A Conversational Agent as Museum Guide – Design and Evaluation of a Real-World Application

Stefan Kopp[1], Lars Gesellensetter[2], Nicole C. Krämer[3] and Ipke Wachsmuth[1]

[1] A.I. Group, University of Bielefeld, P.O. Box 100132, 33501 Bielefeld, Germany
{skopp,ipke}@techfak.uni-bielefeld.de
[2] IPD, University of Karlsruhe, Adenauerring 20a, 76128 Karlsruhe, Germany
lars@ipd.info.uni-karlsruhe.de
[3] Dept. of Psychology, University of Cologne, Bernhard-Feilchenfeld-Str. 11,
50969 Köln, Germany
nicole.kraemer@uni-koeln.de

Abstract. This paper describes an application of the conversational agent *Max* in a real-world setting. The agent is employed as guide in a public computer museum, where he engages with visitors in natural face-to-face communication, provides them with information about the museum or the exhibition, and conducts natural small talk conversations. The design of the system is described with a focus on how the conversational behavior is achieved. Logfiles from interactions between Max and museum visitors were analyzed for the kinds of dialogue people are willing to have with Max. Results indicate that Max engages people in interactions where they are likely to use human-like communication strategies, suggesting the attribution of sociality to the agent.

1 Introduction

Embodied conversational agents (ECAs) begin to show impressive human-like capabilities of natural face-to-face dialogue. Agents of this kind have been successfully developed for various target applications. Yet, it is noteworthy that they are normally designed for specific settings and have rarely made the step out of their laboratories into real-world settings. One problematic consequence of this is that we still have little data on how such agents do in real-world settings and which factors influence acceptance and success in such scenarios. But, to make ECAs ultimately a useful and successful application, we need to make them capable of interacting with naïve, uninformed humans in everyday situations.

Originally started out as platform for studying the generation of natural multimodal behavior, we have extended the agent *Max* in following projects to a conversational assistant in Virtual Reality construction tasks [13] or to a virtual receptionist that welcomes people in the hallway of our lab [12]. In January 2004, we have brought Max to an application in the *Heinz Nixdorf MuseumsForum* (HNF), a public computer museum in Paderborn (Germany), thus venturing the step from a lab-inhabiting research prototype to a system being confronted daily with real humans in a real-world setting. In this setting (shown in Figure 1), Max is visualized in human-like size on a static screen, standing face-to-face to visitors of the museum. The agent is equipped with camera-based visual perception and can notice visitors that are passing by. Acting as a museum guide, Max's primary task is to engage visitors in conversations in which he provides them in comprehensible and interesting ways with information about the museum, the exhibition, or other topics of interest. Visitors can give natural language input to the system using a keyboard, whereas Max will respond with a synthetic German voice and appropriate nonverbal behaviors like manual gestures, facial expressions, gaze, or locomotion. In doing so, he should be as natural and believable as possible a communication partner, being entertaining and fun to talk with. He should not give talks in a teacher-like manner, but tailor his explanations to contextual factors like the visitor's interests and respond to questions, interruptions, or topic shifts. To create the impression of an enjoyable, cooperative interaction partner, the agent should also be capable of coherent small talk which helps reduce the social distance between the interlocutors [3].

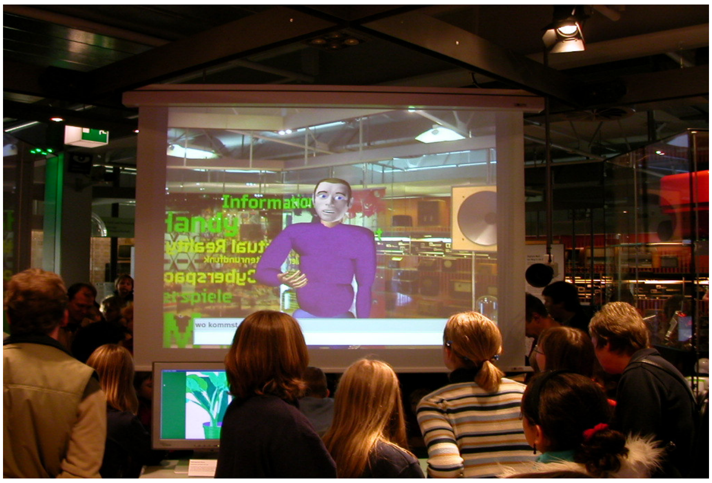

Fig. 1. Max interacting with visitors in the Heinz-Nixdorf-MuseumsForum.

After discussing related work in the next section, we start to describe the design of our system by explaining shortly the overall architectural layout in Sect. 3. In Sect. 4, we then focus on how Max's conversational capabilities are achieved. Finally, we have studied the communications that take place between Max and the visitors. Some anecdotal evidence on Max's capabilities to engage visitors in communicative episodes in the hallway setting was already reported in [12]. Now we were interested in the kind of dialogues that the museum visitors—unbiased people with various backgrounds, normally not used to interact with an ECA—are willing to have with Max and whether these bear some resemblance with human-human dialogues. We describe results of our first studies in the last section of this paper.

2 Related Work

Systems capable of spoken dialogue, either text-based or in natural language, have been around for quite a period of time and the approaches differ in many respects, from the modeling of linguistic structure and meaning to their efficiency, robustness, or coverage of domains. Already Weizenbaum's virtual psychotherapist Eliza [24], although not even trying to understand its 'patients', often managed to make them feel taken care of, thus demonstrating the effects achievable with rule-based, adeptly modeled small talk. During the last years, this genre of conversational agents revived as so-called chatterbots on the web, still making use of the 'Eliza-effect'. To name the most elaborated one, ALICE [23] utilizes a knowledge base containing 40.000 input-response rules concerning general categories, augmented with knowledge modules for special domains like Artificial Intelligence. This approach was also employed in other domains, e.g., to simulate co-present agents in a virtual gallery [7].

With enhancement of virtual agent technology and a growing awareness of the fact that a dialogue contribution is usually an ensemble of verbal and nonverbal behaviors, ECAs have become prominent. Some ECAs take a deep generation approach to generation, like the real estate agent REA [4] that was capable of understanding speech and gesture and of planning multimodal utterances from propositional representations of meaning. Keeping a model of interpersonal distance to the user, REA used small talk to reduce this distance if she noticed a lack of closeness to the client [3]. Systems like BEAT [5] or Greta [20] have addressed the generation of complex multimodal behavior from aspects like information structure, semantic-pragmatic aspects, or certainty and affect. Other ECAs have been designed based on more practical approaches aiming at robustness, efficiency or coverage of multiple, yet shallowly modeled domains. For example, MACK [6] could give directions to visitors of the MIT Media Lab based on a repository of user queries and system responses. August [9] was a talking head that has been used for six months as an information kiosk at the Stockholm Cultural Center. The system replied to spoken utterances by predefined answers in synthetic speech, facial expression, head movement, and thought balloons. Similar systems have been proposed as virtual museum guides, e.g. in [25]. The virtual H.C. Andersen system [2] uses spoken and gestural interaction to entertain children and educate them about life and work of HCA. Conversational skill is modeled by fairy tale templates and topic-centered mini-dialogues, while paying attention to the rhapsodic nature of non-task-oriented conversation and conversational coherence. These main tenets have been confirmed in user tests.

3 System Architecture

To comply with the requirements in the HNF setting, we have designed the overall architecture of the system as shown in Fig. 2. It resembles what has been proposed as reference architecture for ECAs [4], but is based on more cognitively motivated tenets [18]. As the agent should be able to conduct natural language interactions, constraints on linguistic content (in understanding as well as in producing utterances) should be as weak as possible. Thus, a keyboard was used as input device, avoiding problems that arise from speech recognition in noisy environments. Note also that this restricts Max to dialogues with only one visitor at a time. Nevertheless, camera-based perception provides the agent with constant visual information about the space in front of the keyboard as well as a greater view at the exhibition area. Real-time capable, standard image processing techniques are employed to scan the video data for skin-colored areas, find regions that probably correspond to faces, and track them over time. That way Max is able to detect the presence of multiple persons and to discriminate between them as long as no overlaps of face regions in the image occur. All speech and visual input are sent to a perception module that utilizes sensory buffers, ultra-short term memories, to compensate for recognition drop-outs and to integrate both kinds of data. It thus detects changes that take place in the scene and distributes them in the form of events, e.g., `person-13-entered` or `person-22-speaks`, to both reactive and deliberative processing.

Reactive processing is realized by the behavior generation component, which is generally in charge of realizing the behaviors that are requested by the other components. On the one hand, this includes feedback-driven reactive behaviors. For example, it hosts a behavior that, based on incoming positioning events, immediately triggers the agent's motor control to perform all eye and head movements needed to track the current interlocutor by gaze. Such reactive behaviors can be activated, deactivated or set to other stimulus objects at any time. Other behaviors concern the agent's secondary actions like eye blink and breathing. On the other hand, the behavior generation component must accomplish the realization of all utterances Max is to make. This includes the synthesis of prosodic speech and the animation of emotional facial expressions, lip-sync speech, and coverbal gestures, as well as scheduling and executing all verbal and nonverbal behaviors in synchrony. This task is realized using our *Articulated Communicator Engine*, a framework for building and animating multimodal virtual agents [14].

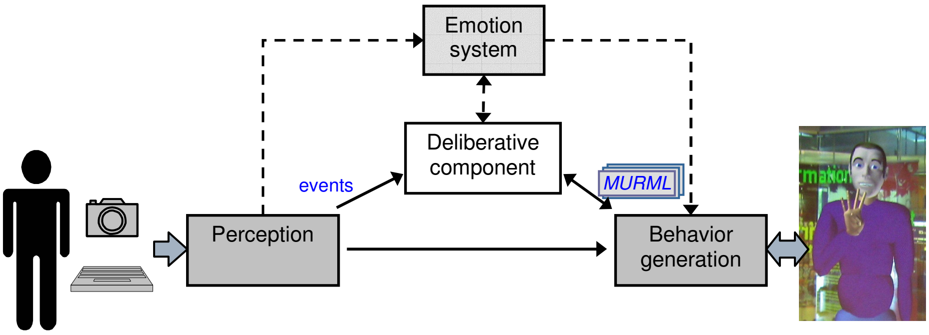

Fig. 2. Overview of the system architecture.

Deliberative processing of events takes place in a central deliberative component (the white box in Fig. 2). This component determines when and how the agent acts, either driven by internal goals and intentions or in response to incoming events, which, in turn, may originate either externally (user input, persons that have newly entered or left the agent's visual field) or internally (changing emotions, assertion of a new goal etc.). It maintains a dynamic spatial memory that contains all objects and persons in the agent's environmental context. This enables Max to directly refer to objects in its real-world surrounding, for example, to point at a robot placed next to the screen when mentioning it. How the deliberative component produces conversational behavior is described in Sect. 4.

Finally, Max is equipped with an emotion system that continuously runs a dynamic simulation to model the agent's emotional state. The emotional state is available anytime both in continuous terms of valence and arousal as well as a categorized emotion, e.g. happy, sad or angry, along with an intensity value (see [1]). The continuous values modulate subtle aspects of the agent's behaviors, namely, the pitch and speech rate of his voice and the rates of breathing and eye blink. The weighted emotion category is mapped to Max's facial expression and is sent to the agent's deliberative processes, thus making him cognitively "aware" of his own emotional state and subjecting it to his further deliberations. The emotion system, in turn, receives input from both the perception (e.g., seeing a person immediately causes positive stimulus) and the deliberative component. For example, obscene or politically incorrect wordings ("nowords") in the user input leads to negative impulses on Max's emotional system (see [1]). Since subsequent stimuli in the same direction accumulate in the emotion system, repeated insults will put the agent in an extremely bad mood, which in turn can eventually result in Max leaving the scene, an effect introduced to de-escalate rude visitor behavior.

4 Generating Conversational Behavior

The deliberative component carries out the three basic steps in creating conversational behavior: interpreting an incoming event, deciding how to react dependant on current context, and producing the appropriate response. Fig. 3 shows the flow of processing in this component, exposing separate processing stages for these steps and the knowledge structures they draw upon. On the one hand, the agent has static (long-term) knowledge that encompasses former dialogue episodes with visitors, informs his capabilities of dialogue management, and lays down his general competencies in interpreting natural language input and generating behaviors for a certain communicative function. On the other hand, there is evolving dynamic knowledge that provides the context in which interpretation, dialogue management, and behavior generation are carried out. A discourse model contains a history of the last utterances as well as up-to-date context information: The currently perceived persons and the active participant (interaction level); the holder of the turn, the goals the dialogue is pursuing and who brought them up, i.e. who has the initiative (discourse level); the current topic and contexts, the rhetorical structure, and the grounding status of information (content level). A user model contains all information that is gained throughout the dialogue. This includes information about the user (name, age, place of residence, etc.), his preferences and interests (determined by topics the user selected or rejected), and his previous behavior (cooperativeness, satisfaction, etc.). Lastly, a system model comprises the agent's world knowledge as well as current goals and intentions (for details see [8]). These structures enable Max's to act proactively in dialogue, e.g., to take over the initiative, rather than being purely responsive as classical chatterbots are.

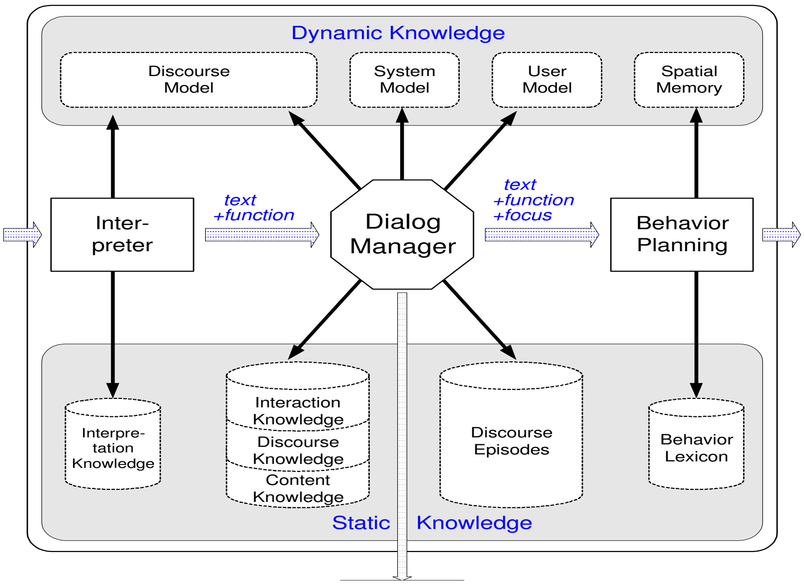

Fig. 3. An interior view at the functioning of Max's deliberative component.

All processes in the deliberative components are carried out by a BDI interpreter, which incessantly pursues multiple plans (*intentions*) to achieve goals (*desires*) in the context of up-to-date world knowledge (*beliefs*). We use an extended version of JAM [10]. Most of the plans implement condition-action rules, one of the underlying mechanisms with which Max's conversational knowledge is modeled. Such rules can test either the user input (text, semantic or pragmatic aspects) or the content of dynamic knowledge bases (beliefs, discourse or user model); their actions can alter the dynamic knowledge structures, raise internal goals and thus invoke corresponding plans, or trigger the generation of an utterance (stating words, semantic-pragmatic aspects, and markup of the central part). All rules are defined in an XML-based specification language that builds on and extends the AIML language from the ALICE system [23]. These XML descriptions were turned automatically into JAM plans (via XSL transformations) and added to the plan library of the BDI system.

4.1 Dialogue acts and conversational functions

In general, our approach to modeling conversational behavior assumes a *dialogue act* to be the basic unit of interaction, comparable to but more specific than Poggi & Pelachaud's [20] communicative act. Every user input as well as every response by the agent is considered to consist of one or more dialogue acts. Following Cassell et al. [4] and building on speech act theory, we consider each dialogue act as goal-directed action performed in context, and we distinguish between the overt behaviors and the functional aspects these behaviors fulfill. That is, every dialogue act fulfills a *communicative function.* It can thereby be effective on different levels of dialogue (cf. [22]), of which we distinguish the following three: the *interaction* level, the *discourse* level, and the *content* level. Actions at the interaction level can take place anytime and concern the establishment and maintenance of a conversation (greeting/farewell). Within an interaction, actions at the discourse level manage the topic and flow of conversation (e.g., the suggestion of a new topic to talk about). At the content level, information about the current topic is conveyed.

A communicative function, which explicates the functional aspects of a dialogue act with respect to these levels, consists of the following independent components:

- *performative*: the action that the dialogue act performs, reflecting part of the speaker's intention—to provide or require (askFor) information
- *reference level*: which level the act refers to—content, discourse, or course of interaction
- *content*: the information that further specifies the performative, e.g., which information the speaker asks for or which interactional signal she emits.

These aspects are collapsed into one single communicative function of the form `<performative>.<reference level>.<content> [arguments]`. Further details can be added either as a forth component or as the optional arguments (in brackets). The resulting function covers, from left to right, a spectrum from pragmatic to semantic aspects of the dialogue act. This allows for grouping the functions, e.g., `provide.content` comprises all propositional contributions regardless of the semantic information they convey. In our current system, about 200 communicative functions are distinguished, including for example

`provide.interaction.greeting`	(e.g. "Hi there!")
`askFor.content.name`	(e.g. "What's your name?")
`askFor.discourse.topic.sports`	(e.g. "Let's talk about sports.")

4.2 Interpretation

Starting out from textual user input, the first stage in deliberative processing is interpretation (see Fig. 3). Its task is to derive the intended communicative function and to pass it along with the original text on to the dialogue manager. A regular parser would constantly fail in the museum setting where mistyped or ungrammatical input is not unusual. We thus opted for simple but robust text analysis techniques that neglect most of syntactic well-formedness. Incoming input is interpreted by dedicated JAM plans in two steps. First, general semantic concepts are identified (negation, agreement, pos./neg. adjective, references) by simple pattern matching rules. To deal with negated expressions, different patterns are matched sequentially. For example, the utterance "I won't agree" contains a negation ("won't") and a signal for agreement ("agree"), therefore resulting in a disagreement. The second step determines the communicative function, again, using dedicated rules whose preconditions match actual words, the occurrence of semantic concepts, or entries of the discourse or user model. Ordering rules by decreasing generality, a general rule can be corrected by a more specialized one. When none of the rules matches, i.e. no function could be recognized, only the text is being passed on and Max can still revert to small-talk behavior using, e.g., commonplace phrases.

Currently, Max has 138 interpretation rules. To demonstrate how they work in detail, we give here two examples of rules—for sake of clarity in the original XML format—that interpret user input for its communicative function. The first example is a rule that checks in its condition part (`match`) for keywords, specified as regular expressions with an asterisk, and `asserts` in its action part a modifier `farewell`. A modifier constitutes an intermediate representation of communicative aspects, which are then further processed by subsequent rules. Note that this rule does not make any assumptions about the performative or reference level.

```
<rule name="interprete.type1.farewell">
  <match>
    <keywords>bye,cu,cya,exit,quit,ciao,ade,adios,hasta*,auf
     wieder*,tschoe,tschues*,tschau,und weg,so long,machs
     gut,bis bald,bis dann,bis spaeter,wiedersehen</keywords>
  </match>
  <assert>
    <convfunction modifier="farewell" filter="yes"/>
</assert> </rule>
```

The second example shows a rule that inspects semantic-pragmatic aspects of the current as well as the former dialogue act, notably, whether the utterance this input is in response to was a request for confirmation and whether the semantics of this input has been analysed by previous rules to be `undecided`. In this case, the rule will assert to Max's beliefs a communicative function meaning that the visitor has provided information indicating that he is undecided regarding the previous question of Max:

```
<rule name="interprete.type4.provide.content.indecision">
  <match>
    <allof>
      <convfunction ref="lastReply" type="askFor.content.
                                         confirmation"/>
      <convfunction modifier="undecided"/>
    </allof>
  </match>
  <assert>
    <convfunction type="provide.content.indecision"/>
</assert> </rule>
```

4.3 Dialogue management

The tasks of the dialogue manager amount to updating the dynamic knowledge bases, controlling reactive behaviors, and—most importantly—creating appropriate utterances. While a simple rule-based approach seems appropriate to model robust small talk, the agent must also be able to conduct longer, coherent dialogues, calling for a more plan-based approach and a profound modeling of the respective domains. We have combined these two approaches employing the BDI interpreter that affords both kinds of processing. A skeleton of JAM plans realize the agent's general, domain-independent dialogue skills like negotiating initiative or structuring a presentation. These plans are adjoined by a larger number of small JAM plans that implement condition-action rules like the ones shown above. These rules define the agent's domain-dependant conversational and presentational knowledge, e.g., the dialogue goals that can be pursued, the possible presentation contents, or the interpretation of input. As currently set up in the museum, Max is equipped with 876 skeleton plans and roughly 1.200 rule plans of conversational and presentational knowledge. At run-time, the BDI interpreter scores all plans dependant on their utility and applicability in context. The most adequate plan is then selected for execution.

Max's conversational behavior is laid down through this collection of JAM plans, which can be differentiated according to the level of dialogue they act upon. The plans at the *interaction level* state how Max can start/end a dialogue and how to react to various input events (e.g., when the user starts or finishes typing). If there is no ongoing conversation, newly perceived persons are greeted and encouraged to start an interaction. If an interaction is concluded, the gained knowledge (models of the discourse and its participants) is compressed into a dialogue episode and stored in long term memory. In future discourses the system draws upon these episodes to answer questions like "How many people were here today?" or to derive user-related answers: If a user states a favorite movie, the system looks up whether it has been stated before, possibly resulting in the response "Seems to be a rather popular movie".

The *discourse layer* deals with mechanisms of turn-taking, topic shift, and initiative. The user can take the turn by starting to type, causing Max to stop speaking as soon as possible and to yield the turn. As the system performs a mixed-initiative dialogue, it needs to know about the user's wish to take the initiative, how to deal with conflicts, and how to establish initiative at all. Initiative is modeled as the raising of obligatory dialogue goals. The system is aware of these goals (discourse model) and disposes of plans for initiating, holding, resuming and releasing them. Dedicated rules analyse the input communicative function, e.g., to determine if the user wants to seize control over discourse and what goal she wants to pursue.

From the point of view of the system, initiative is the key for maximizing the coherence of the dialogue. If Max has the goal of coming to know the interlocutor's name, he will try to seize control over dialogue and to ask for the name. If the user refuses to answer but brings up another topic to talk about, Max will accept this "intermezzo", giving away the initiative temporarily but will re-seize it and return to his goal at the earliest time possible. This is one instance where a rule-based, merely responsive approach to dialogue would break down. Max can handle these cases by utilizing longer-term plans and the notion of desires to influence plan execution in the BDI framework: the agent's own desire for initiative increases when it is available and neither of the participants is about to take it. He then seizes control when a threshold is reached. Instead of being only reactive to user input, Max is thus able to keep up the conversation himself and to conduct a coherent dialogue.

The *content layer* comprises general conversational knowledge that comprises a dictionary of given names, a lexicon of “no-words” according to the museum’s policies, and 608 rules that state how to react to keywords or keyphrases in a given context, forming Max’s small talk capabilities. This also encompasses rules for a guessing animal game where Max asks questions to find out an animal that a visitor has in mind based on discriminating features. In addition, the content layer contains plans that form Max’s presentation knowledge. This knowledge is chunked into units that are categorized (e.g., ‘technically detailed’, ‘anecdotal’) and organized according to the rhetorical relations between one another (e.g., ‘elaborates’, ‘explains’). Three top-level units (introduction, overview, summary) form the skeleton of a presentation. The remaining units form a tree with the overview unit on top. After giving an overview, the ongoing presentation can be influenced by the user as well as the system. Upon finishing a unit, Max offers the user possible units to elaborate. Explained units are noted in the discourse model. If the user is reluctant to select a unit or a certain unit might be of interest to the user, Max may also proceed with this unit himself. Such evidence comes from the user model and is gained either explicitly in previous dialogue or is inferred when the user rejects or interrupts the presentation of a unit of a certain type. In general, Max knows all possible dialogue goals of a certain domain, their preconditions, and the dialogue acts to open, maintain and drop them. When taking the initiative, Max can thus select one of these goals and initiate a presentation, small talk, or a guessing game himself.

4.5 Behavior planning

Behavior planning receives the words, the communicative function of the dialogue act, and the focus of the utterance to produce, and it is always informed about the current emotional state. It adds to the utterance nonverbal behaviors that support the given communicative function. Behaviors are drawn from a lexicon containing XML-based specifications in MURML [14]. At the moment, 54 different behaviors are modeled.

The mapping of communicative functions onto nonverbal behaviors is not easy, nor clearly definable for all cases. One reason for this that behaviors like hand gestures or facial expressions may serve fundamentally different semiotic functions. Additionally, there is barely a one-to-one mapping as multiple behaviors can often realize one function, just as one behavior can fulfill several functions [4]. To account for most of the flexibility and complexity of this mapping, the indexing of nonverbal behaviors in our lexicon can address single parts of the hierarchical structure of a communicative function. For examples, defined mappings are

`provide.interaction.greeting`	→ hand wave
`provide.discourse.agreement`	→ head nod
`provide.content.ironical`	→ eye blink
`provide.content`	→ raise hand
`*.content.number-two`	→ handshape two fingers stretched

The functions’ hierarchical structure allows to suitably represent the functions of more general behaviors, like the quite generic, yet frequent metaphorical gesture of simply raising a hand in front of the body (example four). Omitting the content part of the function (`provide.content`), our mapping assumes that this gesture signals that some content is being brought up, independent of the content itself. That is, while this gesture focuses on pragmatic aspects, it can be chosen to accompany words and other nonverbal behaviors that probably inform better about the content itself. On the other hand, a nonverbal behavior can serve a semiotic function of conveying a certain meaning, regardless of pragmatic aspects like whether this meaning is part of a request or an inform type dialogue act. Using an asterisk symbol as shown in the last example, the symbolic gesture for the number of two, single aspects of the function can be left open for such behaviors. In result, Max can choose this gesture whenever he needs to refer to this meaning, in statements as well as in questions.

When augmenting a dialogue act with nonverbal behaviors, the generation component picks behaviors whose functions cover most of the semantic-pragmatic aspects of the dialogue act (trying to increase informativeness). Yet, there will often be too large a number of possible behaviors. As in other systems [5], this conflict is resolved partly based on information about the scope of each behavior (the occupied modality) and partly by random choice. Behavior planning also allocates the bodily resources and can thus take account of the current movement and body context. For example, a greeting gesture that can potentially be made with either hand is performed with, say, the left hand if this hand has been mobilized before and has not returned to its rest position yet. Drawing upon the spatial memory, behavior planning also refines deictic gestures by translating symbolic references like `camera` into world coordinates.

6 Evaluation of Max´s communicative effects

We wanted to see (1) if Max's conversational capabilities suffice to have coherent, fluent interactions with the visitors to the museum, and (2) whether the dialogues bear some resemblance with human-human dialogues, i.e. if Max is perceived and treated as human-like communication partner. Recent findings demonstrate remarkable effects of an agent on the user's (social) behavior: An embodied agent may lead people to show increased impression management and socially desirable behaviors [21,15]; may influence the user's mood [16] or affect the user's task performance (social facilitation/inhibition [21,17]). Also, agents have proven to affect the communication of the human user: When interacting with ECAs, people are more inclined to use natural language than when interacting with text- or audio-based systems [17,15], children accommodate their speech to that of the virtual character [19], and people engage in small talk with a virtual character and take its social role into account [11]. Yet, none of these studies has been conducted in a real-world setting.

Study 1

A first screening was done after the first seven weeks of Max's employment in the Nixdorf Museum (15 January through 6 April, 2004). Statistics is based on logfiles, which were recorded from dialogues between Max and visitors to the museum. During this period, Max on average had 29.7 conversations daily (SD=14), where "conversation" was defined to be the discourse between an individual visitor saying hello and good bye to Max. Altogether there were 2259 conversations, i.e. logfiles screened. On the average, there were 22.60 (SD=37.8) visitor inputs recorded per conversation, totalling to 50,423 inputs recorded in the observation period. The high standard deviation (SD) reveals a great variation in the length of the dialogues, with extremely short interactions as well as long ones of more than 40 visitor inputs. The data were further evaluated with respect to the successful recognition of communicative functions, that is, whether Max could associate a visitor's want with an input. A rough screening among these further pertained to whether visitors would approach Max politely or whether they would employ insulting, obscene, or "politically incorrect" wordings. Finally, we looked at how often visitors would play the guessing game with Max.

We found that Max was able to recognize a communicative function in 32,332 (i.e. 63%) cases. Note that this is the absolute number of classifications, including possibly incorrect ones. We can thus only conclude that in at most two-thirds of all cases Max conducted sensible dialogue with visitors. In the other one-third, however, Max did not turn speechless but simulated small talk behavior by employing commonplace phrases. Among those cases where a communicative function was recognized, with overlaps possible, a total of 993 (1.9%) inputs were classified by Max as polite ("please", "thanks"), 806 (1.6%) as insulting, and 711 (1.4%) as obscene or politically incorrect, with 1430 (2.8%) no-words altogether. In 181 instances (about 3 times a day), accumulated negative emotions resulted in Max leaving the scene "very annoyed". The guessing animal game was played in 315 instances, whereby 148 visitors played the game once, 34 twice, and 26 three or more times. A qualitative conclusion from these findings is that Max apparently "ties in" visitors of the museum with diverse kinds of social interaction. Thus we conducted a second study to investigate in what ways and to what extent Max is able to engage visitors in social interactions.

Study 2

We analysed the content of user utterances to find out whether people use human-like communication strategies (greetings, farewells, commonplace phrases) when interacting with Max. Specifically, we wanted to know if they use utterances that indicate the attribution of sociality to the agent, e.g., by asking questions that only make sense when directed to a human. We analysed logfiles of one week in March 2005 (15th through 22nd) containing 205 dialogues. The number of utterances, words, words per utterance, and specific words such as “I/me” or “you” were counted and compared for agent and user. The content of user utterances was evaluated by means of psychological content analysis and following criteria of qualitative empirical approaches: using one third of the logfiles, a scheme was developed that comprised categories and corresponding values as shown in Table 1. Two coders coded the complete material and counted the frequencies of categories and values, with multiple selections possible. We chose this method since a solid theoretical foundation and a thorough understanding of the kinds of social interactions one could expect to take place between Max and the visitors is currently lacking. We thus developed the categories data-driven instead of deduced from theory. In order to achieve a maximum of inter-coder reliability, the coders jointly coded parts of the material and discussed unclear choices.

The quantitative analysis showed that the agent is more active than the user is. While the user makes 3665 utterances during the 205 dialogues (on average 17.88 utterances per conversation), the agent has 5195 turns (25.22 utterances per conversation). Not only does the agent use more words in total (42802 in all dialogues vs. 9775 of the user; 207.78 in average per conversation vs. 47.68 for the user), but he also uses more words per utterance (7.84 vs. 2.52 of the user). Thus, the agent in average seemed to produce more elaborate sentences than the user does, which may be a consequence of the use of a keyboard as input device. Against this background, it is also plausible that the users utters less pronouns such as “I/me” (user: 0.15 per utterance; agent: 0.43 per utterance) and “you” (user: 0.26 per utterance; agent: 0.56 per utterance). These results might be due to the particular dialogue structure that is, for some part, determined by the agent’s questions and proposals (e.g., the guessing game leaves the user stating “yes” or “no”). On the other hand, the content analysis revealed that 1316 (35.9 %) of the user utterances are proactive (see Table 1). Concerning human-like strategies of beginning/ending conversations, it turned out that especially greeting is popular when confronted with Max (used in 57.6% of dialogues). This may be triggered by the greeting of the agent. But, given that the user can end the conversation by simply stepping away from the system, it is remarkable that at least 29.8% of the people said goodbye to Max. This tendency to use human-like communicative structures is supported by the fact that commonplace phrases—small talk questions like "How are you?"—were uttered 154 times (4.2% of utterances).

Table 1. Contents of user utterances and their frequencies.

Category & corresponding values	Examples (translated to English)	*N*
Proactivity		
Proactive utterance		1316 (36%)
Reactive utterance		1259 (34%)
Greeting		
Informal greeting	Hi, hello	114
Formal greeting	Good morning!	4
No greeting		87
Farewell		
Informal farewell	Bye	56
Formal farewell	Farewell	5
No farewell		144
Flaming		**406** (11%)
Abuse, name-calling	Son of a bitch	198
Pornographic utterances	Do you like to ****?	19
Random keystrokes		114
Senseless utterances	http.http, dupa	75
Feedback to agent		**83** (2%)
Positive feedback	I like you; You are cool	51
Negative feedback	I hate you; Your topics are boring	32
Questions		**746** (20%)
Anthropomorphic questions	Can you dance? Are you in love?	132
Questions concerning the system	Who has built you?	109
Questions concerning the museum	Where are the restrooms?	17
Commonplace phrases	How are you?	154
Questions to test the system	How's the weather?	146
Checking comprehension	Pardon?	139
Other questions		49
Answers		**1096 (**30%)
Inconspicuous answer		831
Apparently wrong answers	[name] Michael Jackson, [age] 125	61
Refusal to answer	I do not talk about private matters	8
Proactive utterances about oneself	I have to go now	76
Answers in foreign language		30
Utterances to test the system	You are Michael Jackson	66
Laughter		24
Request to do something		**108** (3%)
General request to say something	Talk to me!	10
Specific request to say something	Tell me about the museum!	13
Request to stop talking	Shut up!	24
Request for action	Go away! Come back!	61

As with all publicly available agents or chatterbots, we observed flaming (406 utterances; 11.1%) and implicit testing of intelligence and interactivity (303; 8.3%). The latter happens via questions (146; 4%), obviously wrong answers (61; 1.7%), answers in foreign languages (30; 0.82%), or utterances to test the system (66; 1.8%). However, direct user feedback to the agent is more frequently positive (51) than negative (32). Most elucidating with regard to whether interacting with Max has social aspects are the questions addressed to him: There were mere comprehension questions (139; 18.6% of questions), questions to test the system (146; 19.6%), questions about the system (109; 14.6%), the museum (17; 2.3%), or something else (49; 6.6%). The vast amount of questions are social, either since they are borrowed from human small talk habits (commonplace phrases; 154; 20.6%) or because they directly concern social or human-like concepts (132; 17.7%). Thus, more than one-third of the questions presuppose that treating Max like a human is appropriate—or try to test this very assumption. Likewise, the answers of the visitors (30% of all utterances) show that people seem to be willing to get involved in dialogue with the agent: 75.8% of them were expedient and inconspicuous, whereas only a small number gave obviously false information or aimed at testing the system. Thus, users seem to engage in interacting with Max and try to be cooperative in answering his questions.

7 Conclusion

Current ECAs have for the most part stayed within the boundaries of their lab environments and there is only little data on whether conversational virtual agents can be successfully employed in real-world applications. We have developed our agent Max to apply him as a guide to the HNF computer museum, where he has been interacting with visitors and providing them with information daily since January 2004 (more than one and a half years by now). To comply with the requirements for human-like, yet robust conversational behavior, our design adopts the rule-based approach to dialogue modeling but extends it in several ways. It takes account of the semantic-pragmatic and context-dependent aspects of dialogue acts, it combines rule application with longer-term, plan-based behavior, and it drives the generation of not just text output but fully multimodal behavior.

The field studies that we have conducted to see if Max, based on this design, is accepted by the visitors as a conversation partner and if he succeeds in engaging them in social interactions yielded promising evidence. Judging from the logfiles, people are likely to use human-like communication strategies (greeting, farewell, small talk elements, insults), are cooperative in answering his questions, and try to fasten down the degree of Max's human-likeness and intelligence. This indicates the attribution of sociality to the agent. Our studies also provide clues to how the design should be enhanced. For example, we realized from many anthropomorphic questions that Max should be capable of flirting behavior as he is tested in this respect quite frequently. The studies will also serve as pre-test for a more experimentally controlled study on Max's social effects and subtle user reactions, which would also include analyses of video data.

Acknowledgement

This work was partially supported by the *Heinz Nixdorf MuseumsForum* (Paderborn).

References

1. C. Becker, S. Kopp, I. Wachsmuth: Simulating the Emotion Dynamics of a Multimodal Conversational Agent. Affective Dialogue Systems (2004)

2. N.O. Bernsen, L. Dybkjær: Domain-Oriented Conversation with H.C. Andersen. Affective Dialogue Systems (2004)
3. T. Bickmore, J. Cassell: 'How about this weather?' Social Dialog with Embodied Conversational Agents. Proc. of AAAI Symposium on Socially Intelligent Agents (2000)
4. J. Cassell, T. Bickmore, L. Campbell, H. Vilhjalmsson, H. Yan: Human Conversation as a System Framework: Designing Embodied Conversational Agents. In: Cassell et al. (eds.) Embodied Conversational Agents, MIT Press (2000)
5. J. Cassell, H. Vilhjalmsson, T. Bickmore: BEAT: The Behavior Expression Animation Toolkit. Proc. of SIGGRAPH 2001, Los Angeles, CA (2001)
6. J. Cassell, T. Stocky, T. Bickmore, Y. Gao, Y. Nakano, K. Ryokai, D. Tversky, C. Vaucelle, H. Vilhjalmsson: MACK: Media lab Autonomous Conversational Kiosk. Proc. of Imagina '02, Monte Carlo (2002)
7. M. Gerhard, D.J. Moore, D.J. Hobbs: Embodiment and copresence in collaborative interfaces. Int. J. Hum.-Comput. Stud. 61(4): 453-480 (2004)
8. L. Gesellensetter: Planbasiertes Dialogsystem für einen multimodalen Agenten mit Präsentationsfähigkeit. (Plan-based dialog system for a multimodal presentation agent) Masters Thesis, University of Bielefeld (2004)
9. J. Gustafson, N. Lindberg, M. Lundeberg: The August Spoken Dialogue System. Proc. of Eurospeech '99, Budapest, Hungary (1999)
10. M.J. Huber : JAM : A BDI-Theoretic Mobile Agent Architecture. Proc. Autonomous Agents'99, Seattle (1999)
11. K. Isbister, B. Hayes-Roth: Social Implications of Using Synthetic Characters. IJCAI-97 Workshop on Animated Interface Agents: Making them Intelligent, Nagoya (1998), 19-20
12. B. Jung, S. Kopp: FlurMax: An Interactive Virtual Agent for Entertaining Visitors in a Hallway. In T. Rist et al. (eds.): Intelligent Virtual Agents, Springer (2003), 23-26
13. S. Kopp, B. Jung, N. Lessmann, I. Wachsmuth: Max-A Multimodal Assistant in Virtual Reality Construction. KI-Künstliche Intelligenz 4/03: 11-17 (2003)
14. S. Kopp, I. Wachsmuth: Synthesizing Multimodal Utterances for Conversational Agents. Computer Animation and Virtual Worlds 15(1): 39-52 (2004)
15. N.C. Krämer, G. Bente, J. Piesk: The ghost in the machine. The influence of Embodied Conversational Agents on user expectations and user behaviour in a TV/VCR application. In: G. Bieber & T. Kirste (eds): IMC Workshop 2003, Assistance, Mobility, Applications. Rostock (2003) 121-128
16. N.C. Krämer, B. Tietz, G. Bente: Effects of embodied interface agents and their gestural activity. In: T. Rist et al. (eds.): Intelligent Virtual Agents. Springer (2003) 292-300
17. N.C. Krämer, J. Nitschke: Ausgabemodalitäten im Vergleich: Verändern sie das Eingabeverhalten der Benutzer? (Output modalities compared: Do they change the input behavior of users?) In: R. Marzi et al. (eds.): Bedienen & Verstehen. 4. Berliner Werkstatt Mensch-Maschine-Systeme. VDI-Verlag, Düsseldorf (2002) 231-248
18. N. Leßmann, I. Wachsmuth: A Cognitively Motivated Architecture for an Anthropomorphic Artificial Communicator. Proc. of ICCM-5, Bamberg (2003)
19. S. Oviatt, C. Darves, R. Coulston: Toward adaptive Conversational interfaces: Modeling speech convergence with animated personas. ACM Trans. on CHI, 3 (2004) 300-328
20. C. Pelachaud, I. Poggi: Multimodal Communication between synthetic Agents. Proc. of Advanced Visual Interfaces, L'Aquila, Italy (1998)
21. R. Rickenberg, B. Reeves: The effects of animated characters on anxiety, task performance, and evaluations of user interfaces. Letters of CHI 2000 (2000), 49-56
22. D.R. Traum, J. Rickel: Embodied Agents for Multi-party Dialogue in Immersive Virtual Worlds. Proc. of AAMAS'02 (2002)
23. R.S. Wallace: The Anatomy of A.L.I.C.E. Tech.report, ALICE AI Foundation (2000)
24. J. Weizenbaum: ELIZA: a computer program for the study of natural language communication between men and machines. Communications of the ACM, vol.9 (1996)
25. X. Yuan, Y.S. Chee: Embodied Tour Guide in an Interactive Virtual Art Gallery. International Conference on Cyberworlds (2003)