International Journal of Social Robotics
The final publication is available at Springer via http://dx.doi.org/10.1007/s12369-015-0288-9

Capturing expertise - Developing interaction content for a robot through teleoperation by domain experts

Kanae Wada · Dylan F. Glas · Masahiro Shiomi · Takayuki Kanda · Hiroshi Ishiguro · Norihiro Hagita

Received: 30 Oct 2012 / Accepted: 16 Feb 2015

Abstract The development of humanlike service robots which interact socially raises a new question: How can we create good interaction content for such robots? Domain experts specializing in the target service have the knowledge for making such content. Yet, while they can easily engage in good face-to-face interactions, we found it difficult for them to prepare conversational content for a robot in written form. Instead, we propose involving experts as teleoperators in a short-cycle iterative development process in which the expert develops content, teleoperates a robot using that content, and then revises the content based on that interaction. We propose a software system and design guidelines to enable such an iterative design process. To validate these solutions, we conducted a comparison experiment in the field, with a teleoperated robot acting as a guide at a tourist information center in Nara, Japan. The results showed that our system and guidelines enabled domain experts with no robotics background to create better interaction content and conduct better interactions than domain experts without our system.

Keywords Communication robots · Techniques · Field experiments

K. Wada · D. F. Glas · M. Shiomi · T. Kanda · N. Hagita
ATR Intelligent Robotics and Communication Laboratories
2-2-2 Hikaridai. Keihanna Science City, Kyoto, Japan
E-mail: dylan@atr.jp
H. Ishiguro
Faculty of Science and Engineering, Osaka University,
1-3, Machikaneyama, Toyonaka City, Osaka, Japan

Fig. 1 Teleoperation of a robot interacting with a visitor at a tourist information center

1 Introduction

1.1 Communicative Knowledge

Recently, robotics researchers have been investigating the use of humanoid robots to provide services for and interact with people in everyday social environments, such as assisting in nursing homes [29], providing route guidance in a shopping center [33], helping people to do their shopping in a supermarket [12], greeting people at a reception booth [11], and interacting with people on city streets [42]. As such robots become common in society, it will be important for them to provide information to people through humanlike interaction (Figure 1).

However, the question arises: how can we develop good content (utterances and gestures) for such robots to provide information in a conversational style? It can

be difficult for a programmer or researcher to create such content if they lack expertise in the target service domain. While people such as drama majors [11] could be considered experts in social interaction in general, they are not actually experts in a target application domain for the robot. It would be difficult to create content for a teacher, doctor, or salesperson robot, for example, without having the specific skills and knowledge for that job, even if one had access to factual information related to that domain.

This is because, aside from the **factual knowledge** required to perform a service effectively (e.g. knowing the route to the nearest convenience store), there is also **communicative knowledge** which is equally important. For example, a teacher needs to know not only curriculum contents (factual knowledge) but also how to capture the attention of and motivate students (communicative knowledge). A doctor needs to understand the mechanisms of disease (factual) but also know how to talk to patients in a compassionate way (communicative). While factual knowledge is often documented and relatively easy for a robot developer to acquire, communicative knowledge may be undocumented and can only be provided by a domain expert.

This concept of communicative knowledge is a special case, focusing on skills used in communicative interactions, of what is more generally referred to as "tacit knowledge" [30]. An analysis of concepts related to this phenomenon, such as implicit learning, and a good summary of related work are presented in [7]. In the field of Human-Robot Interaction, the concept of "design patterns for sociality" [15] encapsulates some common examples of everyday communicative knowledge that can be used in the design of social behaviors. In this study, we are concerned not with these general patterns, but with patterns specific to a given field of expertise, which might not be obvious to a non-expert attempting to design an interaction.

1.2 Teleoperation and Content Development

So, can service experts make good interaction content for humanoid robots? According to our experience, no. In preliminary studies, we discovered that it is not intuitive for domain experts to sit at computers and create textual content for a robot to provide a service in their own natural conversational style.

This is because communicative knowledge is often implicit and difficult to codify into explicit rules to store for future use. Thus, we believe that content incorporating this knowledge can be most easily created through an iterative process of content generation and conversational interaction through teleoperation, where the expert can use their communicative knowledge in an intuitive way.

We consider this teleoperation phase to be quite important in the content development process. As previous studies have revealed, situation coverage, representing the amount of knowledge stored in the robot, can be increased over time through teleoperation [18]. Content prepared in advance is typically premature, and through observing real people's reactions, an operator can identify content that is missing or needs to be changed. Hence, we believe domain experts can gain useful feedback through the process of teleoperation. We aim to integrate this step of real interaction through teleoperation into an iterative process for the development and refinement of interaction content.

1.3 Applications and objectives

The generation of behavior content for conversational interactions has several potential applications. The first and most immediate goal of our study is to create a system that enables tele-work by expert users. The system should be fast and easy to use, e.g. by minimizing the need for typing, so it is important for the majority of spoken content and associated gestures to be entered ahead of time, rather than during operation.

In this study, we consider the case where the content entered by one operator is used by that same operator. In the future, this scenario could be extended to include collaborative development of content by a group of operators, or even development of content by expert operators to be used by non-experts.

In addition to pure teleoperation, large stores of interaction content can provide a basis for automating parts of an interaction, enabling the development of semi-autonomous systems. Such systems could still be controlled by expert operators, but with a lower workload. A possible application of this type of system would be multirobot teleoperation, like that shown in our previous studies [8].

Finally, in the more distant future, the collection of enough interaction data may make it possible to train fully-autonomous interactive social robots to perform the tasks of domain experts, in which case the operators would take on the role of trainers for the service robots.

1.4 Scenario - Sightseeing Guide Robot

In this study we consider the scenario of a robot providing information for tourists. The interaction content for such a robot would need to include a large amount

This is a preprint manuscript.
The final publication is available at Springer via http://dx.doi.org/10.1007/s12369-015-0288-9

of factual knowledge about the tourist attraction, its history, etc., as well as communicative knowledge like how to capture and sustain the interest of the tourists and how to tell stories in an engaging and exciting way. The ideal domain experts who could provide this knowledge would be people currently working as guides at the target location in question. In Japan, there are many guide associations where senior citizens volunteer their time to work as guides and provide information to tourists about sightseeing attractions.

In this study, we worked with one such volunteer guide association in the city of Nara, Japan. The members of this group are all retired senior citizens, with an average age of 68.4. As walking around and providing information is physically demanding, the guides need to take time to rest and cannot work every day. Consequently they are often understaffed during busy seasons, and they were quite receptive to the idea of using robots to help reduce their workload.

Working with senior citizens provided some challenges, as most participants were not frequent computer users or fast typists. We did make accommodations for this, such as using large font and button sizes in our software interface. However, the focus of this study was not about their limitations, but rather on accessing their rich factual and communicative knowledge developed over years of experience as guides.

2 Related Work

2.1 Dialog Construction

In the studies of dialog, there have been a couple of dialog models developed. Typically, a dialog model assumes "tasks" in a dialog. That is, it assumes typical flow and/or set of information to be exchanged. For example, in case of ticket selling, a dialog system would expect to receive information about a customer's request on departure time, destination, and number of passengers. For such task-oriented dialogs, a state-transition model or information-frame model fits well [24]. There are also authoring tools [9] [23], frameworks [1], and description languages [20] [27] that support the preparation of such task-based dialogs.

However, we aim to realize a chat-like conversation, where it is hard to anticipate a typical flow or set of information. There are agent-based models that can handle flexible dialogs, but it is difficult for people who are not experts in dialog systems to construct such dialog models [24]. Overall, these dialog studies did not reveal a way to convert knowledge from domain experts (who are not experts in dialog systems) into data useful for a robot's conversation.

Toward the problem, an alternative approach would be the modeling of novice people's dialog. Chernova and her colleagues developed an on-line game to collect people's dialog, and converted the collected dialog data into behaviors for a robot [5]. While such approaches based on large datasets hold great promise, it can be difficult to collect such data in advance, particularly if a dialog requires specific domain expertise that non-experts would not have.

In other work, the behavior of domain experts has sometimes been analyzed, e.g. to develop dialogue management systems based on a belief-desire-intention framework [39], or for training knowledge-based systems to mimic an expert's decision processes [2].

2.2 Iterative Content Development

Iterative approaches to content development for robots have often been explored in the context of human-robot interaction. The work by Kuo et al. illustrates a typical example of iterative design for service robots [21], and Lohse et al. explored the conceptual requirements for an architecture to support iterative user-driven design [22].

The idea of short-cycle iterative design has parallels in robotic learning-from-demonstration scenarios as well, e.g. [31, 44], although these studies focused on tasks other than conversational human-robot interaction.

2.3 Telepresence and Partial Autonomy

Previous studies have revealed a number of ways to provide service from distant locations. Telephone and video conferencing are already in widespread use, and recently telepresence robots have also come into use [28] [37] [43]. Studies have begun to investigate support techniques for telepresence robots [35]. In all of these telepresence approaches, it is a human user/operator who engages in a single channel of dialog.

In contrast, our approach uses partial autonomy, a flexible approach which can eventually enable multiple conversations to be supervised by a single operator [8]. In this approach, the ultimate goal is to let the system handle the majority of the dialog, with operators designated to support the system only when a situation is not covered by autonomy. While many situations can be automated for simple information providing dialogs [45] and greeting and information-providing services [17], previous studies have not shown how to prepare and update dialog contents, or how to involve domain experts in the loop.

This is a preprint manuscript.
The final publication is available at Springer via http://dx.doi.org/10.1007/s12369-015-0288-9

Although full or partial teleoperation may be appropriate for many applications, the content created through this procedure could conceivably be used in autonomous systems as well, using learning-based approaches for action selection [3].

2.4 Guidelines for Dialogue Design

There are several research works which have focused on dialogue design for conversational agents. For example, Hollingsed et al. have investigated the effectiveness of the short-term response behaviors by using a tutorial system [36]. Moreover, Ward et al. have reported some usability issues in spoken dialog applications such as responsiveness, feedback and so on [41]. Jung et al. considered a set of guidelines for developing behaviors for social robots, although their focus was on motion, rather than dialog [14]. These research works tried to identify important rules for dialog systems, but they did not focus on how to create a situation where domain experts can effectively create dialogue on their own, using their own intuitive rules.
By contrast, this paper aims to enable domain experts themselves to create content for a conversational robot.

3 Interaction guidelines

In this study, we propose an iterative procedure, alternating between the creation of conversational content and teleoperation using that content. To explore the effectiveness of this technique, we conducted several preliminary trials in which we brought a small focus group of members of the guide organization into our laboratory periodically over a period of several weeks. We asked them to create and test interaction content using early prototypes of our system. We watched them and talked with them to try to understand which aspects of the procedure they found difficult, and we brainstormed new features for the interface which could assist them in creating better content and smoother interactions.
We observed that the iteration process often did not provide feedback as we had hoped; operators could not create good dialog content, and the interactions they conducted were unnatural, slow, and awkward. We think this means that they were not able to create good interactions and content in a natural way. When the conversational content and/or operational technique were below some basic threshold of quality, it was not possible to conduct a smooth interaction, and so the operator did not get useful feedback to improve the content.
To help operators avoid these problems, we categorized a number of common mistakes that we observed. Based on these observations, we compiled a set of guidelines to assist future operators in producing good interactions. These guidelines can be classified into three main categories as shown in Table 1: Responsiveness, Initiative, and Interactivity. Responsiveness is important at all times, whereas Initiative and Interactivity are complementary, and they must be balanced carefully against each other.
As a note of clarification, in this paper, we use the term "behavior" to refer to some combination of utterance and/or gesture. The primary focus of this study was on spoken dialog content, so we are chiefly concerned with the utterance aspect of behaviors. In some places we will use the term "utterance" and "behavior" somewhat interchangeably, but where the system implementation is concerned we will use the term "behavior," as our implemented system does support gestures as well as speech.

3.1 Responsiveness

The first problem we observed was lack of responsiveness in the robot's interactions. The problem of timing and responsiveness, e.g. in turn-taking has been studied both in psychology [32] and in robotics [4], and it has been found that proper response timing is correlated with higher social skills [26].
One example of poor responsiveness is when the robot responded to the visitor slowly, after a long silence. Sometimes this happened when the operator was taking time to search for a proper utterance from a list and didn't seem to feel any time pressure. Other times, the operator didn't find an appropriate behavior in the system and instead took a very long time to type a new utterance.
Another problem with responsiveness is when the robot responded promptly, but appeared to ignore what the visitor was saying. For example, when the robot asked one visitor, "Where are you from?", and he answered, "I'm from Hokkaido!" the next utterance from a robot was "I'll explain about Nara." The visitor felt that the robot was not listening to what he said, or didn't care. Guidelines for responsiveness emphasize the importance of listening to the visitor and responding quickly and appropriately.

Reaction time Studies have shown that the length of natural pauses in human dialogue ranges from 0.62 to 0.77 seconds [13], and that delays longer than 2 seconds during human-robot interaction make people feel frustrated [34]. In teleoperation, an operator requires time to search through content or type utterances, so it can

This is a preprint manuscript.
The final publication is available at Springer via http://dx.doi.org/10.1007/s12369-015-0288-9

be difficult to respond so quickly [10].
Our guidelines recommend that the operator react quickly to the visitor. This includes being aware of the passage of time and minimizing typing whenever possible. Many features of our system (see Sec. 4) were developed to support this guideline.

Topic-independent utterances Many content-rich utterances can be classified under topics, such as "history of Nara Park." However, in natural conversation, people often use phrases such as, "oh, really?", "that's right", or "thank you," which do not fit into a specific topic. We use the term "topic-independent utterances" to refer to short utterances for making responses which cannot be classified as greetings, farewells, or topic-specific informational content. These include backchannel utterances such as "uh-huh," "okay," "yeah, I see," and so on.
Numerous studies have been conducted regarding backchannel utterances, e.g. [40]. These utterances are usually necessary for a smooth dialogue. Some behaviors serve to inform the speaker that the listener is paying attention and has understood what was said. They also play a role in turn-taking. However, while the use of these utterances may be intuitive and nearly unconscious for human speakers, it was nonintuitive for our operators to explicitly actuate them through a computer interface.
In our pre-trials, most participants prepared only topic-specific utterances, and they did not prepare topic-independent utterances. This resulted in awkward, one-way conversations where the robot seemed unresponsive to things the visitor said. Our guidelines explicitly recommend that operators prepare topic-independent utterances, as their use can lead to smoother, more natural interactions where the robot appears more responsive.

3.2 Initiative

In an ideal conversational situation, the dialogue literature would suggest that the robot and the visitor be given equal footing in terms of taking control of the conversation, and that a truly mixed-initiative system would result in better interactions than a fully robot-driven dialogue. However, there is an asymmetry in the system – for the operator to type a response to an unexpected question incurs a cost in terms of waiting time which would not occur in face-to-face conversation.
When the visitor takes the initiative in a conversation, it is likely that the conversation will move into topics and questions that are not covered in the robot's content store. There is an important trade-off here, similar to the "exploration vs. exploitation" trade-off found in reinforcement learning systems – a small number of such unprepared situations could be acceptable, and indeed informative, as they provide an opportunity for the operator to input new and useful content. However, too many unprepared situations will result in the robot's responses being unacceptably slow, as the operator must type every response.
Consider this example from our preliminary tests. The operator had created a rich set of utterance content talking about the deer in Nara Park, and was prepared to explain many things about their history, their life, and their involvement in local festivals and traditions. However, at the beginning of one interaction, the operator took a long time to choose the first utterance. During this time, the visitor tried to think of a question to fill the silence, and asked "what is a good souvenir to get from Nara?" Since this question was outside of the set of prepared utterances, the operator needed to type a response, making the visitor wait for several seconds. After a long, awkward pause, the robot finally answered, "how about sushi?"
Whereas entering a single utterance like this might be considered valuable, because the operator was able to increase the amount of content in the system, the visitor continued to ask questions in this topic, e.g., "What type of sushi is good," and, "Where is a good shop to buy it?" Each time, the operator needed to make the visitor wait for several seconds to type a new response, making the visitor more and more impatient.
Rather than letting the visitor drive the conversation so far outside of the robot's area of expertise, the operator needs to take the initiative, directing the conversation towards topics that it can speak about. While this does not produce an ideal interaction based on equal footing, directing the topic of conversation to the robot's area of expertise should result in higher responsiveness than an open-ended conversation would. Similar strategies have been proposed elsewhere, e.g. in [16].

Setting expectations In pre-trials, many participants designed behaviors only to say "Hello" or "Please ask me any question" at the beginning of an interaction, leaving the visitor confused as to what to do. This often forces the visitor to take the initiative without clearly understanding the robot's purpose.
People interacting with the robot for the first time will not have clear expectations of the robot's abilities or role, so our guidelines recommend that the operator start off each interaction by establishing the robot's role and abilities, as well as initiating the first topic of conversation.

Initiating topics Whenever the conversation stops progressing smoothly, the robot should initiate a new topic

This is a preprint manuscript.
The final publication is available at Springer via http://dx.doi.org/10.1007/s12369-015-0288-9

of conversation, rather than leaving this task to the visitor. Doing so not only makes the visitor more comfortable because the robot is leading the conversation, but it can also ensure that the topics are limited to the robot's prepared conversation content.

Minimize silence time If the robot makes the visitor wait too long, the visitor may choose not to wait for the robot to respond. As long pauses are uncomfortable, the visitor will often jump to an unrelated topic during long silences. Just as in the "responsiveness" guidelines, the best way to avoid these situations is for the robot to respond quickly and avoid long silences.

3.3 Interactivity

Many participants in our preliminary trials tended to create long monologues for the robot. Such one-sided interactions should be avoided, as visitors will feel bored and lose interest in the conversation if they are only listening and not participating. For example, one utterance prepared for the robot said "Let me tell you about Todaiji temple. Todaiji temple and the Big Buddha were first established in 743. This was because, due to earthquakes, hunger, and war, the emperor Shomu thought that Buddhism might be help to save the country. It needed too many ...". This was too long of a one-sided explanation and boring for the people interacting with the robot.

Asking questions To enable an interactive conversation without exposing the robot to many unexpected questions, we recommend that the robot should actively ask questions to the visitor. This allows the visitor to participate in the conversation while the robot keeps the initiative. An additional benefit of asking questions to the visitors is that replies to the questions are often predictable, so it is possible to prepare responses to expected replies.
For example, while the robot is explaining about the Great Buddha statue, it might ask, "How tall do you think the statue is?" instead of simply stating the statue's height in meters. Visitors will probably respond with an estimate which is correct, too low, or too high, or they will simply say they don't know. It is straightforward to prepare content to respond to each of these four cases, e.g. for the case that the guess is too high, the operator could prepare as "Well, it is very tall, but it's not THAT tall!"

4 System implementation

In pre-trials, we observed that operators often did not follow the guidelines when we only told them orally. For example, some operators continued to type every behavior even after we told them "typing takes too much time, so please use the existing behaviors". Other operators continued to make long explanatory behaviors, even after we told them "please make behaviors short so you can see people's reactions and not make the interaction boring".
We thus designed a software system to assist operators in following the guidelines, for both content development and teleoperation of the robot. The objectives of this system were to minimize the effort and search time of the operators during live teleoperation, to streamline the processes of content development and review/improvement of content, and to provide reminders and assistive interfaces encouraging the operators to follow the guidelines.
In order to support the operator's tasks of content development, teleoperation, and review of interactions, the software system provides separate graphical interfaces for each of the following three phases in the development process:

Design phase Operators prepare a basic set of content to enable simple conversations when operation begins. Until this basic content has been developed, it will not be possible to begin gathering feedback through teleoperation. Also, when a robot is in the field interacting with real people, some basic level of conversational ability is necessary. Failed conversations due to insufficient content preparation will generally be unacceptable.

Operation phase Operators teleoperate a robot in interactions with real people using the content they have prepared. These interactions need to be conducted at a natural conversational speed for smoothness, and they should strike a careful balance between exploring new content (e.g. taking the time to type answers to new questions) and presenting prepared content (which is much faster to actuate).

Consolidation phase Operators improve content through feedback from interactions. It is important to review the quality of the interactions and identify areas where the content base can be improved. This can be achieved by reviewing a transcript of the interaction, watching video of the interaction, and responding to system-generated prompts about recommended content changes.

This is a preprint manuscript.
The final publication is available at Springer via http://dx.doi.org/10.1007/s12369-015-0288-9

Table 1 Main guidelines

Name	Main guidelines	For Design/Consolidation phase	For Operation phase
Responsiveness	React to what the visitor says during the interaction (make responses, change topics, greetings at the end)	**(A1)**Make behaviors short **(A2)** Write one idea in one behavior **(A3)** Make topic-independent utterances	**(A4)** Use topic-independent utterances **(A5)** Watch and listen to the visitor carefully
Initiative	Lead the interaction (set expectations, initiate topics, minimize waiting time)	**(B1)** Design behaviors in a flow so the robot can lead the conversation	**(B2)** Choose behaviors smoothly at first **(B3)** Avoid typing too much **(B4)** Keep the conversation focused on prepared topics
Interactivity	Help the visitor participate in the interaction by asking questions	**(C1)** Make questions and prepare for the likely responses	**(C2)** Proactively ask questions

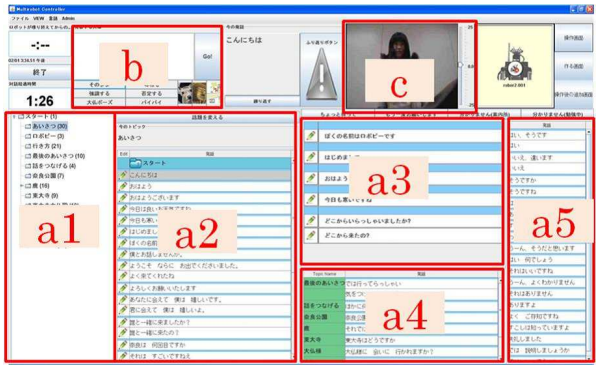

Fig. 2 Operation view. The interface features a hierarchical topic tree (a1), a flat list of all behaviors within the selected topic (a2), a recommended list of behaviors to choose next (a3), a list of behaviors which can introduce other topics (a4), a list of topic-independent behaviors (a5), a tool for entering new text and choosing gestures for a new behavior (b), and a video feed from the robot's eye camera which shows the person interacting with the robot (c).)

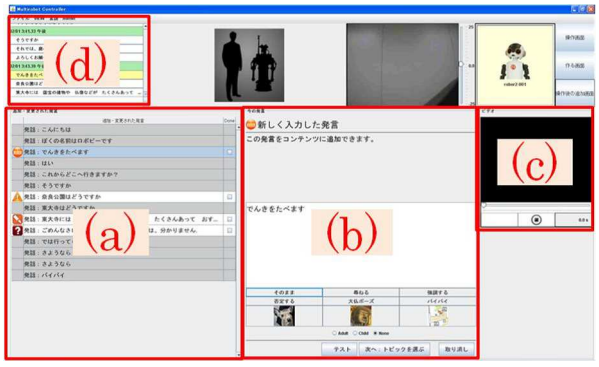

Fig. 3 Consolidation view. The interface features a list of all behaviors actuated during the selected interaction, including icons marking behaviors which were entered, edited, or flagged during teleoperation (a), a "wizard" panel showing a custom interface for each consolidation task, such as editing a behavior (b), a video panel for reviewing the interaction (c), and a panel for selecting other past interactions (d).

4.1 Design phase

The primary tasks of the operator in the design phase are to create, edit, and organize interaction content for the robot. This content takes the form of "behaviors" which can include both utterances and gestures.

Topic organization In our system, these can be organized into a structure of topics and subtopics.

Add new behavior wizard The system also provides a wizard interface to guide the operator through the process of creating a new behavior and categorizing it. Based on feedback from the operators, we determined a limit of 10 behaviors in one topic to be "too many" to search and choose smoothly during interaction. Thus, if a topic contains more than 10 behaviors, the system suggests to divide it into subtopics. Keeping individual behaviors short is also important for responsiveness, so if an utterance is too long, the wizard suggests to divide it into shorter behaviors. While our choice of 10 behaviors was determined informally, there is some similarity to the "seven plus or minus two" limit in human cognitive processing [25].

4.2 Operation phase

The Operation view, shown in Figure 2, enables the operator to teleoperate the robot in conversational interactions with people.
In many ways, this interface is similar to those used in "Wizard of Oz" experiments [8,6,38], but in our system several features have been specifically developed to support the operator in teleoperating the robot quickly and efficiently.

This is a preprint manuscript.
The final publication is available at Springer via http://dx.doi.org/10.1007/s12369-015-0288-9

Basic interface The video stream from the robot's camera showing the visitors is shown in area (c). The operator can select topics in area (a1) and execute behaviors in area (a2) by clicking on them. The operator can type text to be spoken and then stored as a new behavior for a robot in area (b).

Links Some interactions include behaviors that occur in a predictable sequence. For example, after the robot says "what's your name?" its next utterance will often be, "nice to meet you!" The system records the frequency of such transitions and uses this data to suggest likely candidates for the next behavior in area (a3). While this method is not accurate enough to directly be used for automating the robot's behaviors, it can help to reduce the operator's search time.

Topic shortcuts Certain key phrases are often used when changing topics. For example, if the operator wanted to start talking about the Great Buddha statue in Nara, a behavior saying "So, have you gone to see the Great Buddha yet?" would be a likely candidate. We allow operators to identify such behaviors and mark them as "shortcuts" to be displayed in area (a4). We have found this to be quite useful for ending interactions smoothly.

Topic-independent utterances As described in Sec. 3.1, we classify utterances such as "oh, really?", "thanks!", or "yes, that's true" as topic-independent utterances. These behaviors tend not to belong to the context of any particular topic, so they are shown in their own list in area (a5). The behaviors in this list are created and updated by the operator in the same way as all other behaviors.

Auto filler When the operator is slow in choosing or typing a behavior, the system helps to fill the delay by automatically inserting conversational fillers such as "hmm..." or "please wait a moment" to avoid awkward silences [34]. Details about our implementation of this feature are explained in Sec. 5.2.7.

Memo button To help the operator remember any interaction problems that occurred during operation, we provide a "memo button" which marks that point in the interaction for review during the Consolidation phase. This button can be used when the operator would like to add further explanation, improve the phrasing of an utterance, improve the connection between explanations, prepare related behaviors, fix the robot's pronunciation, or change its gestures. The point at which the memo button was pressed is recorded and shown in the Consolidation view.

4.3 Consolidation phase

After the end of every interaction, the system guides the operator through a "consolidation" process of self-evaluation, review of the interaction, and improving the content based on that feedback (Figure 3). This process is intended to help improve both the robot's interaction content and the operator's operation technique.

Guideline checklist panel After each interaction, the system first presents a dialog box, which asks the operator to self-evaluate their performance on a checklist of guidelines.

Consolidation procedure After this self-evaluation, the Consolidation view is shown. The most recent interaction is automatically selected, and the system prompts the operator to watch the video of that interaction. After watching the video, the operator is guided through each of the consolidation tasks in sequence, and the system offers possible actions for each task in area (b). For example, if a new utterance was typed, the operator can add it as a new behavior, edit it, or ignore it. If the memo button was pressed, the operator can add a new behavior or edit an existing behavior.

Interaction selection panel In area (d), the operator can choose an interaction to view, for cases where interactions happen back-to-back with no time for review in between.

Video In area (c), operators can watch a video of the selected interaction. Operators tend to focus on operational tasks while controlling the robot, so reviewing a video of the interaction helps the operator to step back and watch the content and timing of the interaction itself. It has been shown that operators can have an impaired awareness of time during teleoperation tasks [10].

Interaction Transcript Operators can see a transcript of the robot's side of the selected interaction in area (a) in Fig. 3. Entries representing likely "consolidation tasks" are highlighted, including new behaviors typed during that interaction, behaviors that were edited, and points where the memo button was pressed.

5 Experiment

We conducted a field experiment to evaluate the effectiveness of the developed system and guidelines in the context of the tourist information scenario explained in

This is a preprint manuscript.
The final publication is available at Springer via http://dx.doi.org/10.1007/s12369-015-0288-9

Section 1.4. For this experiment, we placed a robot in a tourist information center in Nara, with the task of explaining local sightseeing information to tourists.

5.1 Experimental design and predictions

In order to demonstrate the necessity and effectiveness of our system and guidelines, we designed an experiment to confirm two predictions.
First, we needed to validate our assumption that using the guidelines would result in better interactions and better interaction content. Thus, we proposed a first prediction as follows:

Prediction 1 Operators with the proposed system and guidelines will make better content and operate the robot more effectively, resulting in a better **overall impression** of the robot than in the case of operators without the system or guidelines.
Next, supposing that the first prediction is supported, we aimed to confirm whether the improvement is due to the fact that the operator followed the given guidelines. We found in our informal preliminary experiments that operators do not tend to naturally follow the guidelines. To confirm that our approach increases compliance with the guidelines, the second prediction we tested was as follows:

Prediction 2 Participants who are provided with the proposed system and guidelines will also show behavior that is more consistent with the guidelines.
To evaluate these two predictions, we conducted an experiment comparing the performance of participants who prepared content and operated the robot. We compared the performance of two groups of participants: one using our proposed system and guidelines, and one using a baseline system without our proposed features or guidelines. Each participant operated the robot in interactions with real visitors at the tourist information center.

5.2 Method

5.2.1 Settings

The experiment consists of two parts: preparation (the Design phase in our proposed flow) and operation (iterating through the Operation and Consolidation phases several times). One day was spent on each part. Thus, each participant took part in this experiment for two days.
On the first day, the experiment was conducted at our laboratory, where the participants created an initial set of content for the robot. This session lasted for six hours. Participants were given one hour of instruction on how to use the system, three hours to create interaction content, and two hours to practice with the system by teleoperating the robot by themselves.
On the second day, the experiment was conducted in a tourist information center in Nara (Figure 4). The participants operated the robot for four hours. In these interactions, they presented information and answered any questions that visitors had. Participants were given breaks between the training and operation phases, and due to the small number of visitors to the center, the overall workload of the operators was relatively low.
The human-robot interactions at the tourist information center were recorded using external video cameras for later analysis. All operations of the operators, such as sending commands to the robot or creating new behaviors, were also recorded with timestamps.

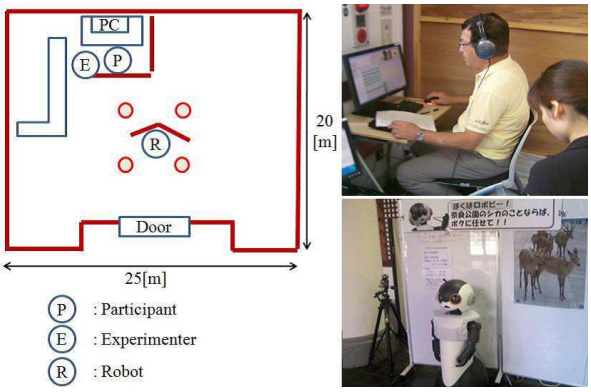

Fig. 4 Field experiment environment

5.2.2 Scenario

Due to the limited amount of time allocated for this experiment, we limited the range of conversation to a single topic. We specified that the robot was to act as a guide specializing in talking about the famous deer in Nara Park, one of Nara's major sightseeing attractions. Its job was to talk with visitors and try to interest them in the deer, as well as to answer any questions they had about the deer. If visitors had questions outside of the topic of deer, the robot was to direct them to the desk staff instead of trying to answer every question itself.

5.2.3 Robot

In the experiment, we used the humanoid robot Robovie R3. It has a human-like appearance with two arms (2*4 DOF), a head (3 DOF), and is 110 cm tall. Its head has two eye cameras, a speaker, and a microphone.

This is a preprint manuscript.
The final publication is available at Springer via http://dx.doi.org/10.1007/s12369-015-0288-9

XIMERA software [19], was used for speech synthesis. The robot is mounted on a mobile base, although locomotion was not used in this experiment.

5.2.4 Control software

The operators in our experiments controlled the robot using the system described in Section 4, implemented in Java and running on a Windows PC. The interaction content created by the operator, including gestures and utterances, was stored in a database. When the operator chose a behavior for the robot to execute, the contents of the behavior were sent to the robot, which synthesized the utterance and executed the appropriate gestures.

5.2.5 Participants

A total of 27 participants (23 men and 4 women, who averaged 68.4 years old, s.d. 3.96) took part in our experiment as operators. All were members of "Suzaku," a volunteer guide association in Nara. They each had 2-15 years of experience, and they were all currently active as volunteer guides at popular sightseeing areas at Nara at the time of the study. They had not previously interacted with our robot and had not had any experience operating any kind of conversational robots.
Each participant provided their age and number of years of experience as a guide, and because many of them were not frequent computer users, we measured their computer ability, in terms of typing speed and speed of controlling a mouse. Based on this information, we assigned participants to conditions in order to balance these factors as closely as possible between conditions.
The participants were not directly compensated for their participation in the experiment, but monetary compensation was paid to the guide organization for their participation in the study, and participation in our study was then treated as part of their normal duties as a member of that organization.
The visitors who interacted with the robot at the tourist information center were not compensated in any way for their participation. They were not told that the robot was teleoperated, although they may have known about the teleoperation via a public media announcement made prior to the experiment. An assistant standing near the robot encouraged people to talk with it if they appeared shy or hestitant, but the visitors were not given any specific instructions, e.g. specifying that they should limit their questions to the topic of deer.

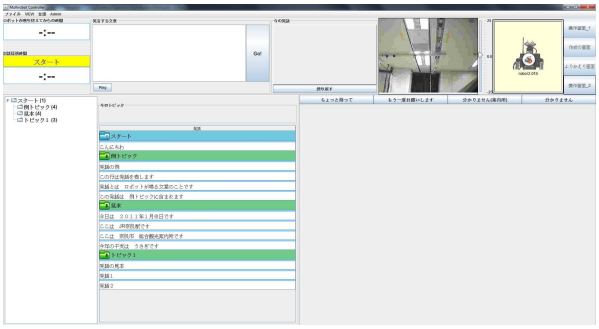

Fig. 5 Simpler interface for *without-assistance* condition.

5.2.6 Conditions

We used a single-factor between-participants experimental design, comparing operator performance in *with-assistance* and *without-assistance* conditions, defined as follows:

With-assistance In this condition, we provided the guidelines and the developed system to participants.

Without-assistance In this condition, we did not provide the guidelines or the features of the system which were designed to support the guidelines. Instead, we provided a basic system to allow them to enter content and operate the robot, and we allowed them to freely create and edit the content. The interface of this simpler system is shown in Figure 5.

5.2.7 Procedure in each condition

The following points explain the differences and similarities between the two experimental conditions.

Differences:

- **Content entry:** In the *with-assistance* condition, participants used the "add new behavior wizard" to create behaviors and organize them into topics. In the *without-assistance* condition, participants created behaviors by typing text into a simple list of utterances. These utterances could be categorized into topics, but none of the assistive features, such as warnings about utterance length or the ability to mark topic shortcuts, were provided.
- **Operation interface:** Of the features described in Section 4.2, links, topic shortcuts, the topic-independent utterance list, instances, and the memo button were not provided in the *without-assistance* condition.

This is a preprint manuscript.
The final publication is available at Springer via http://dx.doi.org/10.1007/s12369-015-0288-9

– **Consolidation:** The Consolidation view and guideline checklist were not provided to participants in the *without-assistance* condition. However, a log of all utterances they had typed during teleoperation was provided so they could easily add their typed utterances to the list of behaviors, rather than typing them from memory.
– **Video review:** In the *with-assistance* condition, the operator was actively prompted to watch the videos, as described in Section 4.3. Participants in the *without-assistance* condition were given video of their interactions to watch if they wished, but they were not actively asked to do so by the system.

Similarities:

– **Conversational fillers:** Automatic conversational fillers were used in both conditions, as we have used these in many studies and consider them to be necessary [10]. The system started automatic conversational fillers based on expected operation time - when switching topics (detected when the operator clicks within area (a1) in Figure 2) it is assumed that the operation will be slow, as the operator must read through the text of several utterances in the new topic. When typing a new utterance (detected when the operator clicks within area (b) in Figure 2), it is expected that operation time will be extremely slow. According to the expected operation time, the robot said appropriate fillers, e.g. "please hold on a second," or "hmm. . . ," in order to fill the silence.
– **Guidance:** In both conditions, the experimenter answered questions from the participants about how to use the system. In the *with-assistance* condition, a document explaining the guidelines was given to the participants. The experimenter read through the document with them and confirmed that the participants understood the meaning of the guidelines.
– **Gestures:** Gestures were not used in this experiment.

As factors such as computer experience, typing speed, and age may affect the results of this experiment, we balanced participants between conditions based on these factors, from the questionnaire and computer ability tests administered before the experiment. Table 2 shows the average ages and typing and mouse usage test scores for participants assigned to each condition.

Table 2 Participants assigned to each condition

	With-assistance	Without-assistance
Age (s.d.)	69.3 (4.0)	67.5 (3.9)
Typing (s.d.)	62.2 (6.2)	69.4 (6.1)
Mouse (s.d.)	32.1 (2.0)	30.7 (2.0)

5.3 Measurement and Results for Prediction 1

Our first prediction was that "operators with the proposed system and guidelines will make better content and operate the robot more effectively, resulting in a better **overall impression** of the robot than in the case of operators without the system or guidelines". We measured the overall impression of interactions between the robot and visitors in which two evaluators, blind to the experimental conditions, watched videos of the interactions and gave subjective quality ratings on a continuous 100-point scale. This evaluation method was chosen instead of directly asking visitors for their impressions, due to the difficulty of getting consistent evaluations from first-time visitors; the robot is still novel and an interaction with the robot is still fun for many people, even with poor interaction content. Thus, we decided to measure the overall impression from a third-person perspective to provide more consistent evaluations and enable better comparison between interactions.

We explained the role of the robot to the evaluators and asked them to rate how well it performed in its role as a guide providing information about deer in Nara Park, and how well it was able to engage in interactive conversation with the visitors. Evaluations were averaged over the final three interactions for each participant.

To provide a consistent scale for the evaluators, we gave reference definitions for 20-point increments, based on a scenario where the evaluator, as an employer, is choosing whether or not to hire the robot. In this scale, 100 is the best, 80 means that the evaluator feels that, as an employer, she could pay the robot slightly more than average, 60 points is normal, and the evaluator feels that she could pay the robot slightly less than average, 40 is not good, and the evaluator feels she would not pay the robot but she could employ him without pay, 20 points is bad, and the robot could be forgiven for his bad actions by saying " I'm training", and 0 points is unacceptable, where the evaluator felt she would not hire him even if he worked for free.

We computed the Pearson correlation of the overall impression scores between the two evaluators to be .645, which we consider to be a good match. Figure 6 shows the result of overall impression scores averaged between the two evaluators. A one-way factorial analysis of variance (ANOVA) was conducted, and a significant main effect was revealed ($F(1, 25)$=4.590, $p = .042$, η^2 =.155). Thus, overall interaction quality was shown to be sig-

This is a preprint manuscript.
The final publication is available at Springer via http://dx.doi.org/10.1007/s12369-015-0288-9

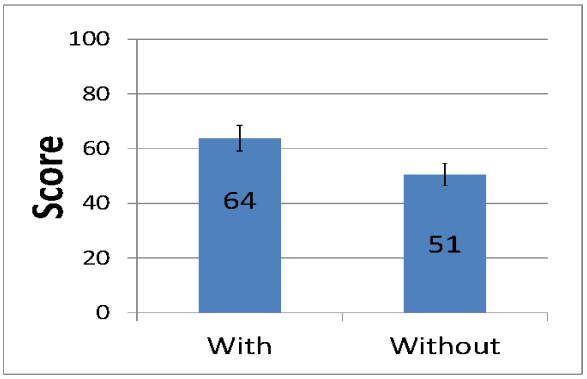

Fig. 6 Overall evaluations of interaction quality

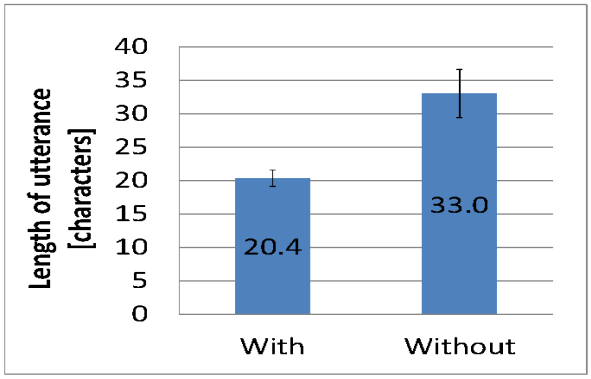

Fig. 7 Length of utterances prepared

nificantly better in the *with-assistance* condition, supporting Prediction 1.

5.4 Measurement and Results for Prediction 2

We believe that Prediction 1 was supported as a result of better compliance with our guidelines. To confirm this, we evaluated our second prediction: "participants with the proposed method will show behavior that is more consistent with the guidelines than participants without the system or guidelines." To evaluate this prediction, we analyzed the degree to which each participant followed each of the guidelines. We evaluated each of the eleven guidelines shown in Table 1 and compared the results between the two experimental conditions.
As the number of interactions conducted by each participant varied from day to day, data analysis was performed only on the final three interactions conducted by each participant.

(A1) Make behaviors short: prevent operators from entering long utterances which could result in boring explanations or one-sided interactions. We evaluated the average length of utterances prepared by each participant, measured by the number of Japanese characters. Figure 7 shows the result of length of utterances created. A one-way factorial ANOVA was conducted, and a significant main effect was revealed ($F(1, 25) = 11.743$, $p = .002$, $\eta^2 = .320$) Length of utterances created was shown to be significantly shorter in the *with-assistance* condition.

(A2) Write one idea in one behavior: When multiple topics are combined into a single behavior, it can seem that the robot is driving the conversation in a one-sided way, rather than reacting to the visitor. So we recommend that operators should prepare behaviors limited only to a single idea. To evaluate whether participants followed this guideline, we counted the average number of distinct ideas per behavior for each participant. In contrast with our preliminary trials, this was not a problem in our comparison experiment. Fewer than 1% of behaviors created in either condition contained more than one idea or topic (0.43% in the *with-assistance* condition vs. 0.25% in the *without-assistance* condition). A one-way factorial ANOVA was conducted, and no significant main effect was revealed ($F(1, 25) = 0.10$, $p = n.s.$) These results showed that operators of both conditions made almost all behaviors with one idea. We think that it is because they only talked about one topic, deer, in this comparison experiment. Thus we did not see the same problem which occurred in our pre-trials, where the robot presented a large amount of content on multiple topics.

(A3) Make topic-independent utterances: As described in Sec. 3.1, we have found topic-independent utterances to be important for making interactions reactive and smooth. In our pre-trials, many operators prepared no such utterances, resulting in awkward interactions. As a measurement, we evaluated the number of participants who made at least one topic-independent utterance. The number of operators who made at least one topic-independent utterance was 13 out of a population of 14 in the *with-assistance* condition, and 6 out of a population of 13 in the *without-assistance* condition. A Chi-squared test revealed a significant difference between conditions ($\chi^2(1) = 4.990$, $p < .05$). Thus, significantly more operators in the *with-assistance* condition created topic-independent utterances.

(A4) Use topic-independent utterances: Our guidelines also recommend using topic-independent utterances during operation (not only creating them at design time). Note that this can be measured independently of A3, because even if topic-independent utterances are not prepared beforehand, operators can still use them in

This is a preprint manuscript.
The final publication is available at Springer via http://dx.doi.org/10.1007/s12369-015-0288-9

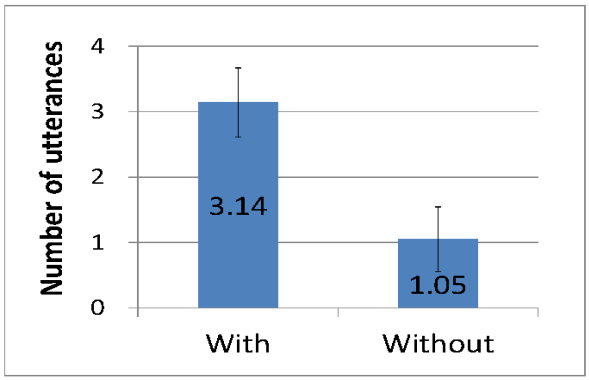

Fig. 8 Number of topic-independent utterances used

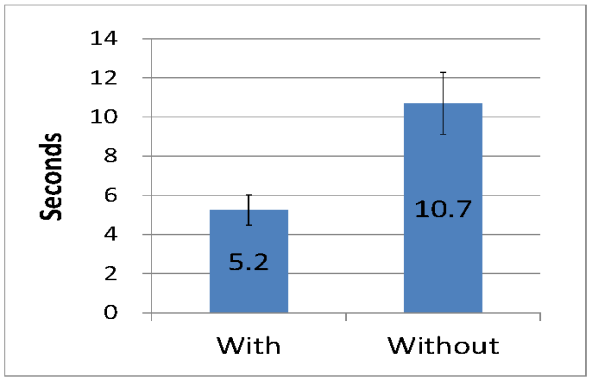

Fig. 9 Average wait time

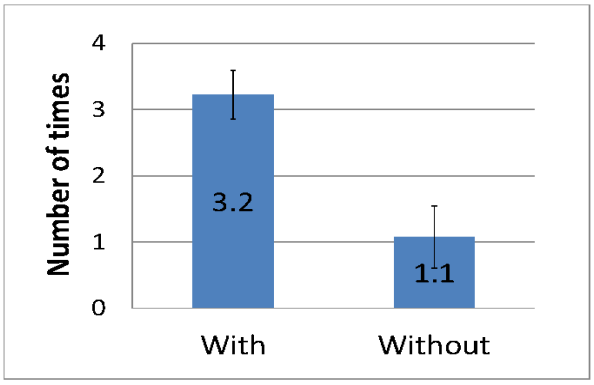

Fig. 10 Number of robot-initiated topics during operation

conversation by typing them in on the fly. We evaluated the number of topic-inde-pendent utterances used during interaction.
Figure 8 shows the number of topic-independent utterances used. The Kolmogorov-Smirnov test was performed to check the normality of the number of topic-independent utterances used; the test indicated that the numbers of topic-independent utterances in both conditions were not normally distributed. Therefore, a Mann-Whitney U test was conducted instead of ANOVA. A significant main effect was revealed with $p < .05$, U=30. Thus, these results show that participants in the *with-assistance* condition used significantly more topic-independent utterances than participants in the *without-assistance* condition.

(A5) Watch and listen to the visitor carefully: In our pre-trials, the operator sometimes explained about many points rapidly without leaving time for the visitors to make any response. To encourage responsiveness, our guidelines remind the operator to listen carefully and quickly react to what the visitors say. To evaluate each participant's responsiveness, we measured the average wait time between a visitor speaking and the response from the robot. Auto-fillers such as "um..." were not counted as responses. Wait time was averaged across the three interactions.
Figure 9 shows the average wait time for visitors during an interaction. A one-way factorial ANOVA was conducted. A significant main effect was revealed ($(1, 25) = 10.193$, $p < .01$, $\eta^2 = .290$), showing that operators in the *with-assistance* condition were more responsive and made the visitors wait less.

(B1) Design behaviors in a flow so the robot can lead the conversation: We saw in our pre-trials that operators were given many questions from visitors, and we observed that it can take a lot of time to search for an answer, during which time the visitor is made to wait. To avoid such delays, we proposed that operator should lead the interaction. To evaluate how well the operator led the interaction, we evaluated the number of robot-initiated topics during each interaction. (e.g. a robot starts to talk about a new topic, like " let's talk about what the deer eat", or a robot asks a question about a new topic " Do you know how many deer are in Nara park?")
Figure 10 shows the average number of times that participants initiated new topics per interaction. A Kolmogorov-Smirnov test performed to check the normality of the number of new topics initiated per interaction indicated that they were not normally distributed. Therefore, a Mann-Whitney's U test was conducted instead of ANOVA. A significant main effect was revealed ($p < .05$, U =27) showing that participants in the *with-assistance* condition initiated significantly more new topics than those in the *without-assistance* condition.

(B2) Choose behaviors smoothly at first: At the start of an interaction, before clear expectations for the content of the dialogue have been established, there is a risk that the visitor will take the initiative and guide the topic of conversation far outside of the intended topic of focus. Our guidelines thus recommend that the operator establish the topic of conversation quickly, without

This is a preprint manuscript.
The final publication is available at Springer via http://dx.doi.org/10.1007/s12369-015-0288-9

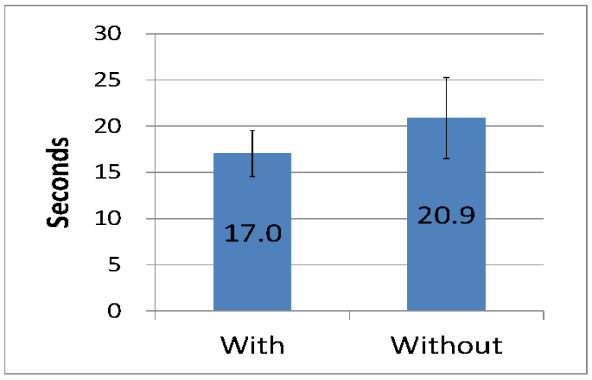

Fig. 11 Average time before the robot or visitor talks about deer

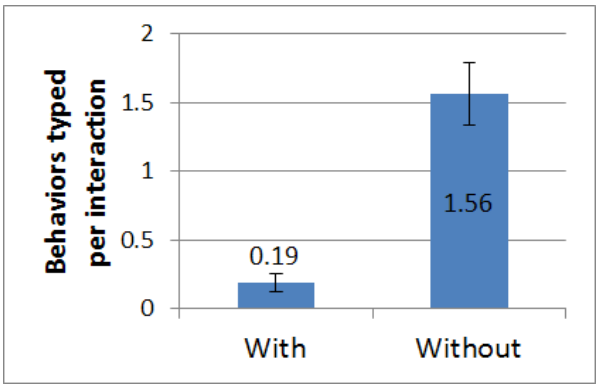

Fig. 12 Average number of behaviors typed per interaction

allowing long silences. To evaluate how well operators followed this guideline, we measured the average time from the beginning of the robot's first utterance in an interaction until the first utterance from the robot or visitor regarding deer (the target topic).
Figure 11 shows the average time before the robot or visitor talked about deer. A one-way factorial ANOVA revealed no significant difference between the two conditions.

(B3) Avoid typing too much: When content has not been prepared to cover a situation, the operator must type a new utterance, a slow process that makes the visitor wait. We counted the total number of utterances typed by the operator in each interaction and averaged this over the three analyzed interactions for each participant. Figure 12 shows the number of typed responses per interaction. The Kolmogorov-Smirnov test was performed to check the normality of number of typed responses; the test indicated that numbers of typed responses in both conditions were not normally distributed. Therefore, a Mann-Whitney U test was conducted instead of ANOVA. A significant main effect was revealed ($p < .01$, $U = 155$).

(B4) Keep the conversation focused on prepared topics: When the conversation moves to a topic for which content has not been prepared, the operator must type a new utterance. This is usually quite time-consuming and makes the visitor wait, so our guidelines recommend keeping the conversation focused on prepared topics. We evaluated this guideline by counting the number of times the operator used two or more behaviors to talk about something aside from the prepared topic, since at least one behavior is necessary to respond to any off-topic question.
This evaluation showed that off-topic diversions were successfully avoided by 14 out of 14 participants in the *with-assistance* conditions, and 9 out of 13 in the *without-assistance* condition. A chi-squared test revealed a significant trend between conditions ($\chi^2(1) = 1.913$, $p < .10$). Two of the four operators who made off-topic diversions were asked questions about the robot by visitors. These operators answered to the questions by typing some behaviors, and the topic continued for a few turns. The other two operators voluntarily chose to type behaviors, in order to make personalized conversation with visitors, but the conversation drifted away from the target topic. All of these interactions included long silence time.

(C1) Make questions and prepare for the likely responses: One way for operators to make conversations more interactive and interest visitors is by asking questions about informational content, rather than simply presenting statements about it. This helps to maintain interactivity while the robot is leading the interaction. To evaluate this point, we counted the number of operators who made at least one question behavior.
The number of operators who made one or more question behaviors was 14 out of a population of 14 in the *with-assistance* condition, and 4 out of a population of 13 in the *without-assistance* condition. A chi-squared test revealed a significant difference between conditions (χ^2 (1) = 11.590, $p < .01$), showing that significantly more operators in the *with-assistance* condition made question behaviors for visitors.

(C2) Proactively ask questions: To confirm that operators not only made questions but also used them, we counted the average number of questions asked by the robot to visitors per interaction.
Figure 13 shows the number of questions asked by the robot per interaction. The Kolmogorov-Smirnov test was performed for checking the normality of number of questions; the test indicated that numbers of questions in both conditions were not normally distributed.

This is a preprint manuscript.
The final publication is available at Springer via http://dx.doi.org/10.1007/s12369-015-0288-9

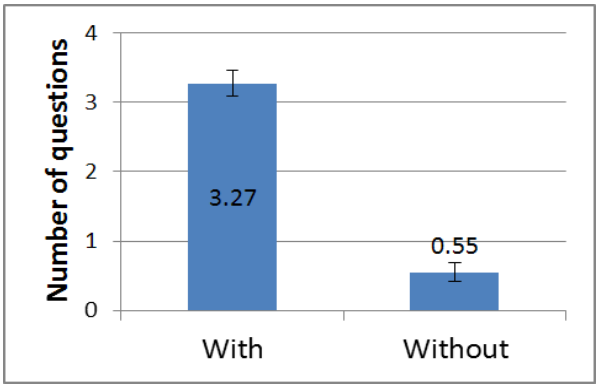

Fig. 13 Number of questions asked by the robot per visitor

Therefore, a Mann-Whitney U test was conducted instead of ANOVA. A significant main effect was revealed ($p < .01, U = 6$), indicating that significantly more questions were asked to visitors in the *with-assistance* condition.

Thus, significant results were obtained for 9 of the 11 guidelines showing that participants in the *with-assistance* condition were more likely to follow the recommendations of the guidelines in the *with-assistance* condition than in the *without-assistance* condition, supporting Prediction 2. Summarized results for all eleven guidelines are shown in Table 2.

5.5 Who was more or less successful?

We further analyzed the results to gain a deeper understanding of the factors affecting interaction quality. Figure 14 shows the interaction quality scores of all participants in descending order. The color represents the condition of each participant. The top five, with scores around 80 points, were all in the *with-assistance* condition.

The best interactions followed the proposed guidelines well. For instance, the example below shows the robot leading the interaction by asking a question with prepared responses:

> Robot: Do you know how many deer are in Nara park? *(leading with a question)*
> Visitor: Hmm, how many? ... maybe 500?
> Robot: More than that. *(prepared response)*
> Visitor: How many?
> Robot: In Nara Park, there are 1200 deer. 25% are male, 75% are female, and 10% are children.

The robot did not make the visitor wait, because all of the situations were expected, and thus the contents for the robot's responses had been prepared and could be actuated quickly.
In contrast, a few participants in the *without-assistance* condition performed very badly. They typically made visitors wait a lot, because of failure in leading the interaction. This is one example of such failure:

> Robot: Where are you from?
> Visitor: I'm from Yamato-Koriyama.
> Robot:(5 seconds passed because the operator was typing) That's nearby.
> Robot: Um,... please wait (the operator is typing, and 17 seconds passed). When will their goldfish festival be held?

The robot asked a question, but had not prepared for the response. What is worse, the operator made the visitor wait and started to deviate from the topic of conversation. Probably the input from the visitor elicited the operator to make the response. It would be much easier to respond in a face-to-face interaction, but it was not easy when he was teleoperating the robot, because typing took so much time. The guidelines for *reaction time*, *silence time*, and *setting expectations* were prepared specifically to prevent situations like this.
While the guidelines helped prevent participants from getting into such awkward situations, it is notable that the majority of participants in both conditions had scores around 50. Our observation is that some participants were very quick learners, and even in the *without-assistance* condition, several participants naturally discovered and used techniques such as those recommended in the guidelines. Top participants in the *without-assistance* condition successfully used topic-independent utterances and took initiative in the conversation by having the robot ask questions.
By contrast, some of the *with-assistance* participants reported that the guidelines and system were too complex. They were required to learn a complex software application as well as pages of guidelines, with only a few hours of studying and very little practice. Some *with-assistance* participants did not succeed in fully following the guidelines, and typically failed to respond to what visitors said, which seemed to be due to excessive cognitive load. Participants in the other condition had a simpler system and so had more free time to think creatively about how to create appropriate content. We expect that the proposed guidelines and system would have provided better performance if we provided more time for training and practice.

This is a preprint manuscript.
The final publication is available at Springer via http://dx.doi.org/10.1007/s12369-015-0288-9

Table 3 Summary of evaluation results. Standard deviations are shown in parentheses.

Guideline	Measurement	With Assistance	Without Assistance	Significance
	Overall			
Interaction quality	Scores from evaluators	65.75 (17.11)	53.54 (14.71)	p <.05
	Responsiveness			
(A1) Make behaviors short	Length of utterances prepared	20.37 (4.51)	33.01 (13.00)	p <.01
(A2) Write one idea in one behavior	Percentage of behaviors containing more than one idea	0.43%	0.25%	n.s.
(A3) Make topic-independent utterances	Number of participants who made at least one topic-independent utterance	13 of 14	6 of 13	p <.05
(A4) Use topic-independent utterances	Number of topic-independent utterances used	3.14 (1.98)	1.05 (1.78)	p <.05
(A5) Watch and listen to the visitor carefully	Average wait time	5.24s (2.87)	10.71s (5.67)	p <.01
	Initiative			
(B1) Design behaviors in a flow so the robot can to lead the conversation	Number of robot-initiated topics during operation	3.22 (1.37)	1.07 (1.69)	p <.05
(B2) Choose behaviors smoothly at first	Average time before robot or visitor talks about deer	17.0s	20.9s	n.s.
(B3) Avoid typing too much	Number of behaviors typed per interaction	0.19 (0.41)	1.56 (1.47)	p <.01
(B4) Keep the conversation focused on prepared topics	Number of participants who successfully avoided off-topic diversions	14 of 14	9 of 13	p <.1
	Interactivity			
(C1) Make questions and prepare for the likely responses	Number of participants who made at least one question utterance	14 of 14	4 of 13	p <.01
(C2) Proactively ask questions	Number of questions asked by the robot per visitor	3.27 (1.21)	0.55 (0.86)	p <.01

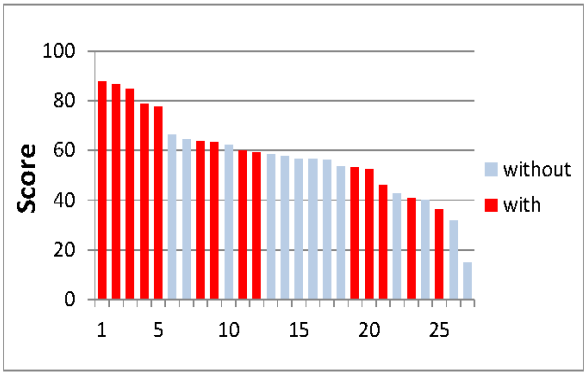

Fig. 14 Interaction quality of each participant

5.6 How was the quality of the information provided by the robot?

We counted the number of utterances containing informational content that were used by each operator. On average, 13.0 unique informational utterances were used per interaction in the *with-assistance* condition, compared with 6.1 in the *without-assistance* condition (Fig 15). A one-way factorial ANOVA revealed a significant difference between the two conditions ($F(1, 25) = 13.278$, $p = .001$, $\eta^2 = .347$))

We also counted the number of visitor's utterances showing surprise or interest (e.g. "Wow!", or "Oh, really?") per interaction (Figure 16). We conducted an ANOVA revealing a significant main effect ($F(1, 25) = 5.69$, $p = .025$, $\eta^2 = .185$). We interpret these results as indi-

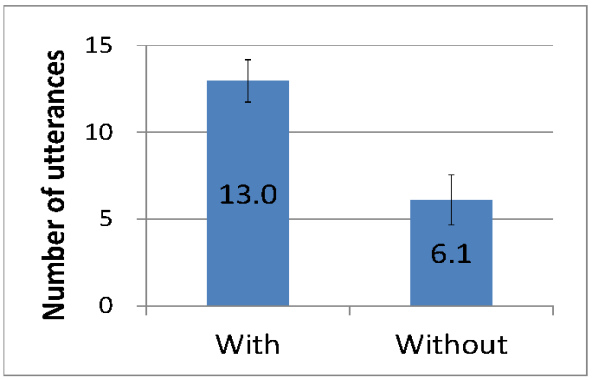

Fig. 15 Number of utterances containing informational content

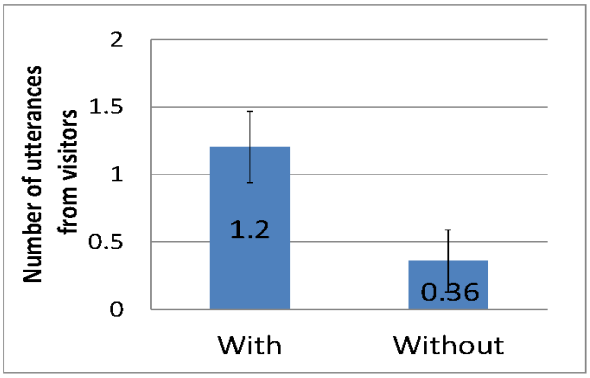

Fig. 16 Number of utterances showing surprise or interest from visitors

cating that operators in the *with-assistance* condition were able to conduct more interesting conversations.

6 Discussion

6.1 Contributions of Individual Components

We introduced many techniques and guidelines at once in this study, and it is not clear to what extent each element contributed to performance. The main focus of this study was to determine whether non-engineering domain experts could effectively be included in the content development process at all, so we did not do a rigorous analysis of each system component.

The contributions of the software system and of the guidelines cannot be meaningfully separated, as the purpose of the software is to assist users in following the guidelines. In the end, participants following all guidelines generally received high evaluation scores.

Aside from explicitly supporting the guidelines, one major contribution of the software system appeared to be the video playback functionality. After watching videos of their operation, participants made a noticeably greater effort to operate the robot quickly and minimize visitor wait time. We attribute this to greater self-awareness, which enabled the operators to more effectively use their intuition and implicit communicative knowledge in interactions.

The effectiveness of other functionalities, such as reminding operators of the guidelines, appeared to be most effective for the less capable operators, but possibly annoying to more capable operators. We attribute this to the idea that greater fluency in using the teleoperation system enables a higher level of immersion in the interaction, which in turn frees the operators to focus on the interaction itself and use their implicit knowledge. Operators who are unable to use the system fluently need to focus on the mechanics of operation and are not immersed in the interaction. In these cases, the guidelines have a more pronounced effect, as they can bootstrap the iterative content development process by assisting an operator in conducting a smooth interaction despite this lack of fluency.

The "Topic-independent utterance list" was quite useful, enabling operators to react to the customers quickly in many situations. Other features, such as the automatic links and topic shortcuts, were not so important for small data sets like those used in our experiment, but we expect that their value will increase for larger sets of content and longer periods of operation (since the links are built based on interaction history).

6.2 Guidelines

The set of guidelines developed in this study was specifically chosen for our scenario of short-term interactions providing information to tourists. We believe that most of these guidelines should be universally applicable, as they reflect general phenomena such as natural turn-taking behavior. However, some guidelines such as "proactively ask questions" or "make behaviors short" may be less applicable in some situations, depending on the nature of the task and on cultural and social factors.

6.3 Gestures

Although the capability of combining gestures and utterances to form behaviors was an important consideration in our system design, the use of gestures was not explored in depth in this study. Our system enabled operators to use specific pre-designed gestures such as emphasizing a point, waving goodbye, or pointing to a poster near the robot, but it included no capacity for generating new poses and gestures for the robot. Techniques for enabling operators to use gestures effectively

This is a preprint manuscript.
The final publication is available at Springer via http://dx.doi.org/10.1007/s12369-015-0288-9

and create new gestures could be interesting topics for future work.

6.4 Limitations

This study was relatively small-scale, as each participant had only 3 hours for content creation and 3.5 hours for operation, covering only one topic. We expect that our system and guidelines would be even more effective on a larger scale, as more operation time would mean participants would be trained better. While a larger set of content would make the operator's task more difficult, many of the features proposed in our system are mechanisms designed to support large content sets.
Whereas six hours could be considered to be quite long for a training period, one of the reasons for this long practice time is that the participants were elderly, and they needed more time to type content and to learn how to use the system than younger people would. We think younger people would need less time for training and practice on Day 1, and it may be possible to prepare more efficient ways to teach elderly people to use the system or explain the interaction guidelines.
Another limitation is that this study only focused on making dialog. We did not consider about robot locomotion, manipulation, or perception. In this paper we have focused on creating information content, and the effects of these tasks are beyond the focus of this paper. But considering of them would be an interesting future work in this direction of research. For example, when locomotion is used, the robot's current location could be used as additional context for predicting the behaviors shown in the "Links" area of the operation view.

6.5 Applicability to other domains

This study demonstrated that it is possible for domain experts who are not engineers or programmers to create interaction content for a conversational robot through an iterative process of content development and teleoperation. While this study focused on guiding tourists as a target application, we can consider many other areas where domain experts would be valuable in creating interaction content for a robot. Knowledge from domain experts might be necessary for robots working in a shop talking with customers and selling products, in a hospital or care home talking with patients and keeping them company, or in an educational setting helping students learn.
These applications differ slightly in their nature. The communicative knowledge needed by a sightseeing guide robot centers around storytelling, engaging listeners, and reacting to what they seem be interested in. A sales or education robot would have different strategies and goals for its interactions. However, requirements such as smoothness of the interactions and responsiveness to the customer or student would be similar, as would be the need for a basic level of interaction quality in order to bootstrap the iterative procedure of teleoperation and content design. Thus, we expect that the developed guidelines and system should be useful for such applications.

7 Conclusion

This paper addressed the challenges of using domain experts to create interaction content for conversational robots. To enable non-robotics domain experts to create content using their implicit communicative knowledge, we proposed an iterative process using teleoperation of the robot in real interactions to provide feedback for improving conversational content.
We presented a system and a set of design guidelines to support domain experts in creating, using, and improving conversational content through teleoperation. We then evaluated how well domain experts could make conversational content and operate a robot using our proposed guidelines and system through a field experiment in a real tourist information center. The results confirmed that with our system and guidelines helped operators in several ways: they were able to provide more timely, natural responses to visitors; they did not need to type as many new utterances during interactions, resulting in less wait time for visitors; and they were able to conduct better interactions overall. We believe these findings will be valuable as conversational robots are developed for new application domains.

Acknowledgements This research was supported by the Ministry of Internal Affairs and Communications of Japan. We wish to thank the staff at the Nara City Tourist information center, and the members of the NPO "Suzaku" for their helpful participation. We also wish to thank Satoshi Koizumi and Masaya Shimoyama for their help.

References

1. Bohus, D., Raux, A., Harris, T. K., Eskenazi, M. and Rudnicky, A. I., Olympus: an open-source framework for conversational spoken language interface research, HLT-NAACL 2007 workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technology (2007), pp. 32-39(2007).
2. Boose, J. H. and Bradshaw, J. M., Expertise transfer and complex problems: using AQUINAS as a knowledge-

acquisition workbench for knowledge-based systems. International Journal of Man-Machine Studies, 26(1), 3-28 (1987).
3. Castro-Gonzlez, ., Malfaz, M., and Salichs, M.A., Learning the Selection of Actions for an Autonomous Social Robot by Reinforcement Learning Based on Motivations, International Journal of Social Robotics, vol. 3, pp. 427-441, (2011).
4. Chao, C. and Thomaz, A. L., Turn Taking for Human-Robot Interaction. Paper presented at the AAAI Fall Symposium: Dialog with Robots (2010).
5. Chernova, S., Orkin, J. and Breazeal, C., Crowdsourcing HRI through Online Multiplayer Games, the AAAI Fall Symposium "Dialog with Robots", pp. 14-19(2010).
6. Dahlbck, N., Jnsson, A., and Ahrenberg, L., Wizard of Oz studies: why and how, International Conference on Intelligent User Interfaces, pp. 193-200 (1993).
7. Eraut, M., Non-formal learning and tacit knowledge in professional work, British Journal of Educational Psychology, 70:113-136(2000).
8. Glas, D. F., Kanda, T,. Ishiguro H., and Hagita,N., Teleoperation of Multiple Social Robots, IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans. Vol. 42, No. 3, pp. 530-544, (2012).
9. Glas, D. F., Satake, S., Kanda, T. and Hagita, N., An Interaction Design Framework for Social Robots, Robotics: Science and Systems Conference (2011).
10. Glas, D. F., Kanda, T., Ishiguro, H. and Hagita, N., Temporal Awareness in Teleoperation of Conversational Robots, IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans, Vol. 42, No. 4, pp. 905-919(2012).
11. Gockley, R., Bruce, A., Forlizzi, J., Michalowski, M., Mundell, A., Rosenthal, S., Sellner, B., Simmons, R., Snipes, K., Schultz, A. C. and Wang, J., Designing Robots for Long-Term Social Interaction, IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS2005), pp. 1338-1343(2005).
12. Gross, H.-M., Boehme, H.-J., Schroeter, C., Mueller, S., Koenig, A., Martin, C., Merten, M. and Bley, A., ShopBot: Progress in Developing an Interactive Mobile Shopping Assistant for Everyday Use, IEEE Int. Conf. on Systems, Man, and Cybernetics , pp. 3471-3478(2008).
13. J. Jaffe and S. Feldstein, Rhythms of dialogue. New York: Academic Press (1970)
14. Jung, J., Kanda, T., and Kim, M.-S., Guidelines for Contextual Motion Design of a Humanoid Robot, International Journal of Social Robotics, Vol.5, Issue 2, pp.153-169, (2013).
15. Kahn, P., Freier, N., Kandau, T., Ishiguro, H., Ruckert, J., Severson, R., and Kane, S., Design Patterns for Sociality in Human-Robot Interaction, ACM/IEEE Int. Conf. on Human-Robot Interaction, pp. 97-104 (2008).
16. Kanda, T., Ishiguro, H., Imai, M., and Ono, T., Development and evaluation of interactive humanoid robots. Proceedings of the IEEE, 92(11), 1839-1850 (2004).
17. Kanda, T., Shiomi, M., Miyashita, Z., Ishiguro, H. and Hagita, N., An Affective Guide Robot in a Shopping Mall, ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI2009), pp. 173-180(2009).
18. Kanda, T., Shiomi, M., Miyashita, Z., Ishiguro, H. and Hagita, N., A Communication Robot in a Shopping Mall, IEEE Transactions on Robotics, vol.26, pp.897-913(2010).
19. Kawai, H., Toda, T., Ni, J., Tsuzaki, M. and Tokuda, K., XIMERA: A new TTS from ATR based on corpus-based technologies, ISCA Speech Synthesis Workshop, pp. 179-184(2004).
20. Krenn, B. and Sieber, G., Functional Mark-up for Behaviour Planning: Theory and Practice, AAMAS 2008 Workshop FML: Functional Markup Language. Why Conversational Agents do what they do(2008).
21. Kuo, I.-H., Jayawardena, C., Broadbent, E., and MacDonald, B. A., Multidisciplinary Design Approach for Implementation of Interactive Services: Communication Initiation and User Identification for Healthcare Service Robots, International Journal of Social Robotics, vol. 3, pp. 443-456, (2011).
22. Lohse, M. and Siepmann, F., A Modeling Framework for User-Driven Iterative Design of Autonomous Systems, International Journal of Social Robotics, 6(1), pp.121-139 (2014).
23. McTear, M. F., Modelling spoken dialogues with state transition diagrams: experiences with the CSLU toolkit, International Conference on Spoken Language Processing (ICSLP1998), pp. 1223-1226(1998).
24. McTear, M. F., Spoken Dialogue Technology: Enabling the Conversational User Interface, ACM Computing Surveys (CSUR), vol. 34, pp. 90-169(2002).
25. Miller, G. A., The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological review, 63(2), 81 (1956).
26. Nagaoka, C., Komori, M., and Nakamura, T., Influence of Response Latencies on Impression Evaluation of Speakers in Dialogues: differences of cues used for evaluation by degree of social skill. Tech Rep IEICE, 104(745), 57-60 (2005).
27. Nishimura, Y., Minotsu, S., Dohi, H., Ishizuka, M., Nakano, M., Funakoshi, K., Takeuchi, J., Hasegawa, Y. and Tsujino, H., A Markup Language for Describing Interactive Humanoid Robot Presentations, International Conference on Intelligent User Interfaces (IUI 2007), pp. 333-336(2007).
28. Petelin, J. B., Nelson, M. E. and Goodman, J., Deployment and early experience with remote-presence patient care in a community hospital, Surgical Endoscopy, vol. 21, pp. 53-56(2007).
29. Pineau, J., Montemerlo, M., Pollack, M., Roy, N. and Thrun, S., Towards robotic assistants in nursing homes: challenges and results, Robotics and Autonomous Systems, vol. 42, pp. 271-281(2003).
30. Polanyi, M., Personal Knowledge: Towards a Post-Critical Philosophy. University of Chicago Press, Chicago(1962).
31. Ross, D., Lim, J., Lin, R.-S., and Yang, M.-H., Incremental Learning for Robust Visual Tracking. International Journal of Computer Vision, 77(1-3), 125-141 (2008).
32. H. Sacks, E. A. Schegloff, and G. Jefferson, A simplest systematics for the organization of turn-taking for conversation, Language, pp. 696-735 (1974).
33. Shiomi, M., Kanda, T., Glas, D. F., Satake, S., Ishiguro, H. and Hagita, N., Field Trial of Networked Social Robots in a Shopping Mall, IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS2009), pp. 2846-2853(2009).
34. Shiwa, T., Kanda, T., Imai, M., Ishiguro, H. and Hagita, N., How Quickly Should Communication Robots Respond?, ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI2008), pp. 153-160(2008).
35. Takayama, L., Marder-Eppstein, E., Harris, H. and Beer, J. M., Assisted Driving of a Mobile Remote Presence System: System Design and Controlled User Evaluation, IEEE Int. Conf. on Robotics and Automation (ICRA2011), pp. 1883-1889(2011).
36. Tasha K. Hollingsed and Nigel G. Ward, A Combined Method for Discovering Short-Term Affect-Based Response Rules for Spoken Tutorial Dialog, ISCA ITRW Speech and Language Technology in Education(SLaTE)(2007).

This is a preprint manuscript.
The final publication is available at Springer via http://dx.doi.org/10.1007/s12369-015-0288-9

37. Tsui, K. M., Desai, M., Yanco, H. A. and Uhlik, C., Exploring Use Cases for Telepresence Robots, ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI2011)(2011).
38. Villano, M., Crowell, C. R., Wier, K., Tang, K., Thomas, B., Shea, N., Schmitt, L. M., Diehl, J. J., DOMER: a wizard of oz interface for using interactive robots to scaffold social skills for children with autism spectrum disorders, ACM/IEEE 6th international conference on Human-robot interaction, Lausanne, Switzerland (2011).
39. Wallis, P., Mitchard, H., Das, J., and ODea, D., Dialogue modelling for a conversational agent AI 2001: Advances in Artificial Intelligence (pp. 532-544): Springer (2001).
40. Ward, N., and Tsukahara, W., Prosodic features which cue back-channel responses in English and Japanese. Journal of pragmatics, 32(8), 1177-1207 (2000).
41. Ward, Nigel, Anais G. Rivera, Karen Ward, David G.Novick, Some Usability Issues and Research Priorities in Spoken Dialog Applications, Technical Report UTEP-CS-05-23(2005).
42. Weiss, A., Igelsbock, J., Tscheligi, M., Bauer, A., Kuhnlenz, K., Wollherr, D. and Buss, M., Robots Asking for Directions: The Willingness of Passers-by to Support Robots, ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI2010), pp. 23-30(2010).
43. Yun, S.-S., Kim, M., and Choi, M.-T., Easy Interface and Control of Tele-education Robots, International Journal of Social Robotics, vol. 5, pp. 335-343, (2013).
44. P. Zang, R. Tian, A. L. Thomaz, and C. L. Isbell, Batch versus interactive learning by demonstration, Development and Learning (ICDL), IEEE 9th International Conference on, pp. 219-224 (2010).
45. Zheng, K., Glas, D. F., Kanda, T., Ishiguro, H. and Hagita, N., How Many Social Robots Can One Operator Control?, in ACM/IEEE 6th Annual Conference on Human-Robot Interaction , ed. Lausanne, Switzerland., pp. 379-386(2011).

Kanae Wada received her master's degree in engineering from Osaka University in 2012 and is currently working at 3D MEDiA Co., Ltd. From 2010-2012, she was an intern researcher at the Intelligent Robotics and Communication Laboratories (IRC) at the Advanced Telecommunications Research Institute International (ATR) in Kyoto, Japan. Her research interests include networked robots, human-robot interaction, and field experimentation.

Dylan F. Glas received his Ph.D. in Robotics from Osaka University in 2013. He received his M.Eng in Aerospace Engineering from MIT in 2000 and two S.B. degrees, one in Aerospace Engineering and one in Earth, Atmospheric, and Planetary Sciences, also from MIT in 1997. From 1998-2000 he was a member of the Tangible Media Group at the MIT Media Lab. He is currently group leader of the Department of Cloud Intelligence at the Intelligent Robotics and Communication Laboratories (IRC), Advanced Telecommunications Research Institute International (ATR) in Kyoto, Japan. His research interests include social human-robot interaction, social behavior design, cloud robotics, ubiquitous sensing, teleoperation, and machine learning.

Masahiro Shiomi received M. Eng. and Ph.D. degrees in engineering from Osaka University in 2004 and 2007. From 2004 to 2007, he was an intern researcher at the Intelligent Robotics and Communication Laboratories (IRC). He is currently a group leader in the Agent Interaction Design department at IRC, Advanced Telecommunications Research Institute International (ATR). His research interests include human-robot interaction, robotics for child-care, networked robots, and field trials.

Takayuki Kanda received the B. Eng., M. Eng., and Ph.D. degrees in computer science from Kyoto University, Kyoto, Japan, in 1998, 2000, and 2003, respectively. From 2000 to 2003, he was an Intern Researcher with the Advanced Telecommunications Research Institute International (ATR) Media Information Science Laboratories, Kyoto. He is currently a Senior Researcher at ATR Intelligent Robotics and Communication Laboratories. His research interests include intelligent robotics, human-robot interaction, and vision-based mobile robots. Dr. Kanda is a Member of the Association for Computing Machinery, the Robotics Society of Japan, the Information Processing Society of Japan, and the Japanese Society for Artificial Intelligence.

Norihiro Hagita received the B.E., M.E., and Ph.D. degrees in electrical engineering from Keio University in 1976, 1978, and 1986. In 1978, he joined Nippon Telegraph and Telephone Public Corporation (Now NTT). He was a visiting researcher in the Department of Psychology, University of California, Berkeley in 1989-90. He is currently Board Director of ATR and ATR Fellow, director of the Social Media Research Laboratory Group and the Intelligent Robotics and Communication Laboratories. He is the chairman of ATR Creative. He is also a visiting professor of Nara Institute of Science and Technology, Osaka University, and Kobe University. His major interests are cloud networked robotics, human-robot interaction, ambient intelligence, pattern recognition and learning, and data-mining technology. He has served as a chairman of technical committee in Network Robot Forum in Japan.

This is a preprint manuscript.
The final publication is available at Springer via http://dx.doi.org/10.1007/s12369-015-0288-9