation of the product.

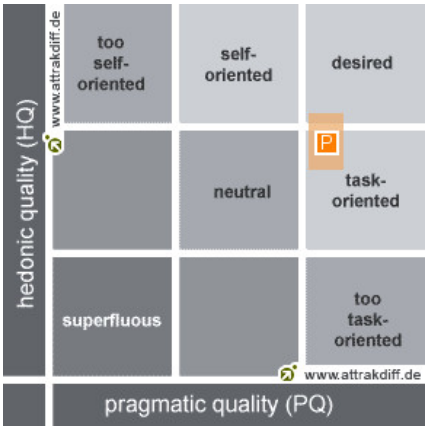


Fig. 3. Portfolio with average values of the dimensions PQ and HQ and the confidence rectangle of trainIT

Depending on the dimensions values the product will lie in one or more "character-regions". The bigger the confidence rectangle the less sure one can be to which region it belongs. A small confidence rectangle is an advantage because it means that the investigation results are more reliable and less coincidental. The bigger the confidence rectangle, the more variable the evaluation ratings are.

Overall, the trainIT system was rated as "*fairly practice-oriented*". The ***pragmatic quality*** is obviously high. The user is assisted by the system, and it is task oriented, but not too much. In terms of ***hedonic quality*** the character classification does clearly not apply because the confidence interval spills out over the character zone. The user is stimulated by the system, however the hedonic value is only slight above average. Since the confidence rectangle is small, the users agree in their evaluation of the system.

Detailed Analysis. The average values of the AttrakDiff™ dimensions for the system are plotted in fig. 4. In this presentation hedonic quality distinguishes between the aspects of stimulation (HQ-S) and identity (HQ-I). Furthermore the rating of product quality (PQ) and attractiveness (ATT) are presented.

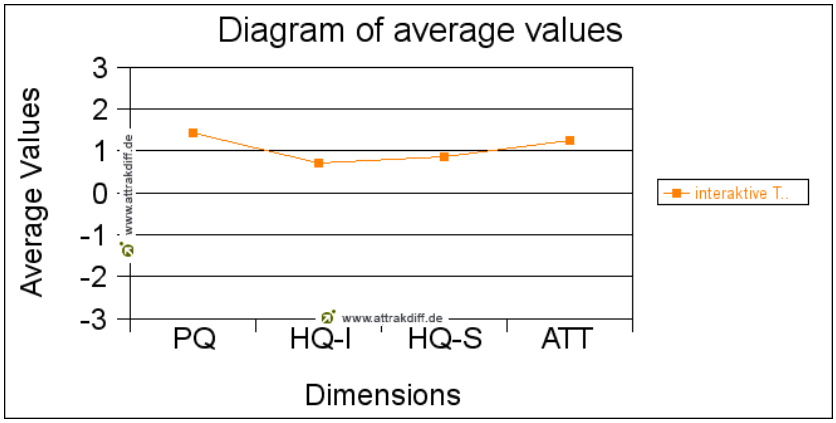


Fig. 4. Mean values of the four AttrakDiff™ dimensions for trainIT

With regards to hedonic quality – identity (HQ-I), the product is located in the average region. It provides the user with identification and thus meets well the standards. With regard to hedonic quality – stimulation (HQ-S), the product is also located in the slightly above average. The product's attractiveness value (ATT) is located in the above-average region, so the system is very attractive.

Description of Word-pairs. The mean values of the word pairs from the online questionnaire are presented in fig. 5. Of particular interest are the extreme values. These show which characteristics are particularly critical or particularly well resolved. Only on the world pairs "separates me – brings me closer" and "cautious – bold" are in the negative sector. The first value is obviously true: A training with a human person is more desirable than a training at home with remote interactions. The second word pair is actually good for this type of interactive system. Since the user should be cautious and should not overextend their training, this is a good indicator that we met one of the intended goals of the system.

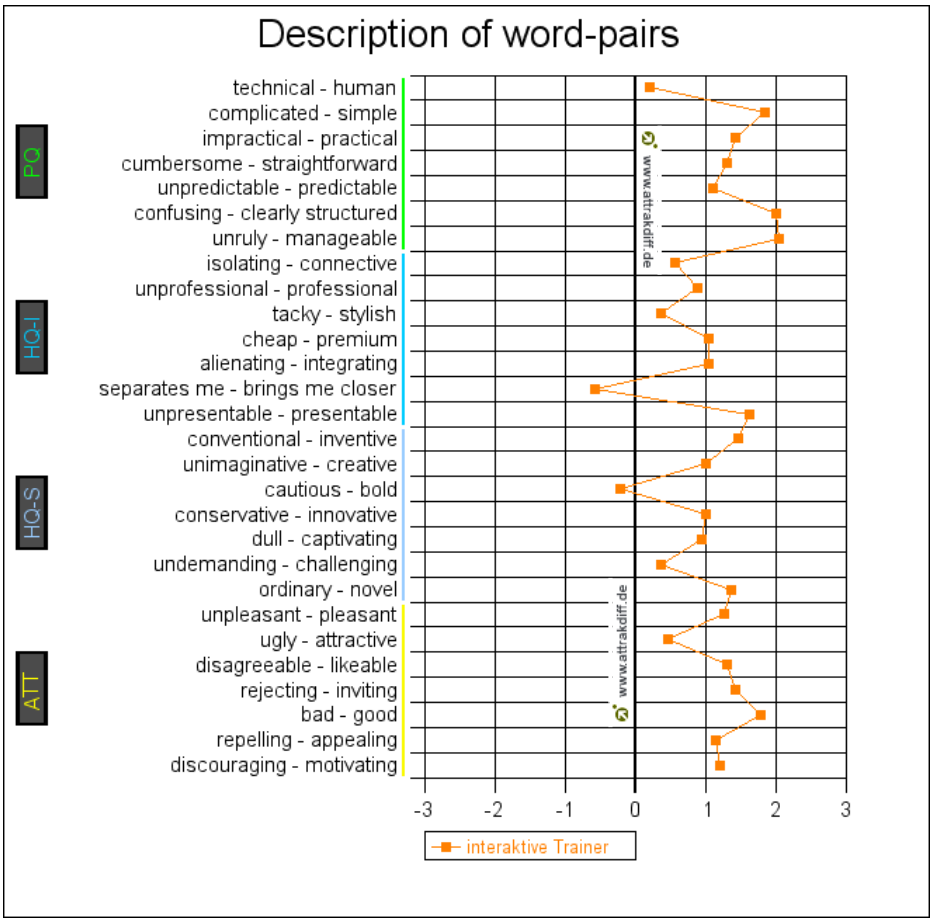


Fig. 5. Mean values of the AttrakDiff™ word pairs for trainIT

4 Social actorship in trainIT

Within the "Computers as a Social Actor" activity, we tried to come up with a common definition with some clear-cut criteria for social actorship. We developed the following definition of Social Actorship[11]:

> *Ability of the system to act in a **social context**, with **an implicit or explicit goal**. From the user perspective, Actorship is a characteristic of the system that makes the user **perceiving it as a human actor** to which s/he can direct*

[11] Source: Internal working document of the EIT ICTLabs Activity „Computers as a Social Actor", 2012.

*their **attention** and have **attention** in return (This can be explained by the Mirror Concept: the system that sense something and acts in response). Although, some systems could be seen as just a mediating actor, like mobile phone and ICT in general that fosters social interaction among people. In this case social actorship is seen as **the ability to influence and support the social life** of people.*

We agreed to focus on specific dimensions that define a system as a social actor. The dimensions are:

1. *Awareness*
2. *Intelligence=Intentionality*
3. *Embodiment: language, face, body*
4. *Social perception*
5. *Task/Goal of the system*
6. *Nature of the system: social tool-mediator-actor*

trainIT addresses most of the dimensions, defining actorship: it is aware of the user through the various sensors, it interacts intentionally, using a dialog strategy that reacts on the users' multimodal input, is embodied by an avatar, and has clear tasks and goals. In the above definition, we highlighted the main issues that are addressed by trainIT. The system was designed to act in the special context of rehabilitation, where people feel weak and sometimes out of touch with their usual social environment. The system's main goal is to engage the user in rehabilitation exercises. Through the use of a therapist avatar that is also talking to the user, the system creates the perception of a personal bond. The sensor feedback is also channeled through the avatar thus influencing the training exercise.

In detail we address:

- ***Social context**, with **an implicit or explicit goal**.*
  trainIT is tailored to help persons in a clear social context: being alone at home and getting back to be healthy again after a stroke. The explicit overarching goal is to go through a training plan set up by a physiotherapist, which broken down in various subgoals implicit in the various training sessions. This addresses especially dimension 5 (Task/Goal of the system). The test shows that the system has a pretty good pragmatic quality, i.e., that the users could interact with it successfully, even tough the test persons were
- *System that makes the user **perceiving it as a human actor***
  This goal is reached through, amongst others, a virtual actor that stands in as the person's physiotherapist. This therapist is able to carry out a spoken dialog about the training, supervises the exercises, as recorded by the sensors, motivates, and provides feedback. This addresses especially dimensions 6 (Nature of the system: social tool-mediator-actor), 2 (Intelligence=Intentionality) and 3 (Embodiment: language, face, body). The Hedonic Quality – Identity category in the tests relates to this dimension. As noted above the "*separates me – brings me*

closer" indicates that the non-human interaction is clearly noticed and valued negatively.

- *To which s/he can direct his/her **attention** and have **attention** in return*
  The person performing the exercise gets immediate attention to its performances, as measured by the sensors, both visually and through speech. E.g. corrections of the posture are signaled by sentences like "Please do not lean that far back". This addresses especially dimension 1 (Awareness). The word pairs from the Hedonic quality – Stimulation category in the tests show that the system stimulates mostly, even though the interaction is seen as *cautious*. Taking into account the target group of the system, namely mostly elderly people, this might actually be a good sign.
- ***The ability to influence and support the social life** of people.*
  Through the system the person gets immediate feedback to her performance and also can be sure that the performance results are channeled back to the telemedicine centre. Both the persons using trainIT and a supervisor at the centre can establish a direct interaction using a high definition videoconference solution built in the system. Thus the persons who have suffered a stroke and are not yet as mobile as before know there is a direct link that supports them, if necessary. As this group of persons often has problems taking up a normal life again, the system provides, besides the training, to support stability. This addresses partially dimension 4 (Social perception). The results in the Attractiveness category of the test mostly relate to this dimension: Only if the user considers the system positively in this category she or he might be willing to integrate the system in the daily life.

5 CONCLUSIONS

The study presented the trainIT interactive training system for stroke rehabilitation. We performed a reasonable sized usability test of the system using a standardized approach. We used it to categorize the social aspects of the system, which, on the one hand provides assurance in the work already done, and on the other hand, shows deficiencies that must be addressed in future versions of the system. As there is still no "standard" way to address social agency, usability tests like the one presented using more product oriented standard test tools are only a first step towards more elaborate testing schemes. Future activities continuing a CASA like theme could be very helpful in leveraging the results of this study and look deeper in the social aspects of personalized, interactive, agent based systems, which will become even more prominent in the immediate future.

The study shows that, even as a result of a research projects, trainIT is already desirable and attractive to a general audience, and thus, hopefully, also to potential customers. The system clearly was successfully tested on various categories that can be related to social actorship.

Acknowledgments

The work presented here was funded in parts by the German Federal Ministry of Education and Research under grant number 16KT0902 (Project SmartSenior) and in parts by EIT ICT Labs Activity 12124 (Computers as Social Actors). The responsibility for this publication lies with the authors. Thanks to Aaron Ruß for helpful comments.

References

1. Dong, L., Tan, M.S., Ang, W.T. and Ng C.K..: Interactive rehabilitation. i-CREATe'09 Proceedings of the 3rd International Convention on Rehabilitation Engineering & Assistive Technology (2009)
2. Gebhard, P., Kipp, M., Klesen, M. and Rist, T.: Authoring scenes for adaptive, interactive performance. Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems, Second International Joint Conference on Autonomous Agents and Multiagent Systems (2003)
3. Hennig B., N. Reithinger: Development of a Multimodal Interactive Training System in Therapeutic Calisthenics for Elderly People. Proc. of KI 2012, Springer Verlag (2012)
4. John, M., Klose, S., Kock, G., Jendreck, M., Feichtinger, R., Hennig, B., Reithinger, N., Kiselev, J., Gövercin, M., Kausch, S., Polak, M. and Irmscher, B.: Smartsenior's interactive trainer - development of an interactive system for a home-based fall prevention training for elderly people. Advanced Technologies and Societal Change, 7:305-316 (2012)
5. McTear, M.F.: Spoken dialogue technology: Enabling the conversational user interface. ACM Computing Surveys (CSUR), 34(1):90-169 (2002)
   Nass, C., Steuer, J., and Tauber, E.R.: Computers are social actors. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94). ACM, New York, NY, USA, 72-78 (1994)
6. Wahlster, W., editor. SmartKom: Foundations of Multimodal Dialogue Systems. Springer-Verlag, Berlin and Heidelberg (2006)

A crowdsourcing toolbox for a user-perception based design of social virtual actors

Magalie Ochs, Brian Ravenet, and Catherine Pelachaud

CNRS-LTCI, Télécom ParisTech
{ochs;ravenet;pelachaud}@telecom-paristech.fr

Abstract. One of the key challenges in the development of social virtual actors is to give them the capability to display socio-emotional states through their non-verbal behavior. Based on studies in human and social sciences or on annotated corpora of human expressions, different models to synthesize virtual agent's non-verbal behavior have been developed. One of the major issues in the synthesis of behavior using a corpus-based approach is collecting datasets, which can be difficult, time consuming and expensive to collect and annotate. A growing interest in using crowdsourcing to collect and annotate datasets has been observed in recent years. In this paper, we have implemented a toolbox to easily develop online crowdsourcing tools to build a corpus of virtual agent's non-verbal behaviors directly rated by users. We present two developed online crowdsourcing tools that have been used to construct a repertoire of virtual smiles and to define virtual agents' non-verbal behaviors associated to social attitudes.

1 Introduction

Virtual agents are increasingly used in roles that are typically fulfilled by humans, such as tutors in virtual learning class, assistants for virtual task realization, or play-mate in video game (e.g. [1, 2]). To embody successfully these social roles, virtual agents have to be able to express socio-emotional behavior during human-machine interaction. Indeed, several researches have shown that the expressions of emotion may enhance not only the believability of the virtual agent [3] but also the satisfaction of the user and his performance in task achievement [4, 5].

One of the key challenges in the development of embodied virtual agents is to give them the capability to display socio-emotional states *through their non-verbal behavior*. Several virtual agents are already able to express emotions or social attitudes through different modalities such as facial expressions, gestures or postures [6–9]. We can distinguish two main approaches to define the virtual agent's non-verbal behaviors associated to socio-emotional states: a *theoretical-based approach* and a *corpus-based approach*.

The theoretical-based approach consists in exploiting the studies in human and social sciences that have highlighted the characteristics of human's non-verbal behavior conveying socio-emotional state. For the expressions of emotion,

most of the computational models are based on the categorical approach proposed by Ekman [10] describing the human facial expressions of the "big six" basic emotions (joy, fear, anger, surprise, disgust, and sadness) [11]. Other psychological theories have been explored to define the emotional facial expressions of virtual agents, such as the dimensional theory [12] or the appraisal theory [13].

To gather more subtle and natural expressions, another approach is based on the analysis of annotated corpora of human expressing socio-emotional states. Based on an annotated corpus of humans expressions, different methods to synthesize virtual agent's non-verbal behavior have been explored. Using a motion capture system, the non-verbal behavior can be synthesized at a very low-level by re-targeting the points tracked on a human face and body to a virtual mesh (e.g. in [14]). Another method consists in applying machine learning technique on the collected data to automatically generate the non-verbal behavior associated to particular socio-emotional state (e.g. in [15]). Finally, the corpus may also be exploited by analyzing in detail the correspondences between the expressed socio-emotional states and the characteristics of the displayed non-verbal behaviors. Rules are then extracted and integrated in virtual agents (e.g. in [16]). Most of the corpus-based models of virtual agent's non-verbal behavior is based on corpus of real humans.

Some researchers have proposed to create corpus of virtual agent's non-verbal behaviors. For instance, in [17], a large amount of expressive virtual faces has randomly been generated. They have then been rated with emotional labels by numerous participants. This method has several advantages. First, it considers directly the user's perception of the virtual agent instead of replicating findings of human's non-verbal behavior on virtual agents. Moreover, a corpus of virtual agent's non-verbal behaviors avoids the problematic of acted human's expressions or the difficulty to collect spontaneous expressions. Secondly, this method may generate one-to-many correspondences between socio-emotional states and non-verbal behaviors. Thus, in [17], several facial expressions for each emotion type have been identified. The main problem with this method is the number of required participants (more than 400) for a repetitive and time-consuming task of rating each facial expression (2904 facial expressions). In this article, we propose an alternative methodology to identify the one-to-many correspondences between socio-emotional states and non-verbal behaviors by building and analyzing a corpus of virtual agent's non-verbal behavior directly rated by users.

One of the major issues in the synthesis of behavior using a corpus-based approach is collecting datasets, which can be difficult, time consuming and expensive to collect and annotate. A growing interest in using *crowdsourcing* to collect and annotate datasets has been observed in recent years [18]. *Crowdsourcing* consists of outsourcing tasks to an undefined distributed group of people, often using Internet to recruit participants informally or through formal paid mechanisms such as Amazon's Mechanical Turk [19]. Online tools for crowdsourcing have been developed to allow people to annotate human behaviors (e.g. in [20]). Moreover, an evaluation of the crowdsourcing workers' annotations showed that their qualities are comparable to expert annotators [20]. In order to build a rated

corpus of virtual agents' non-verbal behaviors, we have implemented a toolbox to easily develop online crowdsourcing tools. The objective of such a crowdsourcing tool is to offer the possibility to users to directly configure the virtual agent's non-verbal behaviors conveying particular socio-emotional states. For instance, the users may have the task to define the virtual agent's gestures and facial expressions corresponding to the expression of certain attitudes such as friendliness or dominance. This method avoids the traditional approach of creating a repertoire of socio-emotional states by asking users to label a set of predefined non-verbal behaviors. Instead, users are placed at the heart of the non-verbal behavior creation process. Even if a finite set of animations is pre-defined, the tool gives the users the impression to create the non-verbal behavior they believe corresponds to a given socio-emotional state.

To create such a crowdsourcing tool, the toolbox covers different functionalities:

- the construction of an audiovisual corpus of virtual agent's non-verbal behaviors containing all the possible combinations of modalities (Section 2);
- the framework to develop and distribute the tool online (Section 3).

As use cases, we present two developed online crowdsourcing tools that have been used to construct a repertoire of virtual smiles and to define virtual agent's non verbal behaviors associated to social attitudes (Section 4). We conclude and discuss the limits of this method in Section 5.

2 The GretaModular Platform to create corpora of virtual agents' non-verbal behaviors

The first step to develop the online crowdsourcing tool is to generate the videos of virtual agents displaying different combinations of facial expressions, gestures, and postures. For this purpose, the platform we are using to animate a virtual agent is *GretaModular*, a significantly improved version of the Greta system [21].

GretaModular offers several modules, each dedicated to particular functionality. The core modules, based on the SAIBA framework [22], include an *Intent Planner*, a *Behavior Planner* and a *Behavior Realizer* to compute multimodal expressions of communicative intentions. Additional peripheral modules endow the system with several useful functionalities. For instance, the *Gesture Editor* and the *Facelibrary Viewer* enable one to easily define new gestures and facial expressions of virtual agents. Others modules, such the BML and FML file readers, the Character Manager, the Ogre3D Player and the Video Capture module, facilitate the creation of a corpus of virtual agent animations. Moreover, the flexible architecture of the platform has been implemented with a graphical user interface that allows a simple manipulation of the modules by drag and drop. New modules may be easily developed and plug in to add new functionalities to the system.

In *GretaModular*, a repertoire of signals contains the description of non-verbal behavior (facial expressions, gestures, postures) in BML (Behavior Markup Lan-

guage) [23]. The links between non-verbal behaviors and communicative intentions are specified in a lexicon. Communicative intentions are encoded with FML (Function Markup Language) [24]. Furthermore, expressivity parameters can be used to modulate the qualitative execution of non-verbal behaviors (e.g. fluidity of gestures) [25].

Using *GretaModular*, one may create a corpus in 5 steps (Figure 1). The ap-

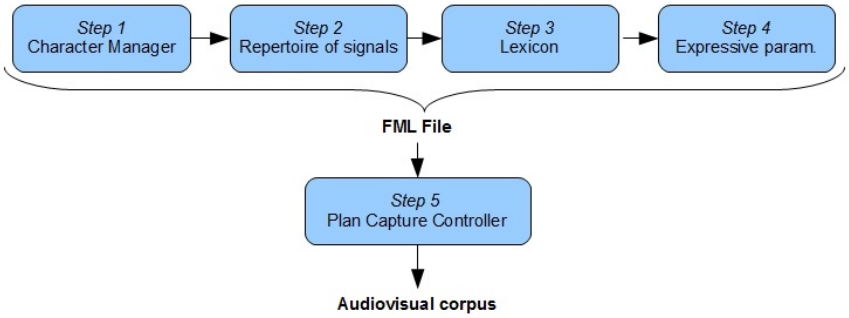


Fig. 1. Steps to create a corpus of virtual agent's non-verbal behaviors using *GretaModular*

pearance of the virtual agent has to be first chosen. The physical appearance of the virtual agent may be selected in the *Character Manager* module of *GretaModular*. The available virtual agents are illustrated Figure 2. The second step

Fig. 2. Virtual agents in *GretaModular*

consists in creating or completing the repertoire of signals in BML. In *GretaModular*, a large set of signals has already been defined for the virtual agents of the platform (Figure 2). One may easily design new signals using the *Gesture Editor* and the *Facelibrary Viewer* of the *GretaModular* platform. Thirdly, the different links between non-verbal behaviors and communicative intentions are defined in the *lexicon*. Moreover, different expressive parameters of the virtual agent's non-verbal behavior may be configured. To generate the videos of the corpus, we

have developed a specific module, named the *Plan Capture Controller Module*, built on the top of the video capture module. This module takes as input an FML file describing the communicative intentions (e.g. emotions, beliefs) that the virtual agent has to express through its verbal and non-verbal behavior [21]. Given that communicative intention may be expressed through different signals and modalities, the module computes and loads all the possible animations corresponding to the FML file (based one the *lexicon* in which the correspondences between communicative intentions and behaviors are described), plays them and captures each one of them in separate video files. For instance, the intention to greet may be expressed by a head nod or a hand shake, with or without a smile. For this communicative intention, the *Plan Capture Controller* module creates 4 different video files of a virtual agent that greets in 4 different manners. The *Plan Capture Controller* module may generate the animations with different values of the expressive parameters. Finally, with *GretaModular*, one may rapidly create a corpus of videos of virtual agents with different appearances displaying multimodal non-verbal behaviors with different expressive parameters.

So far, we have created two corpora of virtual agent's non-verbal behaviors. A first corpus has been dedicated to the virtual agent's *smiles*. One hundred and ninety two different animations of a smiling virtual agent face have been generated. The smiles varied on several morphological and dynamic parameters defined from the theoretical and empirical research on human smiles [26–28]: the cheek raising (Action Unit 6 - AU6), the lip press (Action Unit 24 -AU24), the amplitude of the smile (Action Unit 12 - AU12), the symmetry of the lip corners, the mouth opening (Action Unit 25 - AU25), the duration of the smile and the velocity of the rise and of the decay of the smile. We have considered two or three discrete values for each of these parameters: small or large smile (for the amplitude); open or close mouth; symmetric or asymmetric smile; tensed or relaxed lips (for the AU24); cheekbone raised or not raised (for the AU6); short (1.6 seconds) or long (3 seconds) total duration of the smile, and short (0.1 seconds), average (0.4 seconds) or long (0.8 seconds) beginning and ending of the smile (for the rise and decay). All the possible combinations of these discrete values have been generated to create the corpus of the virtual agent's smiles.

A second corpus has been created to study the non-verbal behaviors conveying the social attitudes of dominance, submissiveness, friendliness and unfriendliness. For this purpose, we have generated 1440 videos corresponding to all the possible combinations of the following parameters identified as cues of social attitudes [29–33]: type of facial expressions (positive: smile, negative: frown or neutral), the activated modalities (arm gestures, head gestures, both or none), the amplitude of arm gestures (small, medium or wide), the power of arm gestures (weak, normal or strong), the head position (upward, downward, tilted aside or straight) and the presence of gaze avoidance (yes or no). The corpus contains animations of two different virtual agents, one with a female appearance and one with a male appearance.

3 Online Crowdsourcing tools for the user design of virtual agent's non-verbal behaviors

In order to create crowdsourcing tools based on virtual agent's non-verbal behaviors corpora, we have created a framework using Flash technology to enable broad distribution on the web. The framework allows one to develop a web application in which users have the task to define the non-verbal behaviors of a virtual agent associated to particular socio-emotional states. The interface of the application is composed of 4 parts (Figures 3 and 4):

1. the upper part contains a description of the task;
2. the left part contains a video showing the virtual agent animation, in a loop;
3. the right part contains a panel with the different non-verbal parameters that the user can change to define the virtual agent's non-verbal behavior. Any time the user changes the value of one of the parameters, a corresponding video is automatically played;
4. the bottom part contains a Likert scale that allows users to indicate their satisfaction with the created animation.

To develop the crowdsourcing tool, one has to define the tasks of the users. The panel of the non-verbal parameters has to correspond to the parameters considered in the creation of the corpus of virtual agents' non-verbal behaviors (Section 2). The videos displayed according to the selected parameters are directly extracted from the corpus. The Flash framework includes a connection with a database to record the responses of the users.

Using this framework, two crowdsourcing tools have been developed: *E-Smiles-creator* and *GenAttitude*. The interfaces of these tools are illustrated in Figures 3 and 4. The objective with the *E-smiles-creator* tool is to study the morphological and dynamic characteristics of different smile types. Through the interface of the *E-smiles-creator* (Figure 3), the users have the tasks to create different types of smile (amused, polite, and embarrassed). To create each of these smiles, the users select the parameters of the smile with the radio buttons (Panel 3, Figure 3). These parameters correspond to those used to create the corpus of virtual smiles. The corresponding video contained in the corpus is automatically loaded and played (Panel 2, Figure 3). With the *GenAttitude* tool (Figure 4), the objective is to identify virtual agents' non-verbal behaviors corresponding to the expression of different attitudes. The users have the tasks to configure the non-verbal behavior of the virtual agent corresponding to the expression of a particular social attitude (dominant, submissive, hostile, or friendly) for a given communicative intention. For instance, the users have the tasks to configure the non-verbal behavior of the agent when it is asking something with a dominant attitude. The parameters of the non-verbal behavior (Panel 3, Figure 4) correspond to those used to create the corpus of videos (Section 2).

Finally, with the crowdsourcing tools, users unconsciously rated videos of the corpus (of the virtual agents' non-verbal behaviors) with pre-defined labels

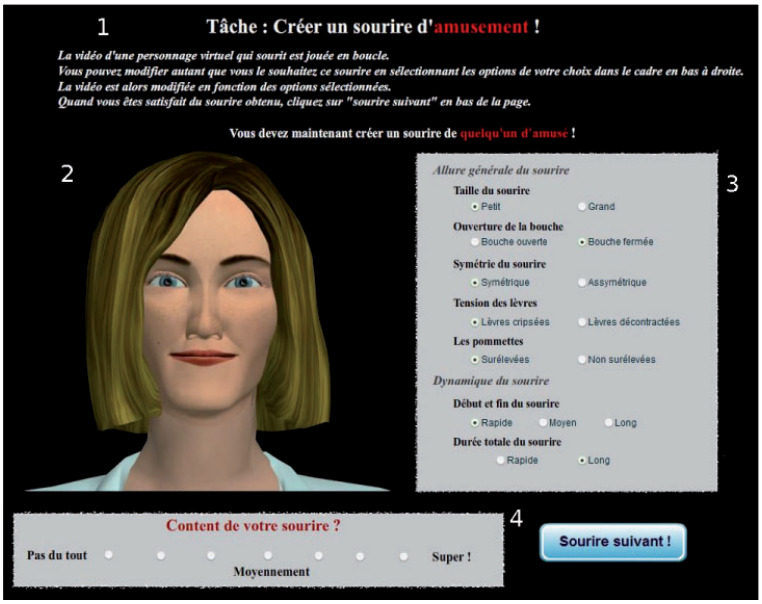


Fig. 3. Screenshot of *E-Smiles-Creator*

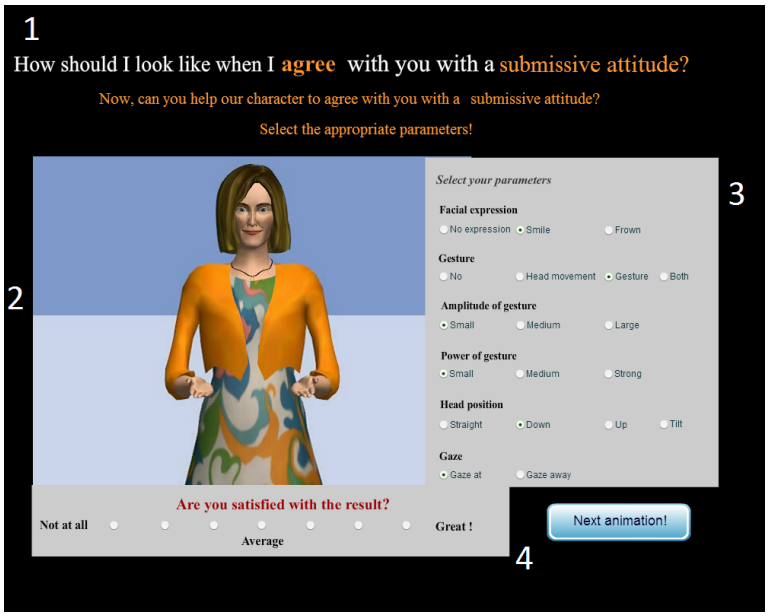


Fig. 4. Screenshot of *GenAttitude*

(emotions or social attitudes). However, not all the videos of the corpus are rated. Only those that appear as relevant for the pre-defined labels are rated.

4 Analysis of the collected data on virtual agents' non-verbal behaviors

Through the developed crowdsourcing tools presented above, we have collected 1044 smile descriptions (from 348 participants among which 195 females; mainly French, with a mean age of 30 years) and 925 non-verbal behavior descriptions corresponding to social attitudes (from 170 participants among which 50 females, mainly French, with a mean age of 29 years), in one week. The participants were recruited via online mailing lists (they were not payed). The average level of satisfaction of the participants (5,3 on a Likert Scale of 7 points, Panel 4) shows that the participants were globally satisfied by the interface to create the animations. Moreover, the positive comments posted by the participants show that their user experience was funny and enjoyable. We have analyzed the collected data to create a repertoire of non-verbal behaviors conveying different emotions and social attitudes.

The levels of satisfaction indicated by the participants (Panel 4, Figures 3 and 4) was used to give higher weight to the non-verbal behaviors with a high level of satisfaction[1]. We made the assumption that the non-verbal behaviors with a high level of satisfaction were more reliable that those with low level. In fact, we *oversampled* the corpora such as each created non-verbal behavior was duplicated n times, where n is the level of satisfaction associated with this non-verbal behavior. For instance, a smile with a level of satisfaction of 7 was duplicated 7 times whereas a smile with a level of satisfaction of 1 was not duplicated. The resulting data sets were composed of 5517 descriptions of smiles and 4947 non-verbal behavior descriptions conveying social attitudes.

To analyze the collected data and construct computational models, different methods have been explored. Concerning smiles, we used a decision tree learning algorithm to identify the different characteristics of the amused, polite, and embarrassed smiles in the corpus. The decision tree has the advantage to be well-adapted to qualitative data and to produce results that are interpretable and that can be easily implemented in a virtual character. The nodes of the decision tree correspond to the smile characteristics and the leaves are the smile types. Different leaves correspond to the same smile type enabling one to identify one-to-many correspondences between smile types and facial expressions. The virtual agent may then express the same type of smile in different manners during an interaction to avoid repetition of the exact smile pattern. Previous research has shown that the non-repetitive behaviors of a virtual agent improves its perceived believability [6]. A perceptive study has validated most of the smiles as appropriate in amused, polite or embarrassed situations. The smiles decision tree and the validation study are described in more details in [34].

[1] Note that the data could be analyzed without oversampling.

For the second corpus, we have explored another method of analysis: the corpus of the virtual agents' non-verbal behaviors associated to social attitudes has been used to create a Bayesian network. A Bayesian network is a directed acyclic graph that represents causal relations between variables, the strength of these relations being represented by conditional probabilities. The structure of the network has been defined based on a statistical analysis of the corpus. The input nodes of the model are the social attitudes (dominant, submissive, friendly, or hostile) and the communicative intentions. The outputs are the characteristics of the non-verbal behavior that should convey a given communicative intention with a particular attitude. The Bayesian network directly represents the cause-effect relations between our input variables (communicative intentions and social attitudes) and output variables (the non-verbal behavior parameters). The parameters of the model (*i.e.* the probability of the edges) are directly extracted from the built oversampled corpus. The probabilistic nature of such a model enables us to introduce variabilities in the outputs, particularly relevant for modeling human-like uncertain behavior. Once again, the model may be used to determine one-to-many correspondences between attitudes and non-verbal behaviors. Moreover, the model provides a probability that the virtual agent's non-verbal behavior is perceived with the expected attitude. Also, the same Bayesian network can be used to infer the probabilities for the input variables given the output values. This could be use to retrieve the most likely attitude and intention given the nonverbal behavior parameter values. The Bayesian model is detailed in [35].

5 Conclusion

In this article, we have presented a toolbox that enables one to create a crowdsourcing tool to build corpus of virtual agents' non-verbal behaviors. We have presented two use cases that aimed at analyzing the links between the characteristics of a signal (the smile) or of a multimodal behavior to the expression of emotions or social attitudes. Instead of asking users to rate a set of virtual agent animations, we have proposed an approach in which the user has the impressions to directly design the virtual agent's non-verbal behavior. The large number of participants, their levels of satisfaction and their positive posted comments indicate that the proposed tasks and interface are satisfying and attractive. The size of the obtained corpus enables one to apply different methods from a statistical analysis to machine learning techniques.

In future works, we aim at improving the interfaces of the crowdsourcing tools. In the current version, the values of the non-verbal behavior's parameters are selected through radio buttons. Continuous values indicated with sliders could enable us to obtain a more fine-grained description of the virtual agent's non-verbal behaviors. Moreover, the Bayesian network model resulting from the *GenAttitude* tool has to be evaluated during interaction with users to ensure that the non-verbal behaviors convey the expected attitudes.

Acknowledgment

This research has been supported by the European projects NoE SSPNet and IP REVERIE. The authors would like to thank *André-Marie Pez* and *Pierre Philippe* for the implementation of the *GretaModular* platform.

References

1. Johnson, W.L., Rickel, J.W., Lester, J.C.: Animated pedagogical agents: Face-to-face interaction in interactive learning environments. International Journal of Artificial Intelligence in Education **11** (2000) 47–78
2. André, E., Klesen, M., Gebhard, P., Allen, S., Rist, T.: Integrating models of personality and emotions into lifelike characters. In: Affective interactions, Springer (2001) 150–165
3. Bates, J.: The role of emotion in believable agents. Communication of ACM **37**(7) (July 1994) 122–125
4. Beale, R., Creed, C.: Affective interaction: How emotional agents affect users. International Journal of Human-Computer Studies **67**(9) (2009) 755–776
5. Partala, T., Surakka, V.: The effects of affective interventions in human-computer interaction. Interacting with Computers **16**(2) (2004) 295309
6. Niewiadomski, R., Hyniewska, S., Pelachaud, C.: Constraint-based model for synthesis of multimodal sequential expressions of emotions. IEEE Transactions on Affective Computing **2**(3) (2011) 134–146
7. Bickmore, T.W., Picard, R.W.: Establishing and maintaining long-term human-computer relationships. ACM Trans. Comput.-Hum. Interact. **12**(2) (June 2005) 293–327
8. Kasap, Z., Moussa, M.B., Chaudhuri, P., Magnenat-Thalmann, N.: Making them remember emotional virtual characters with memory. IEEE Computer Graphics and Applications (March 2009) 20–29
9. Albrecht, I., Schroder, M., Haber, J., Seidel, H.: Mixed feelings : expression of non-basic emotions in a muscle-based talking head. Special issue of Journal of Virtual Reality on Language, Speech and Gesture **8(4)** (2005) 201–212
10. Ekman, P., Friesen, W.V.: Unmasking the Face. A guide to recognizing emotions from facial clues. Prentice-Hall, Inc., Englewood Cliffs, New Jersey (1975)
11. Ostermann, J.: Face animation in mpeg-4. In Pandzic, I., Forchheimer, R., eds.: MPEG-4 Facial Animation - The Standard Implementation and Applications. Wiley, England (2002) 17–55
12. Mehrabian, A.: Basic Dimensions for a General Psychological Theory: Implications for Personality, Social, Environmental, and Developmental Studies. Oelgeschlager, Gunn & Hain, Cambridge, Mass (1980)
13. Scherer, K.: Appraisal processes in emotion: Theory, methods, research. (2001) 92–119
14. Niewiadomski, R., Pelachaud, C.: Towards multimodal expression of laughter. In: The 12th International Conference on Intelligent Virtual Agents, Santa Cruz, USA (2012) 231—244
15. Ding, Y., Radenen, M., Artiere, T., Pelachaud, C.: Speech-driven eyebrow motion synthesis with contectual markovian models. In: To appear in Proc. of ICASSP. (2013)

16. Castellano, G., Mancini, M., Peters, C., McOwan, P.: Expressive copying behavior for social agents: A perceptual analysis. IEEE Transactions on Systems, Man and Cybernetics **42**(3) (2012) 776–783
17. Grammer, K., Oberzaucher, E.: The reconstruction of facial expressions in embodied systems. ZiF : Mitteilungen (2006)
18. Yuen, M.C., King, I., Leung, K.S.: A survey of crowdsourcing systems. In: Proceedings of the IEEE International conference on Social Computing (SocialCom). (oct. 2011) 766 –773
19. Mason, W., Suri, S.: Conducting behavioral research on Amazon's Mechanical Turk. Behavior Research Methods **44**(1) (June 2011) 1–21
20. Park, S., Mohammadi, G., Artstein, R., Morency, L.P.: Crowdsourcing micro-level multimedia annotations: the challenges of evaluation and interface. In: Proceedings of the ACM multimedia 2012 Workshop on Crowdsourcing for multimedia. CrowdMM '12, New York, NY, USA, ACM (2012) 29–34
21. Niewiadomski, R., Bevacqua, E., Mancini, M., Pelachaud, C.: Greta: an interactive expressive ECA system. In: Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 2. AAMAS '09, Richland, SC, International Foundation for Autonomous Agents and Multiagent Systems (2009) 1399–1400
22. Kopp, S., Krenn, B., Marsella, S., Marshall, A.N., Pelachaud, C., Pirker, H., Thrisson, K.R., Vilhjlmsson, H.: Towards a common framework for multimodal generation: The behavior markup language. In: Proceedings of the international conference on Intelligent Virtual Agents (IVA), Springer-Verlag Berlin, Heidelberg (2006) 21–23
23. Kopp, S., Krenn, B., Marsella, S., Marshall, A., Pelachaud, C., Pirker, H., Tharisson, K., Vilhjalmsson, H.: Towards a common framework for multimodal generation: The behavior markup language. In Gratch, J., Young, M., Aylett, R., Ballin, D., Olivier, P., eds.: Intelligent Virtual Agents. Volume 4133 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2006) 205–217
24. Heylen, D., Kopp, S., Marsella, S., Pelachaud, C., Vilhjlmsson, H.: The next step towards a function markup language. In: Proceedings of the international conference on Intelligent Virtual Agents (IVA). (2008)
25. Mancini, M., Pelachaud, C.: Dynamic behavior qualifiers for conversational agents. In Pelachaud, C., Martin, J.C., Andre, E., Chollet, G., Karpouzis, K., Pele, D., eds.: Intelligent Virtual Agents. Volume 4722 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2007) 112–124
26. Ambadar, Z., Cohn, J.F., Reed, L.I.: All Smiles are Not Created Equal: Morphology and Timing of Smiles Perceived as Amused, Polite, and Embarrassed/Nervous. Journal of Nonverbal Behavior **17-34** (2009) 238–252
27. Ekman, P.: Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage. W.W.Norton & Company edn. (2009)
28. Hoque, M., Morency, L.P., Picard, R.W.: Are you friendly or just polite? - analysis of smiles in spontaneous face-to-face interactions. In: Proceedings of the international conference on Affective Computing and Intelligent Interaction (ACII), Berlin, Heidelberg, Springer-Verlag (2011) 135–144
29. Briton, N.J., Hall, J.A.: Beliefs about female and male nonverbal communication. Sex Roles **32** (1995) 79–90
30. Burgoon, J.K., Buller, D.B., Hale, J.L., de Turck, M.A.: Relational Messages Associated with Nonverbal Behaviors. Human Communication Research **10**(3) (1984) 351–378

31. Burgoon, J.K., Le Poire, B.A.: Nonverbal cues and interpersonal judgments: Participant and observer perceptions of intimacy, dominance, composure, and formality. Communication Monographs **66**(2) (1999) 105–124
32. Carney, D., Hall, J., LeBeau, L.: Beliefs about the nonverbal expression of social power. Journal of Nonverbal Behavior **29** (2005) 105–123
33. Hess, U., Thibault, P.: Why the same expression may not mean the same when shown on different faces or seen by different people. In Tao, J., Tan, T., eds.: Affective Information Processing. Springer London (2009) 145–158
34. Ochs, M., Niewiadomski, R., Brunet, P., Pelachaud, C.: Smiling virtual agent in social context. Cognitive Processing, Special Issue on"Social Agents" **13**(22) (2012) 519–532
35. Ravenet, B., Ochs, M., Pelachaud, C.: From a user-created corpus of virtual agent's non-verbal behavior to a computational model of interpersonal attitudes. In: To appear in the proceedings of the Intelligent Virtual Agents (IVA) conference. (2013)

The Intentional Interface

Peter Wallis

Centre for Policy Modelling
Business School
Manchester Metropolitan University
Manchester, United Kingdom
pwallis@acm.org

Abstract. The SERA project put "robot rabbits" in older peoples homes and recorded what happened. The challenge is now to use that data to develop better rabbits, but how? We are currently working on a methodology for distilling this data down into explanatory narratives, but in the mean time we are working on the idea that the essential nature of the SERA interface (and other conversational agents) is that it is *intentional* - it is an interface that sets out to have people ascribe beliefs and desires to it. According to Tomasello, this is not enough however. An intentional interface also needs to intend to help - it needs to be *cooperative*. What this means in detail is fleshed out in the context of an IVR system - a computer that answers the telephone.

1 Introduction

A year on from the SERA project - Social Engagement with Robots and Agents - this paper looks back on what we did, and attempts to put the lessons learned in a historical context. Our vision was to use a talking robot rabbit (an augmented Nabaztag) as long term "companion". Obviously it is beyond us to create a perfect simulation of a human conversational partner, but was current technology able to capture the *essence* of what is needed? The answer was no, but the experience certainly prompted some thinking about that essence. This paper develops that thinking and in doing so, offers a "grand unified theory" of HCI based on Dennett's *Intentional Stance* [1]. The theory is then used to develop an IVR system (Interactive Voice Response) that answers the telephone and, as such, is inevitably treated as a social actor.

2 SERA

The SERA project was funded under the FP7 Theme 2.2: Cognitive Systems, Interaction and Robots, and the aim was to put real robots in real people's hallways and kitchens and record what happens. The work was done with the School of Health and Related Research at Sheffield (ScHARR) which had extensive experience recruiting subjects from the broader community, and which was working with "smart homes" with a view to "life-style reassurance" in which

people living alone could be assured that, should something happen to them, help would be at hand. Building a state-of-the-art companionable robot is a project in itself [2] and so, rather than building a mobile autonomous robot, we decided to use a commercial off the shelf Nabaztag [3] that behaved as if it could sense it's environment, but which actually used the smart home sensors. Although the "robot" was not be mobile, it was able to sense its environment and was thus able to initiate action in a way that is expected of robots. It is in this sense of "robot" - autonomous action based on sensing the environment - that our nominally simple interface addressed the call. Figure 1 shows the setup in use.

Fig. 1. Sarah and Harvey.

That was the set-up, but another challenge was deciding what the rabbit should say. We settled on the popular scenario of an "exercise companion". Using the Trans Theoretical Model of behaviour change (TTM) [4] which places people in one of 5 stages, the system could introduce the advantages of being fit at the appropriate moment if the user was in the pre-contemplation stage, or identify progress if the user was in the maintenance stage and so on. The idea was to use "key-word spotting" speech technology and develop a system that was primarily "system initiative" with conversation being initiated by the following events:

- Keys off (participant going out)
- Keys on (participant returning home)
- PIR & first appearance in the morning
- PIR & last activity of the day has been done
- PIR & a new message/recommendation
- participant initiation - "Hey rabbit!"

If the keys came off when the subject had an entry in the diary for some exercise, then the rabbit could say things like "Going swimming? Have a good time" - which at least one subject found quite impressive even if she knew how it worked.

Before introducing the theory, the shared conclusions from the SERA project were, first, don't try to use ASR in the wild - the Siri publicity (formal and viral) is a dream. That kitchen is not my kitchen: there are no kids practicing the tin whistle, no oil sizzling on the hob, no radio playing, no extractor fan, no traffic and no refrigerator humming away. Indeed she is perfect as well with a nice East Coast accent with no Yorkshire clipping or Australian vowels. And the recipe - no star annis, or dried paw paw; nothing out of the ordinary. Put a speech recognition system in a kitchen and, we discovered, word error rates are too low for even a handful of key phrases.

Second, managing attention is a big issue when the interface is sensing the environment, is always on, and can be proactive. The classic HCI interface is passive (see below) and needs to be "poked" before it behaves. Our existing model for a proactive interface is the alarm that *demands* attention. In between is the telephone that could be demanding when its location was fixed but mobile phones would, ideally, be more "socially aware" of the context. The SERA rabbits used a PIR security sensor to detect with people were near but is she in a hurry? Is she making an omlette? [5] or is she just after a glass of water in the middle of the night?

Third, people have "idiosyncratic" behaviour. Where one person gets cross and yells, another laughs while another roles her eyes. Another subject may frown or not respond at all. Naturally the notion of a response to "the same" event is also problematic without a framing theory but this issue is well known; what stood out for us all was the huge range of responses across socio-economic backgrounds.

Finally, we can say that there is no consensus on what to do with the data we collected. We could all publish papers, but how can the data be used to advance the state-of-the-art? There has been some work on a better methodology for looking at the data [6], but in this paper a theory of HCI is introduced which may help "frame" the questions one might ask of the data and so help identify the issues and suggest improvements. The proposed theory is based on how people view artifacts around them.

3 How the mind works[1]

The proposal is that the SERA rabbits were not simply a conventional human-computer interface with a speech recognition front end, but were instead an attempt at an *intentional interface.* This is not to say that the SERA interface was unique – many have attempted similar things – the point is to introduce a class for interface for which SERA is an example. In order to compare and contrast, the observation is that we can classify human-computer interfaces based

[1] Thanks to Steven Pinker for the title of this section and the next.

on how the user goes about understanding the computer, and that interesting distinctions can be drawn by looking at Dennett's position on intentional systems.

Dennett, argues that the study of minds is different to the study of brains, and that the wide spread use of "folk psychology" in the Social Sciences is perfectly valid as science. For realists there is little doubt that minds reside in the hardware of brains, but studying brains is not necessarily going to provide explanations for why things are the way they are. As a scientist one might have a theory of id, ego and super-ego, or as a mathematician one might have an elegant Bayesian model of how brains work that is meant to explain things, but Dennett's line is that the psychology we use in our everyday lives is equally valid as a scientific theory *and more efficient* .

Dennett argues that humans use three different approaches, or stances, when trying to predict the behaviour of something. When a system is fairly simple - balls on a level table perhaps - then we can use a causal model to predict future events. Tapping the white ball in a particular way will cause it to role over to the red ball and knock it into the centre pocket. Taking this **physical stance**, people use their knowledge of hundreds (if not thousands) of highly reliable "facts" about the way things behave to assemble chains of causal events to predict the future. Dennett was writing at the time of good old fashioned AI and so the nature of these facts, as we now know from the work in computer science, is problematic and (apparently) based on situated action. But possible enumerations and classification of the base facts is not the point; the point is we can and do reason causally. Taken to its extreme, this is the idea of a clockwork, deterministic universe and that ultimately "there is only physics".

Another way we humans predict the future is by knowing what something is designed to do. Pressing the brake pedal when driving, one does not reason about hydraulic fluid, but simply knows what that pedal is *meant* to do. An alarm clock is too complex to follow the internal workings in a causal sense but, knowing what it is designed to do, one can set it in the evening and predict that it will wake you in the morning. This is of course where classic HCI is based with advice on how to create good interfaces being things like making sure that the system works as designed, and that the user has a clear idea of the function of the design (e.g. Interaction Design: beyond human-computer interaction (2ed) [7]). When we use this **design stance** , note how it licences the notion of something being "broken".

The **intentional stance** is what we use when a system is too complex to predict with the physical stance, and the purpose of system - what it is designed for - is inaccessible to us. We humans have a strong tendency to assume that something capable of autonomous action will do what it believes is in its interests. That is, that the system will have desires, and that it can plan its actions to achieve (some of) those desires given its beliefs about the current state of the world. This tendency is very strong in us. Seeing two children tugging at a teddy bear, the casual observer will assume they both *want* it. When playing chess against a computer, I do not reason about the causal behaviour of registers and

electricity, but rather predict the future by reasoning along the lines of it *wanting* to take my bishop. The consequences of a rational agent wanting something do not need to be spelt out for us; we just know. We are also likely to explain things that are not rational action with this model and Dennett gives a lovely example of someone explaining that electricity normally *wants* to take the shortest path but sometimes it "gets confused".

3.1 The Human-Computer Interface from the stances

Current HCI best practice can be critiqued as using a "tool" metaphor in which the computer is wielded by the user to achieve his or her goals. This is fine as far as it goes and has the advantage that as long as the tool does what it is *designed* to do, the user is responsible for outcomes. Hit your thumb with a hammer and there is only yourself to blame. Using such a metaphor, the guidance on HCI design is about making the design clear, and the consequences of an action explicit and immediate [7]. In retrospect this is Dennett's design stance. The human is expected to understand what the interface is designed to do, and then wield it appropriately.

Extending the metaphor, the sexy human-computer interfaces are those based on the physical stance. On the surface there is a class of interface that exploits the "facts" we have about the physical world with desk tops as a place to "put" things temporarily, folders that "contain" stuff, and recycle bins for the things we don't want any more. Today's touch screens allow things to be "flicked" and multi touch screens allow things to be "stretched" in a way that tend to obey our facts about the (physical) world. At a deeper level, many modern interfaces - especially those designed for new markets such as children - not only allow, but actively encourage exploration. In effect they encourage the user to discover things about causality in the virtual world that mirror the "hundreds or thousands of facts" we know that support the physical stance in the physical world.

This exploration process is clearly what Suchman points to in her classic work on situated action and the photocopier [8].

The proposal is that the *essence* of our rabbit interface is that (it looks as if) it behaves in accordance with our intentional stance. At first blush the distinguishing feature of the SERA rabbits was the speech recognition. On reflection the distinguishing feature ws that the PIR meant that our rabbits were proactive about initiating a conversation. In accordance with the call, our engineering aim was indeed to sense and react to the environment as a robot is expected to do and this meant that it is hard not to think of the rabbit as *wanting* to do things. From the user's perspective, a rabbit has its own agenda and the user slips very easily into taking an intentional stance. Once a conversation was started - as was clear from the video evidence - the rabbits were never good at negotiating *shared* goals. The problem was that the system did not take an intentional stance on the functioning of its user and was thus not able to negotiate a shared intention.

It turns out that the intentional nature of human-human communication is well recognised in linguistics proper. What our rabbits need is a better approach to dialogue management.

4 The language instinct

Computer science as applied to natural language moved out of the arm chair in 1989/90 and that research community generally accept that data driven research is the way forward. The critical mass however use statistical models and, like the behaviourists of old, abhor any notion involving "mental attitudes". In the last 10 years this has been applied to dialogue and so, the argument goes, we do not need to study how language works because (given enough data) machine learning techniques will enable computers to simulate conversational behaviour without theory. Partially Observable Markov Decision Processes (POMDP) have been applied to the dialogue problem [9] and the claim is that the goal of the user - the conversational partner's intent - can be treated as a "hidden state" in a POMDP.

This is a noble aim but in practice human intervention is required to make these systems work. In practice ML techniques are not trained on raw speech or text, but rather on tag sequences where the tags are from a set of dialogue acts or DAs. There is no consensus on what should go into these sets of tags and in general each annotation scheme adapts an accepted set of dialogue acts to the particular application domain. From a linguistics perspective the methodology is sequence analysis [10] which has been unfavourably critiqued by Levinson [11] (page 289), and which in practice produces results with questionable repeatability [12]. Indeed Eduard Hovy at ISI has for some time been pointing out just how much theory is embedded in the choice of DAs and argues for more public discussion of underlying theory [13].

Much of Linguistics and the social scientists however do data driven research but take the line that mental attitudes are causal in human affairs and, as Dennett argues, that a valid science can be based on such concepts. The argument is made beautifully by ten Have [14] but the remainder of this paper is based on the hypothesis by Michael Tomasello [15] that human communication is not only intentional in nature but also a fundamentally cooperative process. Rather than the language instinct being some hard wired ability to recognise mathematical patterns [16], it is the hard wired ability to recognise the intention of others, and the propensity to cooperate in the communicative process.

As is often the case there are notable exceptions - classically Grosz and Sidner [17] talk of attention and intention, and the people working on Max, an embodied conversational agent that has been deployed in the wilds of a museum [18], have talked about attention and mixed initiative at the "discourse level," and in this paper we use a model of intention recognition that has been used in military simulation based on the pre-compiled plans of a BDI architecture [19]. This is discussed further in the next section but first a brief discussion of cooperation may be required.

The need for cooperation is made clear when we take a closer look at what linguists have said about the mechanism of language. Conversation Analysis [20, 21, 14] (CA) is a methodology that enables researchers to notice the detail of language in use and the approach has certainly been prolific. Seedhouse [22] summarises the findings of CA as follows. At any point in a conversation, an utterance will go **seen but unnoticed** in that it is (one of a small set of) expected response, it will go **noticed but accounted for** where it wasn't the expected response but the recipient could figure out why it was said, or the utterance will **risk sanction**. When talking to computers, the sanction is swearing [23] or users not wanting "to use the system on a regular basis" [24]. The point is that the "accounting for" requires us to work hard at recognizing the intent of the speaker. Consider the text book example from Eggins and Slade with which they introduce the notion of sequential relevance:

> **A:** What's that floating in the wine?
> **B:** There aren't any other solutions.
> You will try very hard to find a way of interpreting B's turn as somehow an answer to A's question, even though there is no obvious link between them, apart from their appearance in sequence. Perhaps you will have decided that B took a common solution to a resistant wine cork and poked it through into the bottle, and it was floating in the wine. Whatever explanation you came up with, it is unlikely that you looked at the example and simply said 'it doesn't make sense', so strong is the implication that adjacent turns relate to each other [25].

The appearance of an utterance immediately after another in an interaction to which the partners are committed (that is, a conversation) causes the hearer to work hard at recognising the intent of the speaker. The social pressure on doing this and cooperating in general is captured by Tomasello. Quoting at length:

> Thus, from the production side, we humans must communicate with others or we will be thought pathological; we must request only things that are reasonable or we will be thought rude; and we must attempt to inform and share things with others in ways that are relevant and appropriate or we will be thought socially weird and will have no friends. From the comprehension side, we again must participate, or we will be thought pathological; and we must help, accept offered help and information, and share feelings with others, or we risk social estrangement. The simple fact is that, as in many domains of human social life, mutual expectations, when put into the public arena, turn into policable social norms and obligations. The evolutionary bases of this normative dimension of human communication in terms of public reputation, will be ... [15]

Looking again at Eggins and Slade's example, apes it seems are not hardwired, preprogrammed and/or socialized into putting the effort in and would simply say "it doesn't make sense" and move on. What our computers need as

social actors is the ability to *account for* the communicative acts of its human companions and to do that requires intention recognition and a willingness to put in the effort.

5 Practical intention recognition

Intention recognition and pro-active cooperation are core to human communication of all kinds but it is not enough to say this or even prove it. If researchers in an engineering faculty are going to embrace it, there needs to be a means of implementing it, and the rest of this paper shows how this can be done for limited, but useful, cases. As is often the case with non-incremental development [26] a holistic solution can result in problems cancelling each other out. The challenge of intention recognition and the challenge of proactively helping can be beneficially addressed together by working from a pool of pre-compiled partial plans as used in BDI agent architectures [27–29]. The limited domain used to demonstrate the process is the very applied task of accessing information in a relational database via the telephone.

For the next project the aim is to demonstrate an intentional interface with an IVR system and the scenario under consideration is the classic directory assistance application. With these systems a caller can ring the institution and talk to a computer which puts them through to the required individual. The classic approach would hold the relevant information in a relational database and, in the spirit of Meaning-Text Theory [30] would focus on the information. Consider:

M/C	Welcome to University of Sheffield directory assistance. Who do you wish to contact?
USR	Mark Hepple please.

An ASR module would produce the text from the voice signal, a parser normalize the grammar, and a language understanding module might map that into a canonical representation of the meaning. In the case of database access, the canonical form might take the shape of the SQL query:

> SELECT ALL FROM phonebook WHERE "familyName='Hepple' & givenName='mark"'

For the University of Sheffield phone book, such a query returns:

givenName	**familyName**	**dept**	**extn**
Mark	Hepple	DCS	21829

From this result a text generation system that uses some form of pronoun rewriting could say "His number is 21829."

This is great when things go well. Trouble occurs however when the user's query does not return a single row.

5.1 Trouble in text

If the caller asks for Mark Hawley in Computer Science, the resulting query returns no rows as Hawley is not in the Department of Computer Science. What should the system do?

Using a classic HCI approach the aim would be to make it clear to the user that he or she is using a relational database and to point out that it is the user's query that is resulting in unhelpful output. For many of the DARPA Communicator systems a common solution to no result (no rows in the resulting table) was to remind the caller that he or she could change their query by adjusting the parameters. The user is using a tool, and it is the user's responsibility to use it as *designed*.

At the other extreme, if the caller asks for Mark in Computer Science, the resulting query returns:

givenName	**familyName**	**dept**	**extn**
Mark	Hepple	DCS	21829
Mark	Stevenson	DCS	21921
Mark	Ellerby	DCS	21856

But which Mark does he or she mean? When the user's query matches multiple rows a graphical user interface can present all the rows and this is sometimes attempted with IVR systems. Once again in the spirit of the Communicator systems the system might say:

M/C:	"There are 16 people with that name, the first is Mark Heppe in Computer Science. Is that who you are after?"
Usr:	No
M/C:	The second is Mark Stevenson in Computer Science. Is that the person you are after?
Usr:	No
	...

The computer as tool metaphor may work, but can an intentional approach provide an alternative?

5.2 An Intentional interface

It seems a computer behaving as a social actor needs not only to be right, but also seen to be helpful, and the challenge in the first instance is to come up with helping strategies that the system can introduce. Introducing a strategy requires mixed initiative, not just at the information level, but at the level of intent. The following discussion shows what this means and does it in terms of conversational strategies implemented as plans in a BDI architecture.

The Belief, Desire and Intention architecture was introduced by the software agents community as a means of balancing reactive and deliberative behaviour in a constantly changing environment. The approach does not *do* planning in the traditional AI sense, but rather manages commitment to plans. The usual

BDI approach is to work from a library of pre-compiled plans and "intention recognition" can be implemented (in a limited sense) as a variant of plan choice.

In the case of 2 or perhaps 3 rows, the HCI approach of presenting the list can be used:

Usr:	Mark in Computer Science
M/C:	Mark Hepple or Mark Stevenson?
Usr:	Stevenson
M/C:	Mark Stevenson is on 219...

This strategy is good as far as it goes, but the machine's question assumes the user knows. Thus, this helping strategy might fail if the user is unsure. For a BDI architecture this is not a problem — the architecture was introduced to handle plan failure — and the system simply looks for another plan. The success or failure of this plan will of course depend on what the user says next. Failure however is not bad; what is important is that the system is seen to be be trying to help. Consider:

Usr:	Mark in Computer Science
M/C:	Mark Hepple or Mark Stevenson?
Usr:	Err I was talking with Mark about doing a masters course
M/C:	Mark Hepple is the masters coordinator.
M/C:	Mark Hepple is on 219...

Which is a successful outcome based on the system having a strategy in the plan library for callers looking for information on the masters programme. Critically however it is socially acceptable (i.e. does not risk sanction) for the user's plan to fail:

Usr:	Mark in Computer Science
M/C:	Mark Hepple or Mark Stevenson?
Usr:	Err I was talking with Mark about doing a masters course
M/C:	Right.
	...

The point here is the "unfolding" of the conversation and, like a game of football, plan failure is routine. What matters is that the system is seen to be trying so it does not "risk social estrangement ... and have no friends".

If the user's query returns no rows, it is the system that knows what it has and the machine can push information:

Usr:	Mark Hawley in Computer Science please
M/C:	Err no Mark Hawley in Computer Science. (1 second)
M/C:	There is a Mark Hawley in Health? (1 second)
M/C:	I can give you the number for the Departmental Secretary in Computer Science?
Usr:	Mark Hawley please
M/C:	Professor Mark Hawley in the School of Health and Related Research is on 219..."

Once again the point is the "unfolding" of conversation and a socially ept intentional interface has a responsibility to help.

The fourth case is where there are many rows in the table - 0,1,2,many - and when this happens there is often a misunderstanding. Consider someone who thinks he is calling the number for the Department of Computer Science and says:

M/C:	Good morning how can I help?
Usr:	Mark please.
M/C:	Err you have called directory assistance for the University of Sheffield.
M/C:	I'm sorry, who are you after?

Putting on one's CA hat, the "work done" but the machine's response is to appeal to the caller's sense of fairness. As Tomasello says, people have a sense of fairness and the strategy here is for the system to explain what its job is, suggesting that it is unfair to expect it to be able to help in this case.

Intention recognition is hard for a machine but we can get some way there by working from a fixed set of plans. At this stage the above conversational strategies have been implemented but the system has not been evaluated in an operational setting at this stage. The point of this paper however has been to introduce an alternate model for HCI, and to demonstrate that it is not just hand waving - Tomasello's claims are concrete and implementable.

6 Conclusion

ICT is amazingly versatile, enabling us to create the information systems we want, with the interfaces we want. Without limitations, the designer is ultimately responsible for *any* problems. It is very tempting in these circumstances for us to favour interfaces that exploit the user's *design stance* which shifts some responsibility to the user - the user *ought* to RTFM (read the manual) and then wield the tool as we designed it to be wielded.

The sexy new interfaces - be it 2010 or 1985 - exploit the user's *physical stance* in which our understanding of cause and effect in the physical world is mapped onto virtual events.

The claim being made is that the *essence* of "human-like" interfaces — from embodied conversational agents to robot companions through chat-bots to speech interfaces and IVR systems — is that the user takes an *intentional stance*. Although making these systems more like humans is interesting in its own right — adding micro movements to ECA, emotion or persona to chat-bots — the feature of human communication that provides an opportunity for HCI is the intentional nature of the human interface. This is not enough however because, according to Tomasello, a social actor in human society also needs to be *cooperative*.

Such claims might be seen as too abstract, but the paper gives an interpretation of these principles in the context of an IVR system providing directory assistance.

References

1. Dennett, D.C.: The Intentional Stance. The MIT Press, Cambridge, MA (1987)
2. : The companions project (2007) http://www.companions-project.org/.
3. : Nabaztag (2010) http://www.violet.net/_nabaztag-the-first-rabbit-connected-to-the-internet.html.
4. Prochaska, J., Velicer, W.: The transtheoretical model of behaviour change. American Journal of Health Promotion **12** (1997) 38–48 TTM or ttm.
5. Wallis, P.: A robot in the kitchen. In: ACL Workshop WS12: Companionable Dialogue Systems, Uppsala (2010)
6. Wallis, P.: From data to design. Applied Artificial Intelligence **25** (June 2011) 530–548
7. Sharp, H., Rogers, Y., Preece, J.: Interaction Design: beyond human-computer interaction (2ed). John Wiley and Sons, Chichester, UK (2007)
8. Suchman, L.A.: Plans and situated actions - the problem of human-machine communication. Learning in doing: social,cognitive,and computational perspectives. Cambridge University Press (1987)
9. Young, S.J.: Spoken dialogue management using partially observable markov decision processes (2007) EPSRC Reference: EP/F013930/1.
10. Bakeman, R., Gottman, J.M.: Observing Interaction: An Introduction to Sequential Analysis. Cambridge University Press (1997)
11. Levinson, S.C.: Pragmatics. Cambridge University Press (2000) discussion of discourse analysis and mark up is page 289.
12. Carletta, J., Isard, A., Isard, S., Kowtko, J.C., Doherty-Sneddon, G., Anderson, A.H.: The reliability of a dialogue structure coding scheme. Computational Linguistics **23**(1) (1997) 13–31
13. Hovy, E.: Injecting linguistics into nlp by annotation (July 2010) Invited talk, ACL Workshop 6, NLP and Linguistics: Finding the Common Ground.
14. ten Have, P.: Doing Conversation Analysis: A Practical Guide (Introducing Qualitative Methods). SAGE Publications (1999)
15. Tomasello, M.: Origins of Human Communication. The MIT Press, Cambridge, Massachusetts (2008)
16. Pinker, S.: The Language Instinct. Penguin Books, London (1994)
17. Grosz, B., Sidner, C.: Attention, intention, and the structure of discourse. Computational Linguistics **12**(3) (1986) 175–204

18. Kopp, S., Gesellensetter, L., Kramer, N., Wachsmuth, I.: A conversational agent as museum guide - design and evaluation of a real-world application. In: 5th International working conference on Intelligent Virtual Characters. (2005) http://iva05.unipi.gr/index.html.
19. Heinze, C.: Modelling intention recognition for intelligent agent systems (November 2004)
20. Sacks, H., Schegloff, E., Jefferson, G.: A simplest systematics for the organisation of turntaking in conversation. Language **50**(4) (1974) 696–735
21. Hutchby, I., Wooffitt, R.: Conversation Analysis: principles, practices, and applications. Polity Press (1998)
22. Seedhouse, P.: The Interactional Architecture of the Language Classroom: A Conversation Analysis Perspective. Blackwell (September 2004)
23. Wallis, P.: Robust normative systems: What happens when a normative system fails? In Antonella de Angeli, S.B., Wallis, P., eds.: Abuse: the darker side of human-computer interaction, Rome (September 2005)
24. Wallis, P.: Revisiting the DARPA communicator data using Conversation Analysis. Interaction Studies **9**(3) (October 2008)
25. Eggins, S., Slade, D.: Analysing Casual Conversation. Cassell, Wellington House, 125 Strand, London (1997)
26. Constant, E.W.: The Origins of the turbojet revolution. The John Hopkins Press Ltd, London (1980)
27. Bratman, M.E., Israel, D.J., Pollack, M.E.: Plans and resource-bound practical reasoning. Computational Intelligence **4** (1988) 349–355
28. Rao, A., Georgeff, M.: BDI agents: from theory to practice. Technical Report TR-56, Australian Artificial Intelligence Institute, Melbourne, Australia (1995)
29. Wooldridge, M.: Reasoning about Rational Agents. The MIT Press, Cambridge, MA (2000)
30. Mel'cuk, I.: Meaning-text models: a recent trend in soviet linguistics. Annual Review of Anthropology **10** (1981) 27–62

Taking Things at Face Value: How Stance Informs Politeness of Virtual Agents

Jeroen Linssen, Mariët Theune, Dirk Heylen

Human Media Interaction, University of Twente
P.O. Box 217, 7500 AE, Enschede, The Netherlands
j.m.linssen@utwente.nl

Abstract. In this paper, we contend that interpersonal circumplex theories and politeness strategies may be combined to inform the generation of social behaviours for virtual agents. We show how stances from the interpersonal circumplex correspond to certain politeness strategies and present the results of a small pilot study that partially supports our approach. Our goal is to implement this model in a serious game for police training.

1 Introduction

The automatic generation of social behaviour has been characterized as a 'crucial need' for artificial agents, robots and other intelligent interfaces capable of human-like interaction [14]. In this paper, we focus on social interaction within the field of law enforcement. To assist in the training curriculum of the Dutch police, we are developing a serious game in which police officers will interact with virtual agents to improve their social awareness. How police officers approach and try to reason with civilians and offenders can determine how certain situations are resolved. The Dutch police strive to enforce the law by dealing with conflicts in a de-escalating way. That is, whenever they approach and try to reason with civilians, their goal is to defuse the situation non-aggressively. Being aware of the other's as well as of their own social behaviour is of importance for police officers during such interactions. Therefore, the curriculum of police trainees includes social awareness training. However, these trainings are mainly theoretical, with only few practical training sessions in the form of interaction with actors. Moreover, only a few police officers in training are able to participate in these sessions due to both monetary and time costs—the remaining trainees are restricted to being an audience.

We take on the view that the behaviour of the agents in our serious game should be informed by theories about social interaction that relate to interpersonal attitudes or *stances*. The current training curriculum of police trainees already includes stance theory. We argue that stances are closely related to *politeness*, and propose a mapping of stances to specific combinations of politeness (or impoliteness) strategies.

This paper presents the basis of this approach. First we discuss some related work in section 2. Two theories about stance and politeness are explained in

section 3. In section 4, we discuss the relation between stance and politeness and show how social behaviour can be informed by the combination of the two. We describe a small user experiment carried out to evaluate our model in section 5 and end with conclusions in section 6.

2 Related Work

Past work on social interaction with or between virtual agents has focused on emotions rather than stance [1, 12]. While emotions certainly influence people's behaviour, our approach focuses on people's attitudes toward each other, based on the interpersonal circumplex theory (see section 3.2). Another serious game implementing this theory for human-virtual-agent communication is deLearyous [13]. This game focuses on training interpersonal communication skills in a working environment setting, letting users interact with virtual agents through written natural language input. One of the findings of this project was that determining the stance of dialogue utterances is a very difficult task, even for human annotators.

In our work, we focus on generating utterances that appropriately express the agent's stance. To this end, we combine interpersonal circumplex theories with Brown and Levinson's politeness strategies [2] (see section 3.1). Walker et al. presented one of the first designs for politeness in virtual agents based on these strategies [15]. Their work revolved around using social and affective character traits to inform linguistic style. Gupta et al. continued this work by implementing Brown and Levinson's politeness strategies in POLLy, a system which features a collaborative task-oriented dialogue [5]. They showed that users' perception of the level of politeness of the strategies was largely consistent with Brown and Levinson's theory. Porayska-Pomsta and Mellish implemented a virtual tutor which relies on case-based reasoning to determine which politeness strategy to use [10]. Unlike the work presented in this paper, these previous approaches did not explicitly involve interpersonal attitudes.

3 Theoretical Background

Our model of politeness of social interactions relies on the combination of two theories: the interpersonal circumplex and face theory, which are discussed below.

3.1 Face and Politeness Strategies

Brown and Levinson's work on politeness [2] is based on the notion of *face*, which is a person's public self-image [4]. Brown and Levinson (hereafter, B&L) distinguish between negative and positive face, which denote one's need for freedom and one's need to be approved of and approving of others, respectively. By taking an action, a speaker potentially imposes on a hearer's face by threatening the latter's needs—such an action is called a face-threatening act (FTA). B&L

discuss which strategies can be used to minimize the imposition of an FTA—in other words, how one can be polite. They distinguish the following four strategy types to do so, ordered from least to most polite:[1]

Bald on-record Being straight to the point, e.g., "Hand me the book."
Positive politeness Taking the other's wants into account, e.g., "Would you like to hand me the book?"
Negative politeness Not hindering the other's autonomy, e.g., "If it's not inconvenient to you, could you hand me the book?"
Off record Being indirect or vague about one's own wants, e.g., "I don't seem to be able to reach that book."

Obviously, these politeness strategies do not take into account that people might not want to minimize imposition of their FTAs. Being able to deal with *impoliteness* is especially important for the law enforcement domain, in which police officers and offenders may not care much about each other's face needs, leading to dominant or (verbally) aggressive behaviour. To account for such behaviour, Culpeper et al. [3] investigated impoliteness strategies that are complementary to B&L's strategies. They focus on impoliteness strategies through which the speaker attacks the addressee's positive and negative face needs. Indeed, these are the inverse of B&L's positive and negative face strategies:[2]

Positive impoliteness Damaging the addressee's positive face wants by excluding him or her, being disinterested, disassociating oneself from the addressee or using taboo words. E.g., "Just hand me the bloody book and leave me alone."
Negative impoliteness Damaging the addressee's negative face wants by being condescending, frightening him or her or invading his or her space. E.g., "Hand me the book now, or I'll come and get it."

3.2 The Interpersonal Circumplex

Originating in Leary's work [9] as a tool for diagnosis in a psychotherapeutic setting, a number of varying interpersonal circumplex (IPC) measures of personality have been developed; see [6] for an overview. The IPC model classifies attitudes people have toward each other along two axes: that of dominance and that of affection.[3] Dominance refers to the concepts of one's own autonomy and control over others, while affection stands for affiliating and being accommodating toward or approving of others.

Evaluations of the IPC show that each degree of dominance and affection corresponds to a *stance* [6]. Scherer defines stances as being "characteristic of

[1] An exhaustive list of instantiations of these strategies can be found in [2].
[2] See [3, p. 1555] for more examples of impoliteness strategies.
[3] We adopt these terms as we feel they are clear and unambiguous; variants include 'agency and communion' [6], 'autonomy and friendliness' and 'dominance and sociability'.

an affective style that spontaneously develops or is strategically employed in the interaction with a person or a group of persons, colouring the interpersonal exchange in that situation," [11, p. 705]. For example, one might be dominant and hostile toward someone (high dominance, low affection), resulting in an 'arrogant' stance, or one may adopt a submissive and affectionate attitude (low dominance, high affection), which results in an 'agreeable' stance. Figure 1 shows an example mapping of these two dimensions to a circle and a division into eight octants, each of these corresponding to a stance with a descriptive adjective based on the Interpersonal Adjective Scales [16].

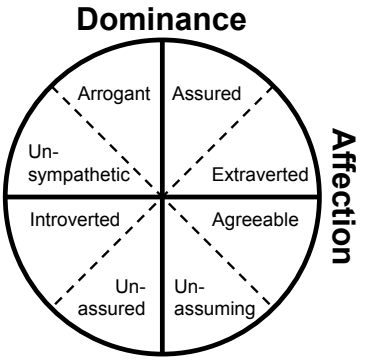


Fig. 1. The interpersonal circumplex, a model which splits social interaction into eight different stances according to the axes of dominance and affection (based on [16]).

4 Stance and Politeness Model

In this section, we propose a model for generating politeness strategies. This model is based on two ideas: (1) politeness strategies can be mixed (section 4.1) and (2) the interpersonal circumplex and politeness theories about face are based on the same principles (section 4.2). In section 4.3, we explain how the model can be used to construct actions for socially interacting agents.

4.1 Mixing Politeness Strategies

Most computational approaches to politeness look at how face-threatening certain acts are by ranking the face threats of those acts in varying ways. For example, following B&L, Walker et al. sum the social distance between the interaction partners, the relative power of one over the other and a static value for imposition of the act [2, 15]. Based on the result, one of the four politeness strategies is then selected to realise the speech act, with the more polite strategies (negative or off record) being used for the bigger face threats.

In our opinion, such a one-dimensional ranking of face threats and politeness strategies disregards the basis of Brown and Levinson's politeness theory, namely

that an act may threaten both positive and negative face. This suggests that a combination of strategies could be used to minimize both impositions. However, B&L oppose the idea of mixing their strategies to express an FTA. They are aware of such mixing occurring in natural discourse, but assert that such utterances express multiple FTAs which need to be ranked separately. Nonetheless, Hasegawa shows that (in Japanese) counterexamples do exist [7]. This view is supported by the observation of Porayska-Pomsta and Mellish that linguistic politeness strategies can address positive and negative face at the same time, and should be classified two-dimensionally [10]. Culpeper et al. show that this also holds for impoliteness strategies, as the positive impoliteness strategy of using taboo words can be mixed straightforwardly with negative impoliteness strategies [3, p. 1561] by simply inserting such words in negative impolite utterances. Therefore, in our model we assume that mixing politeness strategies is possible.

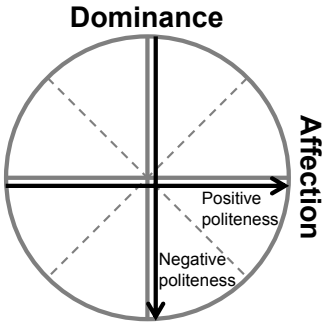


Fig. 2. The relation between the two dimensions of the IPC (dominance and affection) and the two types of politeness (positive and negative). Positive politeness and affection are directly proportional, while negative politeness and dominance have an inverse relation.

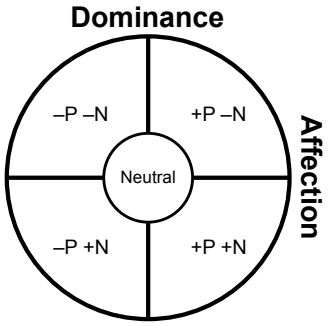


Fig. 3. The mapping of politeness strategies to the IPC. N and P denote 'negative' and 'positive', while the $+$ and $-$ signs denote politeness and impoliteness respectively.

4.2 Combining Face and Stance

Intrinsic to both IPC theories and B&L's politeness theory is that they feature interpersonal relations. Moreover, attitude and stance toward interaction partners play a key part in the choice of actions and the way they are carried out. Dominance and affection, the two dimensions of the IPC, are very similar to the concepts of negative and positive face, respectively. Clearly, dominance revolves around the notion of a person's autonomy. Where the IPC is concerned, this dimension signifies the person's own autonomy, whereas negative politeness strategies address the other's autonomy. As the autonomy of both parties is inversely related, we equate a low value for dominance in the IPC to a high negative face value and vice versa. In other words, when a speaker expresses

little agency, he acts submissively and only threatens the hearer's negative face to a small degree. Similarly, we correlate the dimension of positive face—striving toward acceptance and being approved by others—to that of affection. In this case, a low value of affection corresponds to being 'disconnected' [8], which is directly related to not taking into account the hearer's positive face. Figure 2 shows how negative and positive politeness can be mapped to dominance and affection. When the intention of a speaker is neither to attack an addressee's face (be impolite) nor to weaken his FTA (be polite), we assume that he or she will use B&L's 'bald on record' strategy. In the IPC, this strategy corresponds to having a 'neutral' stance, which is found at the origin of the IPC's axes.

Our model does not include off record strategies at this point. Gupta et al. showed that off record strategies are not necessarily the most polite, as claimed by B&L [5]. Culpeper et al. suggest a structure parallel to that of politeness to resolve this [3, p. 1554], but this is outside the scope of this paper. Figure 3 shows the mapping of the different combinations of politeness strategies as well as the inclusion of the neutral 'bald on record' strategy.

4.3 Utterance Realisations

As shown above, the different politeness strategies addressing negative and positive face can be mapped straightforwardly to IPC stances. Thus, we can construct actions for a given stance by combining the politeness strategies that correspond to that stance. We limit our approach by only taking five stances into account, namely the four combinations of high and low dominance or affection and a fifth 'neutral' stance which represents the origin of the two axes of the IPC. That is, we do not divide the IPC into eight stances as in Fig. 1, but take the stances of each of the four quarters of the IPC (for example, 'arrogant' and 'unsympathetic') together as one stance, as shown in Fig. 3.

Based on the intention of a speaker and a given stance, we can realise an utterance within a given scenario. For example, in a situation in which a few loitering juveniles are playing loud music on a square, the police officer's intention will probably be to reduce the noise level. He then needs to carry out the act of asking the juveniles to turn down the volume, which both limits their freedom (a negative face threat) and implies disapproval (a positive face threat). When the police officer has a dominant yet affectionate stance, he will, according to our theory, use a positive politeness strategy combined with a negative impoliteness strategy ($+P\ -N$ in Fig. 3).

We mix different politeness strategies by creating complex sentences consisting of two clauses, each of which is an instantiation of one type of politeness strategy. Since each clause expresses a different dialogue act, this approach seemingly reflects B&L's opinion about how strategies cannot be mixed in one utterance (see section 4.1). However, we see the compound sentence, taken as a whole, as capturing the intention of being dominant and being affectionate concurrently (or, equivalently, being negatively impolite and positively polite at the same time). This is in line with the findings of Porayska-Pomsta and Mellish

[10]. In a corpus of tutoring dialogues they observed complex strategies that consisted of a main strategy used to express the main message of an act, combined with an auxiliary strategy used to express redress.

Table 1. Example utterances based on five different stances and corresponding (im)politeness strategies (from [2, 3]) in different scenarios. A and D stand for affection and dominance, respectively, with the $+$ and $-$ signs and 0 denoting the value of these dimensions (positive, negative and neutral).

Scenario description	**Stance**	**Politeness strategies**	**Utterance**
Loitering juveniles have just told the police officer to go away. The police officer refuses.	$+A+D$	$+P-N$ (convey cooperation, condescend/ridicule)	"As if I would take orders from you! We can work this out together."
Juveniles are smoking in a shopping mall; the police officer wants to inform them this is not allowed.	$+A-D$	$+P+N$ (raise common ground, question)	"I like a smoke now and then as well, but did you know that smoking isn't actually allowed here?"
Loitering juveniles are playing loud music; the police officer wants them to dim the noise.	$-A+D$	$-P-N$ (unsympathetic, invade space)	"What a racket! You have to stop this immediately."
The police officer has just asked the juveniles to move away, but after a short discussion he decides to let them stay against his will.	$-A-D$	$-P+N$ (disassociate, apologize)	"I'm sorry to have bothered you, but this is going nowhere anyway."
Juveniles are bothering passers-by in a shopping mall. The police officer wants to make clear that people are feeling harassed.	$0A\,0D$ (neutral)	$0P\,0N$ (bald on record)	"Some people feel harassed by you."

In the example scenario, the positive politeness strategy of a police officer would for instance be to say "I understand that you want to chill and listen to music," through which he tries to claim common ground and attend to the juveniles' interests. The negative impoliteness strategy could be instantiated by saying "You have to stop this immediately," which shows the police officer's resolve to impose on the juveniles' autonomy. Taken together, these two sentences will be the police officer's utterance when he takes an affectionate but dominant stance: "I understand that you want to chill and listen to music, but you

have to stop this immediately." Table 1 lists a variety of example utterances (translated from Dutch) that we constructed based on different scenarios and different stances of a police officer toward a group of loitering juveniles. These and other utterances were used in a small user experiment to evaluate our model, as described below.

5 Pilot Study

To validate our ideas about the relations between a person's stance and the politeness of that person's utterances, we conducted a small user experiment. By means of a survey we intended to find out whether politeness strategies indeed correspond to stances as proposed in our model.

5.1 Method and Measures

We carried out an online survey in which participants were asked to give their opinion about the stance of a police officer who is addressing a group of loitering juveniles in various scenarios. For this survey, we constructed a collection of utterances for the police officer based on five different stances, as described in the previous section. In the design of these utterances, we took two additional factors into account, namely that both speech act types as well as contextual content of utterances may influence the face-threat of an act, as noted by Walker et al. [15]. Therefore, we designed utterances for four different speech act types, namely *inform*, *request*, *reject* and *acknowledge*. For each of these speech act types, we conceived two scenarios with a different context to provide a broad collection of situations. For example, for the *request* speech act type, we let the police officer ask juveniles to turn down their loud music in one scenario and let him ask the juveniles to move away from their hangout place in another. Per scenario, we constructed six utterances. Five of these were constructed as explained in the previous section and one was a 'distractor' item. The latter was devised to offer more variety in the survey as well as to make it harder for participants to see through the pattern of the survey questions. In total, we created 48 utterances across 8 scenarios, with each scenario containing 6 utterances of which 5 according to different stances and one being a distractor.

At the beginning of the survey, we explained to the participants that they had to judge the stance of the police officer based on his utterances. We explained that they should do so by rating the police officer's intended dominance and affection toward the juveniles. Participants could indicate their ratings of dominance and affection on two distinct Likert-scales ranging from 1 to 5, where 1 stood for 'not at all' and 5 for 'completely'. Furthermore, we made clear that only verbal actions were included in the scenarios and that none of these utterances should be taken to be sarcastic or ironic.

After having read the instructions and having indicated they understood them, the participants were presented with one of the eight different scenarios and the six corresponding police officer utterances. After rating the intended

dominance and affection of the police officer, participants were asked if they had any comments or critique on the utterances, which they could write down in a text input field. Then, they could continue to the next scenario. Finally, we collected information on the participants' age and gender. We also asked them about their familiarity with interpersonal circumplex theories and theories on politeness and face, as such familiarity might have influenced their judgements.

Table 2. Mean ratings of utterances per stance ($n = 144$; 8 utterances per stance $\times$ 18 participants). T-test values are indicated where significant; * means $p < .05$, ** means $p < .005$.

Stance	**Politeness**	**Means (SD)**		**T-test** ($t(17), value = 3$)	
		Affection	**Dominance**	**Affection**	**Dominance**
$+A\ +D$	$+P\ -N$	3.10 (.68)	2.78 (.42)	n.s.	-2.24 *
$+A\ -D$	$+P\ +N$	2.96 (.67)	2.83 (.49)	n.s.	n.s.
$-A\ +D$	$-P\ -N$	2.49 (.55)	3.35 (.63)	-3.94 **	2.35 *
$-A\ -D$	$-P\ +N$	2.68 (.54)	2.71 (.63)	-2.51 *	n.s.
$0A\ 0D$	$0P\ 0N$	2.57 (.54)	3.21 (.53)	-3.41 **	n.s.

5.2 Results

A total of 18 participants took part in our survey, of which 9 males, 8 females and one person who did not wish to indicate his or her gender. The average age of the participants was 29.9 ($SD = 9.3$). The majority of the participants (13) indicated that they did not know or had only heard of the IPC; 5 knew the basics of the theory or had more in-depth understanding. Almost all participants (15) indicated that they had never heard of B&L's politeness strategies.

We calculated the means of the participants' ratings of the utterances for each of the five described stances; $n = 144$ utterances per stance (8 scenarios, 18 participants). Then, we performed one-sample t-tests to investigate whether the mean ratings of utterances were significantly different from the neutral values for dominance and affection (in both cases, the neutral value was 3, the middle of our Likert-scales which ran from 1 to 5). Table 2 shows the means, standard deviations and one-sample t-test results. Only the most impolite ($-P\ -N$) utterances had average ratings for both dominance and affection that differed significantly from the neutral value. In all other cases, at most one of the ratings differed from neutral, and not always in the predicted direction. Interestingly, the mean affection rating of neutral stance utterances ($0A\ 0D$) did differ significantly from the (in this case desired) neutral value.

Next, we investigated the differences between the means of the different stance utterances through paired-samples t-tests. Here, the most obvious dif-

ference was between the most impolite ($-A$ $+D$, $-P$ $-N$) and the most polite ($+A$ $-D$, $+P$ $+N$) utterances; $t(17) = 7.34, p < .001$ for dominance and $t(17) = -6.73, p < .001$ for affection. These results show that, indeed, utterances combining two (negative and positive) impoliteness strategies were rated as more dominant and less affectionate than utterances that combined two politeness strategies. Similar results were achieved when comparing the most impolite ($-A$ $+D$) utterances with the other utterances; the only type of utterance that did not differ significantly from the $-A$ $+D$ category was the neutral type.

Most of the purely polite or impolite ($+P$ $+N$ and $-P$ $-N$) utterances proved to differ significantly from the utterances that combined polite and impolite strategies ($+P$ $-N$ and $-P$ $+N$, used to express the $+A$ $+D$ and $-A$ $-D$ stances respectively). Specifically, they differed in the mean rating of the stance dimension that was varied between the utterances. For example, the purely impolite $-A$ $+D$ utterances were rated as significantly more dominant than the 'mixed' $-A$ $-D$ utterances; $t(17) = 6.74, p < .001$. Similarly, ratings of affection for the impolite $-A$ $+D$ utterances were significantly lower than for the $+A + D$ utterances; $t(17) = -8.04, p < .001$. Yet in the latter comparison, the ratings of dominance were also significantly higher for the $-A$ $+D$ utterances than for the $+A + D$ utterances, even though this dimension was not varied between the two cases; $t(17) = 5.83, p < .001$. This unexpected difference is caused by the low dominance ratings of the mixed utterances expressing the $+A$ $+D$ stance (as shown in Table 2). The opposite effect did however not occur when comparing the ratings of the purely polite $+A$ $-D$ utterances to those of the mixed $+A$ $+D$ utterances; these ratings did not differ significantly.

5.3 Discussion

The results of our pilot study show that, on average, the utterances we constructed were rated close to the neutral middle of the dominance and affection scales. This lack of 'extreme' utterances may explain why utterances that were intended to be neutral were rated as being as dominant and unaffectionate as those that were intended to be the most dominant and the least affectionate. However, this rating may also be an indication that neutral utterances are, in their directness, indeed always very bald to the point of being impolite.

Some participants commented that they found it hard to judge the police officer's stance based on the presented utterances, as there was (1) no information about the intonation of the utterances and (2) insufficient context to determine how dominant or affective the police officer 'ought' to be. This may be the case because the extreme ends of the scales could be interpreted as the police officer being overly dominant or affectionate, as one participant indicated. Moreover, some of the participants expected the police officers to behave in a much more dominant and much less affectionate fashion than included in the survey.

Nevertheless, the results do confirm our hypothesis that, at least for dominant/unaffectionate and submissive/affectionate stances, B&L's politeness strategies can be used to construct utterances that reflect these stances. Although

mixing positive strategies with negative strategies generally worked well, mixing polite strategies with impolite strategies sometimes resulted in successfully expressing the predicted stance, but at other times resulted in ambiguous utterances (as participants commented). Mixing politeness and impoliteness also caused dominant/affectionate utterances to be rated as much less dominant than predicted.

Based on these findings we see various ways to improve our model. First and foremost, we need to support a wider variety of utterances that cover more gradations of dominance and affection. To do so, we plan to gather more domain knowledge from both police officers and (former) loiterers. This will also help to provide a richer context for the scenarios. Furthermore, we plan to investigate how politeness strategies can be mixed so that they are perceived as less ambiguous. Lastly, we believe that to make utterances better express stances, we should look at the processes underlying the adoption of stances, for example by investigating how people appraise events in terms of face values. In future work, we will first investigate the possible correlations of speech acts and different contexts with the ratings of utterances, as these correlations are not addressed in this paper due to space constraints.

6 Conclusion

Being socially aware is of great importance to police officers during their day-to-day dealings with civilians. To assist them in attaining this awareness, we are designing a serious game that will include virtual agents with which police officers can train their social skills. This paper outlines the first steps we have taken toward creating models that will inform the behaviour of these agents.

Our approach combines the interpersonal circumplex theory [6] and Brown and Levinson's theory about politeness [2]. We assert that both these theories share the same fundamentals of social interaction, namely that people have needs for autonomy (dominance) and for affection. In our model, we state that stances (following the interpersonal circumplex theory) correspond to politeness strategies. That is, an agent with a dominant stance will use negative impoliteness while an agent with an affectionate stance will use positive politeness. Conversely, a submissive stance is expressed through negative politeness and an unaffectionate stance though positive impoliteness. Our second assertion was that these politeness strategies can be mixed to account for all different stances.

To determine the validity of our model, we conducted a small user study in which we let participants rate utterances of police officers on the dominant and affectionate stance dimensions. The results from our experiment support our model in the case of utterances mixing either both positive and negative polite or positive and negative impolite clauses. However, ratings of utterances based on combinations of impolite and polite strategies did not completely meet our expectations, as they were sometimes perceived as ambiguous. To overcome such ambiguity, we intend to investigate in more detail how such utterances influence an addressee's autonomy and affection. Additionally, we plan to gather more

domain knowledge to extend the range of possible utterances. We also need to determine how our agents should react to different utterances. In the end, social interaction does not consist of merely taking stances at face value—this is only the first step.

Acknowledgements This publication was supported by the Dutch national program COMMIT.

References

1. Aylett, R., Dias, J., Paiva, A.: An affectively-driven planner for synthetic characters. In: Proc. of ICAPS. pp. 2–10 (2006)
2. Brown, P., Levinson, S.C.: Politeness: Some universals in language usage. Cambridge University Press, Cambridge (1987)
3. Culpeper, J., Bousfield, D., Wichmann, A.: Impoliteness revisited: with special reference to dynamic and prosodic aspects. J. of Pragmatics 35(10), 1545–1579 (2003)
4. Goffman, E.: The presentation of self in everyday life. Garden City, New York (1959)
5. Gupta, S., Walker, M., Romano, D.: How rude are you? Evaluating politeness and affect in interaction. In: Proc. of ACII. pp. 203–217 (2007)
6. Gurtman, M.B.: Exploring personality with the interpersonal circumplex. Soc. Personal. Psychol. Compass 3(4), 601–619 (2009)
7. Hasegawa, Y.: Simultaneous application of negative and positive politeness. Proc. of CLS 44(1), 125–140 (2008)
8. Horowitz, L.M., Wilson, K.R., Turan, B., Zolotsev, P., Constantino, M.J., Henderson, L.: How interpersonal motives clarify the meaning of interpersonal behavior: A revised circumplex model. Personal. Soc. Psychol. Rev. 10(1), 67–86 (2006)
9. Leary, T.F.: Interpersonal diagnosis of personality. Ronald Press, New York (1957)
10. Porayska-Pomsta, K., Mellish, C.: Modelling politeness in natural language generation. In: Proc. of INLG. pp. 141–150 (2004)
11. Scherer, K.R.: What are emotions? And how can they be measured? Soc. Science Inform. 44(4), 695–729 (2005)
12. Swartout, W.: Lessons learned from virtual humans. AI Mag. 31(1), 9–20 (2010)
13. Vaassen, F., Wauters, J.: deLearyous: Training interpersonal communication skills using unconstrained text input. In: Proc. of ECGBL. pp. 505–513 (2012)
14. Vinciarelli, A., Pantic, M., Heylen, D.K.J., Pelachaud, C., Poggi, I., D'Ericco, F., Schroeder, M.: Bridging the gap between social animal and unsocial machine: A survey of social signal processing. IEEE Trans. Affect. Comput. 3(1), 69–87 (2012)
15. Walker, M.A., Cahn, J.E., Whittaker, S.J.: Improvising linguistic style: Social and affective bases for agent personality. In: Proc. of AA. pp. 96–105 (1997)
16. Wiggins, J.S.: A psychological taxonomy of trait-descriptive terms: The interpersonal domain. J. of Personal. and Soc. Psychol. 37(3), 395 (1979)

Capturing the Implicit – *an iterative approach to enculturing artificial agents*

Peter Wallis and Bruce Edmonds

Centre for Policy Modelling
Manchester Metropolitan University
Manchester, United Kingdom
pwallis@acm.org, bruce@edmonds.name

Abstract. Artificial agents of many kinds increasingly intrude into the human sphere. SatNavs, help systems, automatic telephone answering systems, and even robotic vacuum cleaners are positioned to do more than exist on the side-lines as potential tools. These devices, intentionally or not, often act in a way that intrudes into our social life. Virtual assistants pop up offering help when an error is encountered, the robot vacuum cleaner starts to clean while one is having tea with the vicar, and automated call handling systems refuse to let you do what you want until you have answered a list of questions. This paper addresses the problem of how to produce artificial agents that are less socially inept. A distinction is drawn between things which are operationally available to us as human conversationalists and the things that are available to a third party (e.g. a scientists or engineer) in terms of an explicit explanation or representation. The former implies a detailed skill at recognising and negotiating the subtle and context-dependent rules of human social interaction, but this skill is largely unconscious – we do not know how we do it, in the sense of the later kind of understanding. The paper proposes a process that bootstraps an incomplete formal functional understanding of human social interaction via an iterative approach using interaction with a native. Each cycle of this iteration entering and correcting a narrative summary of what is happening in recordings of interactions with the automatic agent. This interaction is managed and guided through an "annotators' work bench" that uses the current functional understanding to highlight when user input is not consistent with the current understanding, suggesting alternatives and accepting new suggestions via a structured dialogue. This relies on the fact that people are much better at noticing when dialogue is "wrong" and in making alternate suggestions than theorising about social language use. This, we argue, would allow the iterative process to build up understanding and hence CA scripts that fit better within the human social world. Some preliminary work in this direction is described.

1 Introduction

This paper is focused upon computers that inhabit roles with human origin. In particular, computers that have to converse with people as social actors, in the course of their interactions with them. This is not the only sort of interface of course and some will argue that computers as we know them have perfectly satisfactory interfaces, e.g. those based on the notion that the computers are a tool facilitated by a physical analogue (e.g.

a desktop). However a "social stance" – considering computers as social actors – may allow for a new range of applications to emerge as well as giving new insights into human behaviour, in particular the current limitations of our models of this.

However, when computers are compelled to work as social actors – for example when they use language as the primary modality – they tend to fail grossly rather than in detail. Indeed people get so frustrated by computers that they often swear at them [1]. When someone swears at a carefully crafted chat-bot, the human is unlikely to have been upset by punctuation or a quirky use of pronouns. The challenge is that existing qualitative techniques are good at the detail, but can fail to find a bigger picture.

In this paper the focus is on performative language and builds on the findings of applied linguistics where the mechanism of language can be seen as part of the same spectrum of communicative acts ranging from "body language" to semiotics. However much of these communicative acts are learned in context with reference to their effect rather than a putative explicit meaning.

This is contrary to approach that characterises human actors as rational actors. Applied to language this motivates the characterisation that natural languages are a "fallen" version of something more pure – a messy version of First Order Predicate Calculus – where elements of the language can be associated with their separate meaning. This meaning-text model [2] has been largely rejected since the late 1980's but has a latent existence in the idea that it is possible to create sets of Dialogue Acts (DAs) that capture in some way the primitive concepts from which any conversation can be constructed. For a comprehensive description of the theory and lack thereof in this area, there are several papers by Eduard Hovy [3].

It is also going in a different direction to those focused on statistical and machine learning techniques [4] that treat mental attitudes as a "hidden state" that can be derived from corpora of human behaviour. The advantage of this approach to engineering dialogue systems is that we do not need to understand how language is used, the machine will figure it out for itself (as far as it is able). The challenge is the amount of training data required to cover all the necessary cases and, unlike a search engine where measured performance as low as 10% is useful, many errors social actors make are noticed and need to be dealt with. The assumption here is that we want to know more about the process of being a social actor, and know enough about it to be able to make a computer to do it with sufficient competency.

Rather this paper is predicated on the notion that there is a wealth of vague, implicit, context-dependent and often unconscious knowledge that is necessary for a social actor to successfully inhabit a society [5, 6], and to show how such knowledge might be incorporated into artificial agents. Such social knowledge is not immediately accessible to an engineer as explicit knowledge and so the classic "waterfall" model of software engineering, in which one starts by developing a detailed specification and follows up with a development phase, is inappropriate. Instead, a process of entity enculturation – learning how a CA should behave in context over a period of time – is required. Design plays a part, but has to be leveraged by a substantial subsequent iterative process of trial and repair [7]. This is not a "one off" method of making socially fluent agents, but a method of repeatedly: (1) analysing records of their interaction in situ, then (2) affecting

a repair on the behaviour for this context. Thus, over time, embedding the agent into the culture it perforce inhabits.

The core of this approach is the leveraging of a common narrative understanding of interactions between people. In this non-scientists are asked to "tell the story" of how a particular situation came about with a conversational agent as a starting point in preparation for an iterative approach to repairing that agent. In subsequent iterations they may be asked to comment upon an existing narrative, possibly entering alternative descriptions at certain points. This interaction will be guided and constrained by a a developer's workbench that allows someone to both "script" future dialogue and analyse recordings of past (real) dialogue with the machine by narrating the action, and would capture the *mechanism* by which we social actors decide what to say and when.

2 Contributory Threads

Given the nature of the proposal, and its contrary direction it is useful to trace the projects and results that have lead us in this direction.

2.1 The KT experiments

The KT experiments were a project to understand the issues and the potential for embodied conversational agents (ECA) acting as virtual assistants [8]. As part of that project, we conducted Wizard-of-Oz experiments, where a human covertly pretends to be the conversational agent conducting the conversations, followed by interviewing the wizard (KT) about her actions using a technique from applied psychology called Applied Cognitive Task Analysis (ACTA) [9]. The aim was to populate a model of KT doing the task, and then use that model to drive a virtual assistant performing the same task. The model was "folk psychological" in that it her beliefs, desires and other mental attitudes were used as theory to explain and identify the "causes" of her behaviour. For these experiments the task was simply to have staff call our agent when they wanted to use one of the cars from the Division's car pool. Ultimately the task was a slot-filling task: specifying which car, who was using it and the time.

The relevant results were twofold. *Firstly*, that politeness was more important than getting the facts right. For various reasons KT's "slot-fill rate" - how often she managed to identify a piece of information in the caller's utterances and enter it in the appropriate slot - was just over 80%. A "fact error rate" of close to 20% might sound high but the point is *nobody minded* and, although we didn't measure it, we expect nobody *noticed*. Why didn't they mind or notice? Because of course KT would make appropriate apologies and gave explanations when she had forgotten what they said their phone number was or where they were going. What is more, looking at the length of utterances, it was easy to see how KT's utterances could convey the same information in a more compact form. Grice's Maxims would suggest shorter is better (a principle popular with call centre industry) but KT did not want to use shorter utterances because, from the interview process, it just wouldn't be polite. As a scientist one might have theories about the concept of face [10], but KT's seems to have some system for doing social interaction that

uses politeness as an atomic concept. She had not read Brown and Levinson's book [10] and didn't need to.

Secondly, it turned out that interviewing people about their everyday behaviour is problematic. Interview techniques such as Applied Cognitive Task Analysis are intended to make explicit expert knowledge that has become automatic. A fireman is likely to be proud of his knowledge and pleased when the interviewer can identify some piece of knowledge he had forgotten was special. Using ACTA to interview KT about her "expertise" (which it is) in the use of language however, KT thinks of her knowledge as just common-sense. The knowledge was implicit knowledge – a set of learnt skills as to how to converse. What we were after was exactly that common-sense knowledge in an explicit form so we could model and use it. Unfortunately it is common sense also in that is common to all – it is knowledge that is shared and KT knows that. Interviewing people about their common-sense knowledge, they quickly become suspicious about the interviewer's motivations. Why is he asking such "dumb" questions?

The lessons from this were that it was precisely the implicit social skills in conducting a conversation that were important but also difficult to get at in an explicit form. Just as one can be able to ride a bicycle but not know how one does it, one can conduct a sensible social interaction whilst not being able to specify how one does this. The very ubiquity of this skill hides its subtlety and complexity.

2.2 Ethnomethods

The CA4NLP project applied an ethnomethodological variant of Conversation Analysis [11] to analysing records of conversations such as those produced by the KT experiments. This approach is predicated upon the notion that the researcher is a "member of the same community of practice" as the discussants, and hence has access to the import of their utterances. Thus, for example, a researcher's introspections about whether or not some communicative act of KT's was polite is valid evidence because both get their knowledge about the purpose of communicative acts from the same common pool. I do not need to ask KT about her internal reasoning because it is the external effect that matters and I have direct access to its significance. KT is right: I could give as good an answer to my own dumb questions as she.

This method also implies a shift from a mechanistic view to a functional view. When it comes to engineering spoken language interfaces, rather than trying to access the internal reasoning of the speaker as the KT experiments attempted, we want to look at and model the way a social agent engages with the community of practice in which it operates. Although engineering more as a process of adaption of function than design will make some engineers uncomfortable, this is common practice for long-standing artefacts that inhabit complex niches, such as sailing yatchs – nobody designs a yatch from first-principles but rather adapts and tunes existing designs, tinkering with each aspect in turn. What matters is how the yatch functions within the complex environment of winds and water. The same applies to computers that act in our social space. A computer that says "no records match your request" might be being informative [12] but is it playing by the rules of social engagement? Using the terminology from Conversation Analysis, what is the *work done* by "no records match your request" and is it all and only what the expression was designed to do in the current context?

The methodology of Conversation Analysis is for the scientist to capture naturally occurring text or speech of interest and ask "Why this, in this way, right here?" Whilst using introspection as a means to assess scientific truth is a bad idea, introspection about community knowledge is fine and provides detailed descriptions of the function of utterances in context. Thus the CA4NLP project illustrated the use of introspection to leverage understanding about utterances. It marks a shift away from attempting to access an internal or foundational model, but rather capitalises upon the function of utterances in context. It is the function of utterances that is constrained by common usage, not the cognitive processes that give rise to them.

The trouble with Conversation Analysis however is exactly its strength in that it provides a valid means of studying anything and everything. It does not provide any guidance on what is *critical* to the structure of a conversation.

2.3 HCI and Grounded Theory

The SERA project put a talking "rabbit" in older people's homes and collected video of the resulting real human-robot interactions. 300 or so recordings of people interacting with their rabbits were collected. The experiment had three iterations of: placing the rabbit, recording the interactions, assessing the success of the system, and improving the system software based on the assessment.

The motivation for the project was to see how different research groups would go about this process. In general, all the groups could find interesting things to write about the data, but the process of improving the system was primarily driven by those with an HCI background who would, in the tradition of design-based engineering, simply have an idea that could be tried. This creative process often worked, and would be followed by a quantitative evaluation, but felt quite unsatisfactory when it came to understanding what is going on.

The understanding that did feel like progress actually came from qualitative methods such as Grounded Theory [13] in the form of detailed analyses of how particular conversations unfolded in those contexts. In particular people are very good at noticing when a conversation is NOT right. As an expert I can tell you that I wouldn't say "no records match your request" in a given context and it is this data that needs to be the raw material on which we base a science of machines in social spaces. However, this micro-level of detail poses a problem when one needs to utilise the knowledge, for example in terms of suggesting improvements to CA scripts. The detail needs to somehow be accumulated in a more comprehensive social ability.

In some preliminary experiments in the SERA project, people were asked to say what happened in a video recording we had of people interacting with one of the SERA rabbits. This initially did not work very well because, although the plot in a film or play is easily identified and summarised, natural recordings are just not that interesting and rather messy. Instead recordings where things go wrong was chosen. This made the 'crux' of the story more salient.

2.4 Summary of threads

From the above experience we draw out several lessons. We see the importance of the shared culture in terms of the common folk theory about what is happening, however we also see that this common knowledge is implicit and not very accessible via direct interrogation. We see the importance of examples learnt in context, in particular in terms of their functional fit to the social circumstances. Finally this suggests that, in order to transfer this implicit knowledge we might have to mimic the learning that usually happens within the social sphere in terms of making mistakes and repairing them.

In order to make better conversational agents they will have to be inducted into the society they are going to inhabit. Clearly, in general, this is extremely hard and takes humans a couple of decades of time but here we might be aiming for an agent that copes tolerably well (on the level of a polite 6 year old) in a single context (or a very restricted range of contexts). Here we aim to imitate the cycle of trial, error and repair on a small scale, hoping to make up for the small number of cycles with a more intelligent repair stage composed of analysis with repair leveraging some of our own innate understanding of social behaviour. Each iteration in a particular context will (on the whole) result in an incremental improvement in social behaviour. The hard part of this cycle (other than the number of times it may have to be done) is the analysis and repair stage. We will thus concentrate on this in this section.

3 Capturing the implicit

The idea presented in Figure 1 is to iteratively improve an *in situ* CA, each iteration through allowing a bit more of the explicit *and* implicit knowledge concerning the appropriate social behaviour to be captured in the knowledge base and hence used to tune the CA rules. Each iteration the CA, in its current state of development, will deployed and new records of its conversation with humans made, since it is difficult to predict the full social effect of any change. This iterative cycle imitates, in a rough manner, the way humans learn appropriate social behaviour: observing others, noticing social mistakes and iteratively adapting their behaviour.

Clearly there are several parts of this cycle that could be discussed in detail. However, here we will concentrate on motivating and outlining how the user-interface that prompts and structures the review of the conversational records by the native expert. The nub of this process is how to elicit the, largely implicit, knowledge about social behaviour using the responses of the third party reading and reacting to the records of the conversation.

3.1 Vygotski

Vygotski's insight used here is that plays and novels exist because they provide plausible accounts of human behaviour. Theatre is the flight-simulator of life [14] and provides a means of exercising our ability to understand the motivations and behaviour of others. We do think about other minds when we communicate – indeed it turns out to be a critical skill [15–17] – and we do it in terms of beliefs, desires and other mental

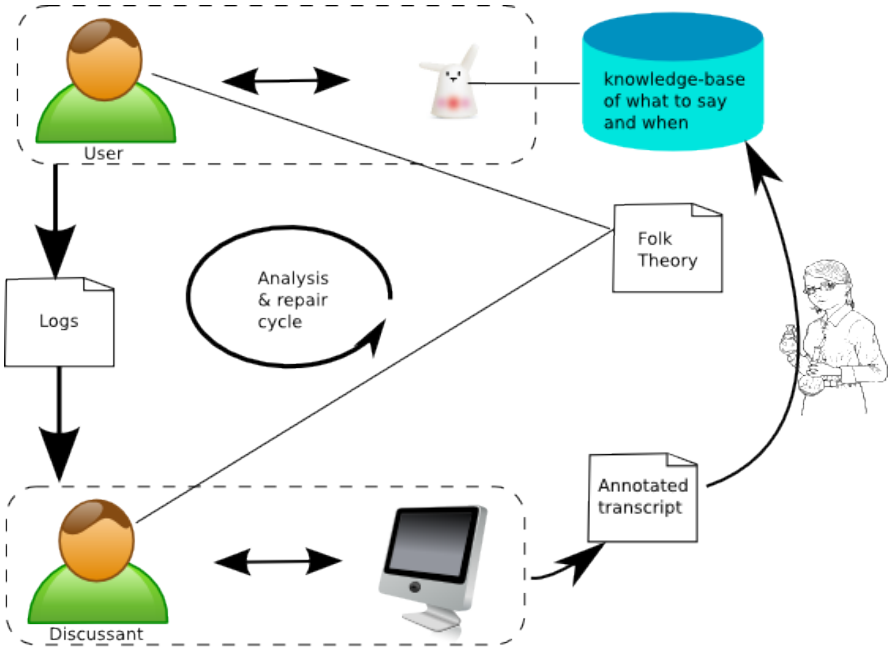


Fig. 1. Summary flow chart of the proposed iterative method

attitudes. What is more, we *expect our conversational partners to do the same*, with the same model. When it comes to communication, the truth of our folk model of other people's thinking doesn't matter; what matters is that it is shared. Rather than looking inside KT's head to see how she would deal with social relations, the idea is to look at some kind of collective understanding of events – what is the shared knowledge that creates the context against which a human social actor figures out the significance of communicative acts? Rather than sitting in an arm-chair and classifying utterances according to the effect they have on a idealised conversational partner, the idea is to look at real interaction data and document the effect in context. Rather than classifying utterances as $REQUEST\_INFORMATION$, or $GREETING$ [18], the idea is to record the "work done" by utterances in the place they are produced. This can be done by any member of the community of communicators and does not require a scientific theory. Consider this example of a conversation between a doctor and a patient taken from the Conversation Analysis literature:

```
Patient: So, this treatment; it won't have any effect on
us having kids will it?
Doctor: [silence]
Patient: It will?
Doctor: I'm afraid the...
```

The “work done” by the silence is of course to disagree and some might be tempted to mark it up as an explicit answer, but there are many different things that the doctor could say at this point, with a wide range of “semantics” but all with the same effect.

The Vygotski argument is that human story telling gives sufficient detail of events that any socialised human (who is part of the same community) can fill in the gaps to produce a set of linked causal relationships for the story to make sense. This requires contextual knowledge (e.g. teddy bears are toys and children like to play with toys) and ”hard-wired” knowledge (e.g. children often *want* things and *act* in ways that bring them about).

One might think that human-machine interactions would be less fraught and thus simpler. Indeed those working on commercial spoken language interface design try very hard to make this true using techniques such as menu choice. However, even in the DARPA Communicator data where the systems were only slightly more natural than those one might find in a bank, there are examples where the work done by an utterance such as “no” goes well beyond what might be seen as the semantics of the utterance [19]. One can not escape the importance of social etiquette.

It is this process, and the interface to support it, that we will now describe.

3.2 An example

Consider some video data captured spontaneously (Figure 2) during the development of the SERA set-up.

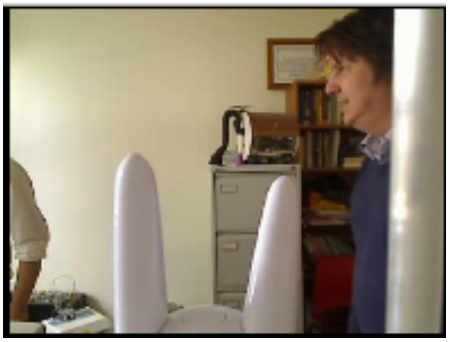


Fig. 2. Mike and the rabbit talking with Peter

To set the context, “Peter and Mike have been talking in Peter’s office where he has a robot rabbit that talks to you and that you can talk to using picture cards.” Two narrators were given this sentence and asked to watch the video. They were then asked to, independently, finish the story in around 200 words. The resulting stories appear in Figure 3.

There are many differences, and many things were left out entirely. There does however appear to be general agreement on core events. Neither narrator mentioned

Narrator 1	Narrator 2
It is time to go home so Peter takes his keys from the rabbit. Mike notices this and says "Isn't it supposed to say hello?" Peter is about to say something when the rabbit says: "Hello, are you going out?" Peter replies that he is (using the card and verbally) and the rabbit tells him to have a good time, bye. Mike picks up a card and shows it to the rabbit, but nothing happens. He thinks this make sense as the rabbit has said goodbye but Peter thinks it should work and shows the rabbit another card. Mike sees that he has been showing the cards to the wrong part of the rabbit and gives it another go. Still nothing happens and Mike tries to wake it up with an exaggerated "HELLO!". Peter stops packing his bag and pays attention. Mike tries getting the rabbits attention by waving his hand at it. Still nothing happens. Mike looks enquiringly at Peter as if to ask "what's happening" He says "that's a new one" and goes back to his packing. Mike takes his leave at this point. Peter finishes his packing, and, as he leaves says to the rabbit "You're looking quite broken."	Peter is about to do something to wake the rabbit up again and as he is about to speak, it says hello. Peter gestures to Mike that it is now talking as expected. Peter presses the video button to record the interaction. Mike laughs as it talks. It asks Peter if he is going out, to which he responds verbally that he is, showing the rabbit the card meaning yes. Seeing Peter's interaction, Mike tries using the cards to interact with the rabbit himself. It does not respond and Mike suggests that this is because it has said goodbye and finished the conversation. Peter tries to reawaken the rabbit with another card. Mike sees that he had put the card in the wrong place. He tries again with a card, after joking that the face card means "I am drunk". Peter laughs. When the rabbit does not respond, Mike says "hello" loudly up to the camera. Peter says he is not sure why there is no response while Mike tries to get a reaction moving his hand in front of the system. They wait to see if anything happens, Mike looking between the rabbit and Peter. When nothing happens, Peter changes topic and they both start to walk away. Mike leaves. As Peter collects some things together, walking past the rabbit, he looks at it. Before leaving the room he says to the rabbit "you're looking quite broken".

Fig. 3. Two narrative descriptions of the same event.

1. Peter is about to say something and is interrupted by the rabbit
2. the rabbit asks if he is going out, Peter's verbal and card response
3. the rabbit says bye
4. Mike's attempt to use a card and the non-response of the rabbit
5. Mike's explanation (that the rabbit has already said bye)
6. and Peter showing the rabbit another card
7. Mike sees that he has been showing the card to the wrong part of the rabbit and has another go
8. the rabbit does not respond
9. Mike says "Hello" loudly
10. Peter acknowledges it doesn't look right
11. Mike tries again by waving his hand in front of the rabbit
12. no response from the rabbit
13. Mike looks at Peter
14. They give up
15. Mike leaves
16. Peter leaves saying "You're looking quite broken" to the rabbit

Fig. 4. The third-party common ground.

the filing cabinet nor the clothes participants were wearing. No comment on accents or word usage; no comment on grammatical structure nor grounding, nor forward and backward looking function. Whatever it is that the narrators attend to, it is different to the type of thing that appear in classic annotation schemes. It does however seem to be shared and, the claim is, shared by the community of practice at large. Both the narrators and the participants are working from a shared theoretical framework – not from raw undifferentiated data – that guides and selects which sense-data is attended to. However this shared framework is implicit.

Accounts of the action in the video data as written down by the narrators are of course *descriptive* in that they are written to 'fit' past events. The claim is that they are also predictive. If Mike wants to use the system, then it would be surprising if others did not want to. If failure to work causes disappointment in Mike, it is likely to also cause it in others. Having a predictive model of events we are well on the way to having *prescriptive* rules that can be used to drive conversational behaviour.

But first however let's look at how we might move more formally from the stories in Figure 3 to the summary in Figure 4.

3.3 An interface to support capture of social knowledge

The problem is that even if they observe the same thing, they may not describe it in the same way and, unless the descriptions are the same, a machine cannot recognise them as the same. In the example above the two observers produced two narrative descriptions and it is claimed they are the same, but how would one *measure* the sameness? Without a machine that can understand what is written, human judgement is involved and claims of researcher bias are possible. How might comparative narratives [20] be produced that are the same to the exacting standards required for machine understanding?

The proposal, should one want to re-do this preliminary experiment properly, is to use the techniques seen in industrial machine translation for the production of operator and repair manuals. Companies like Mitsubishi and Caterpillar [21] have systems that allow them to produce manuals in one language and then, at the push of a button, produce the same manual in all of the languages for countries to which they export. The way this is done is to have the author of the manual write in a restricted version of the source language and provide the tools to guide the writing process. The process of authoring with such tools will be familiar to us all because modern text editors provide spelling and grammar checking assistance in much the same way. The primary differences being of course that the list of recognised words is much smaller and the grammar rules much stricter, and the process of breaking those rules is not simply for the system to ignore it, but to ask the user to add the new word or expression to the system. For instance the author might really want to use the term "airator" and the system would allow that but ask the author if it is a noun or an adjective, a count noun, what its semantic preferences are, and if it is masculine or feminine in French. The word would be added to the lexicon and, the next time an author wanted to use it, the system would have enough detail to translate it correctly or ask this new author how it should be used in the current context.

If one wanted to re-do the experiment above more formally, the approach would be to reproduce a "translators work bench" and, rather than having it translate to another

language, have it "translate" to a different style in the same language. This authoring process works for machine understanding for translation; there is no reason to think it wouldn't work for this new application if one really wanted to do it. But why bother? The ultimate aim is to script dialogue for synthetic characters and the proposal is that, rather than stopping at narrative descriptions, the system would go on to explore counterfactuals.

3.4 Narrative descriptions capturing context

The aim is to classify utterances as the same in context and hence be able to program an agent to give a particular response to any input from the same class. Using a classic annotation scheme one might decide that if its conversational partner produces something in the class of $QUESTION$, then the agent should produce an $ANSWER$. This functionalist model of sameness applies to everything from chatbots in which something like regular expressions are used to recognise inputs are the same, through to full planning systems such as TRAINS [22] in which input recognition is set against the current goals of the system. The variation proposed here is that the functionalist definition of sameness is embedded in narrative. Two expressions are the same if and only if, for every narrative in which expression #1 occurs the outcome of the story would not change if expression #2 was used.

Given such a definition of sameness, it is only in trivial cases that expressions will be universally the same. It is far more likely that expression #1 and #2 will be equivalent for some narratives and not others – the equivalence is context dependent, and this provides an opportunity to question an observer about the features of the context that determine when an existing response to input might or might not be appropriate for another input.

As an example of the type of thing we have in mind Figure refCTAprobes gives a table showing the type of question that was asked of KT. It would appear that some of these questions would be a useful way to explore context with our observers and, importantly, the questioning could be automated. An observer might provide a narrative description of a particular recording of an interaction and, at some point in that description the computer might say S where the rules being used by the machine might have equally produced S'. An annotator's work bench could ask the human if S and S' would be functionally equivalent in the narrative given. If not, the workbench could ask what (in the context) makes S' inappropriate, and perhaps ask the annotator to develop a rule that distinguishes the context for S and S'. Similarly the system could ask the observer if he or she can formulate an alternative to S and S' that would be better, and develop a rule to distinguish the alternative utterance from S and S'.

The above gives a flavour for the proposed work bench designed to enable non scientists to use their expert knowledge of language use to create context dependent rules so the system can decide what to say when. The aim is to combine the direct contact with the data normally seen in an annotation tool such as Anvil [23] with the creative process of scripting conversation for the agent. In effect the aim is to formalise the process (and add some theory) that people use when they script chat-bots using AMIL by pouring over log files.

Fig. 5. O'Hare et al 1998 - the revised CDM probes.

Goal specification	What were your specific goals at the various decision points?
Cue identification	What features were you looking at when you formulated your decision?
Expectancy	Where you expecting to make this type of decision during the course of the event?
	Describe how this affected your decision-making process
Conceptual model	Are there any situations in which your decision would have turned out differently?
	Describe the nature of these situations and the characteristics that would have changed the outcome of your decision.
Influence of uncertainty	At any stage, wee you uncertain about either the reliability or the relevance of the information that you had available?
	At any stage, were you uncertain about the appropriateness of the decision?
Information integration	What was the most important piece of information that you used to formulate the decision?
Situation awareness	What information did you have available to you at the time of the decision?
	What information did you have available to you when formulating the decision?
Situation assessment	Did you use all the information available to you when formulating the decision?
	Was there any additional information that you might have used to assist in the formulation of the decision?
Options	Were there any other alternatives available to you other than the decision that you made?
	Why were these alternatives considered inappropriate?
Decision blocking - stress	Was there any stage during the decision-making process in which you found it difficult to process and integrate the information available?
	Describe precisely the nature of this situation.
Basis of choice	Do you think that you could develop a rule, based on your experience, which could assist another person to make the same decision successfully?
	Why/Why not?
Analogy/generalization	Were you at any time, reminded of previous experiences in which a *similar* decision was made? Were you at any time, reminded of previous experiences in which a *different* decision was made?

4 Conclusion – Towards the Iterative Embedding of Implicit Social Knowledge

This proposed approach seeks to take seriously the subtlety of social behaviour, resulting from the "double hermeneutic" which relies on the fact that encultured actors will have a ready framework of how to interpret the social behaviour of others, including the expectations that others will have of them. In particular it is important how it is that social knowledge is embedded within a complex of social relations and knowledge, which makes it hard to formalise *in general*. We do not expect that this will be easily captured in a "one-off" analysis but require a iterative approach based on repair. The difficulty of the task means that a number of approaches will need to be tried to leverage little bits of social knowledge each iteration. The key parts of this are the interactive capture of social information from a third party and the use of that knowledge to inform an update of the CA rules. We have not talked about the latter here – currently it will require significant programming skill. The ultimate aim would be to eliminate this programmer, so that this iterative process could be used by non-experts, utilising their own implicit expertise, to socially "educate" their own CA. This is illustrated in Figure 6.

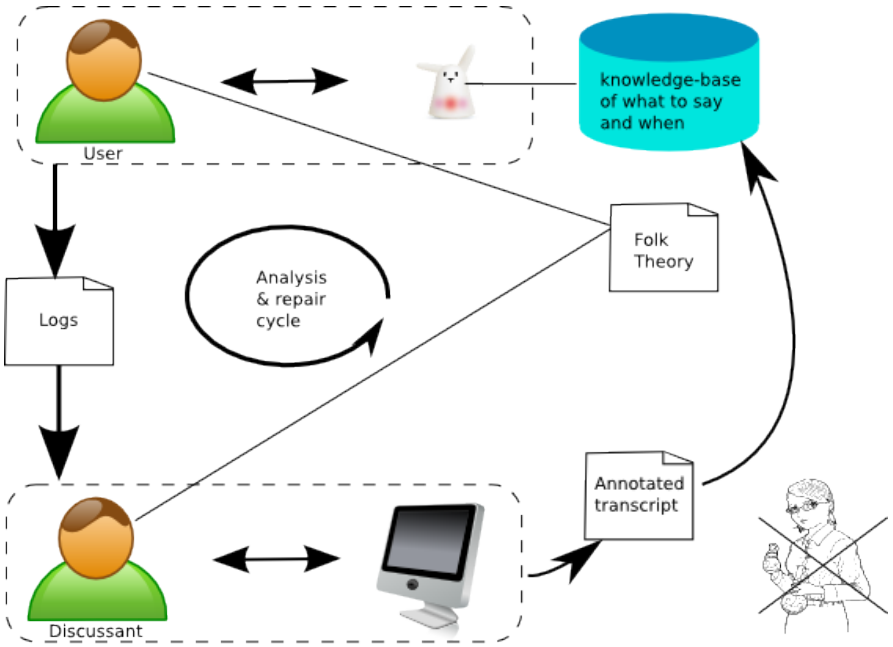


Fig. 6. Flow chart of the process without the programming expert

References

1. de Angeli, A.: Stupid computer! abuse and social identity. In de Angeli, A., Brahnam, S., Wallis, P., eds.: Abuse: the darker side of Human-Computer Interaction (INTERACT '05), Rome (September 2005) http://www.agentabuse.org/.
2. Mel'cuk, I.: Meaning-text models: a recent trend in soviet linguistics. Annual Review of Anthropology **10** (1981) 27–62
3. Hovy, E.: Injecting linguistics into nlp by annotation (July 2010) Invited talk, ACL Workshop 6, NLP and Linguistics: Finding the Common Ground.
4. Young, S.J.: Spoken dialogue management using partially observable markov decision processes (2007) EPSRC Reference: EP/F013930/1.
5. Edmonds, B.: Complexity and scientific modelling. Foundations of Science **5** (2000) 379–390
6. Edmonds, B., Cershenson, C.: Learning, social intelligence and the turing test: Why an 'out-of-the-box' turing machine will not pass the turing test. In Cooper, S.B., Dawar, A., Lwe, B., eds.: Computers in Education. Springer (2012) 183–193 LNCS 7318.
7. Edmonds, B., Bryson, J.: The insufficiency of formal design methods – the necessity of an experimental approach for the understanding and control of complex mas. In: Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multi Agent Systems (AAMAS'04). ACM Press, New York (July 2004) 938–945
8. Wallis, P., Mitchard, H., O'Dea, D., Das, J.: Dialogue modelling for a conversational agent. In Stumptner, M., Corbett, D., Brooks, M., eds.: AI2001: Advances in Artificial Intelligence, 14th Australian Joint Conference on Artificial Intelligence, Adelaide, Australia, Springer (LNAI 2256) (2001)
9. Militello, L.G., Hutton, R.J.: Applied cognitive task analysis (ACTA): a practitioner's toolkit for understanding cognitive task demands. Ergonomics **41**(11) (November 1998) 1618–1641
10. Brown, P., Levinson, S.C.: Politeness: Some Universals in Language Usage. Cambridge University Press (1987)
11. ten Have, P.: Doing Conversation Analysis: A Practical Guide (Introducing Qualitative Methods). SAGE Publications (1999)
12. Traum, D., Bos, J., Cooper, R., Larson, S., Lewin, I., Matheson, C., Poesio, M.: A model of dialogue moves and information state revision. Technical Report D2.1, Human Communication Research Centre, Edinbrough University (1999)
13. Urquhart, C., Lehmann, H., Myers, M.: Putting the theory back into grounded theory: Guidelines for grounded theory studies in information systems. Information Systems Journal **20**(4) (2010) 357–381
14. Bennett, C.: But mr darcy, shouldn't we be taking precautions? Keith Oatley (quoted in) (July 2011)
15. Grosz, B., Sidner, C.: Attention, intention, and the structure of discourse. Computational Linguistics **12**(3) (1986) 175–204
16. Eggins, S., Slade, D.: Analysing Casual Conversation. Cassell, Wellington House, 125 Strand, London (1997)
17. Tomasello, M.: Origins of Human Communication. The MIT Press, Cambridge, Massachusetts (2008)
18. Jurafsky, D., Shriberg, E., Biasca, D.: Switchboard swbd-damsl shallow- discourse-function annotation coders manual. Technical Report 97-01, University of Colorado Institute of Cognitive Science, Colorado (1997)
19. Wallis, P.: Revisiting the DARPA communicator data using Conversation Analysis. Interaction Studies **9**(3) (October 2008)

20. Abell, P.: Comparing case studies: an introduction to comparative narratives. Technical Report CEPDP 103, Centre for Economic Performance, London School of Economics and Political Science, London, UK (1992)
21. Kamprath, C., Adolphson, E., Mitamura, T., Nyberg, E.: Controlled language for multilingual document production: Experience with caterpillar technical english. Proceedings of the Second International Workshop on Controlled Language Applications **146** (1998)
22. Allen, J.F., Schubert, L.K., Ferguson, G., Heeman, P., Hwang, C.H., Kato, T., Light, M., Martin, N.G., Miller, B.W., Poesio, M., Traum, D.R.: The TRAINS project: A case study in defining a conversational planning agent. Journal of Experimental and Theoretical AI **7**(7) (1995) 7–48
23. Kipp, M.: Gesture Generation by Imitation - From Human Behavior to Computer Character Animation. PhD thesis, Saarland University, Saarbruecken, Germany (2004)