Interactive Problem-Solving with Humanoid Robots and Non-Expert Users

Duygun Erol Barkana[1] [a], Mattias Wahde[2] [b] and Minerva Suvanto[2] [c]

[1] *Yeditepe University, Kayışdağı, Ataşehir, İstanbul, Turkey*

[2] *Chalmers University of Technology, SE-412 96 Gothenburg, Sweden*

duygunerol@yeditepe.edu.tr, {mattias.wahde, minerva.suvanto}@chalmers.se

Keywords: Interpretable Models, Conversational AI, Human-Robot Interaction, NAO Robot.

Abstract: We study conversational interaction between a humanoid (NAO) robot and a human user, using a glass-box dialogue manager (DAISY). The aim is to investigate how such an interaction can be organized in order for a non-expert user to be able to interact with the robot in a meaningful way, in this case solving a scheduling task. We compare two kinds of experiments, those that involve the NAO robot and the dialogue manager, and those that involve only the dialogue manager. Our findings show a rather clear preference for the setup involving the robot. Moreover, we study the level of linguistic variability in task-oriented human-machine interaction, and find that, at least in the case considered here, most dialogues can be handled well using a small number of patterns (for matching user input) in the agent.

1 INTRODUCTION

Robots are taking an active role in many fields, such as education, health care, and assistance (Youssef et al., 2022). Humanoid robots are among the most dominant categories in spoken human-robot interaction (Reimann et al., 2024). These robots are designed specifically to interact with human users (Breazeal et al., 2016), for example by using spoken dialogue. The interaction of humans with social robots plays a crucial role (Baraka et al., 2020; Fong et al., 2003) in the context of elderly care (Broekens et al., 2009) and education (Belpaeme et al., 2018). Different robotic platforms have been used with varying interaction modalities for human-robot interaction (Youssef et al., 2022; Reimann et al., 2024) bearing in mind that spoken dialogue systems are not restricted to specific robots with certain shapes or features. The small, commercially available humanoid robot NAO is preferred in many studies due to its affordability compared to other robots with similar functionality, its wide range of capabilities, and its ability to act as a platform for evaluating human-robot interaction (Amirova et al., 2021; Pino et al., 2020).

However, in most cases, robots are demonstrators rather than usable tools. Going from a carefully orchestrated demonstration to a fully operational and reliable robotic system is a daunting task for several reasons, including issues related to reliability and safety, as well as the ability of the robot to handle new and unexpected situations. Another reason is that the intended end users - for instance, medical doctors or teachers - are rarely robotics experts. Instead, to be useful from their point of view, the robot should seamlessly be integrated into the daily work activities of its users. Those activities may involve a need to tune the robot's behaviors, or to add new behaviors, both of which are very hard to achieve for a non-expert. In this paper, we will describe a method for conversational interaction between a non-expert user and a small humanoid robot (NAO). In other words, we will illustrate how a non-expert user, with our method, can carry out meaningful interactions with the robot to complete the task at hand.

At present, studies on human-machine verbal interaction are dominated by the use of large language models (LLMs), popularized in chatbots such as ChatGPT. While such systems have indeed had a massive impact on the development of dialogue capabilities in artificial systems, they also suffer from several drawbacks (Wahde, 2024), one such drawback being their lack of interpretability: LLMs are essentially black boxes that struggle with accountability, reliability, and safety. Moreover, they are prone to

[a] https://orcid.org/0000-0002-8929-0459
[b] https://orcid.org/0000-0001-6679-637X
[c] https://orcid.org/0009-0003-1751-151X

Erol Barkana, D., Wahde, M. and Suvanto, M.
Interactive Problem-Solving with Humanoid Robots and Non-Expert Users.
DOI: 10.5220/0013369000003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 1*, pages 870-879
ISBN: 978-989-758-737-5; ISSN: 2184-433X

what is sometimes referred to as hallucinations, i.e., a propensity to make up answers that are factually incorrect. This problem, in particular, limits their usefulness in high-stakes applications (Rudin, 2019) such as, for example, in healthcare.

In this paper, as the primary task, we will study the specific problem where a doctor wishes to generate a daily schedule for a patient with the help of an interactive system. Instead of an LLM-based system, we will use our precise dialogue manager DAISY (Wahde and Virgolin, 2023) that, unlike the currently popular LLM-based chatbots, does not hallucinate. On the other hand, DAISY's conversational capabilities are much more limited than those of an LLM-based chatbot. Hence, as a secondary task, we will investigate to what degree a dialogue manager of this kind is capable of handling the variability (as induced by the human users) of dialogue that occurs naturally in human-machine interactions, even in cases such as ours, where the task is rather narrowly defined.

2 INTERACTIVE SYSTEM PLATFORM

The overall system architecture is shown in Figure 1. As can be seen in the figure, a human operator gives input to NAO in the form of speech. The recorded speech is then transferred to an external computer (connected to NAO through a wireless network) that carries out speech recognition, i.e., converting the speech recording to text. Next, the text is passed to DAISY, which then carries out its processing in three steps: Language understanding (input matching), cognitive processing (thinking, of sorts), and, finally, response generation. The (textual) response and a suitable animation (movement) sequence are then passed to NAO. The robot then carries out the animation (if any) while, at the same time, presenting its output in the form of speech.

The following subsections provide a more detailed description of the robot, the dialogue manager, and their communication.

2.1 Robot: NAO

The NAO robot is 0.58 m tall and weighs 4.5 kg. The version used in this study is NAO v.6. The robot has a Python SDK (Naoqi) available, allowing researchers to easily control the hardware and to design robot behaviors involving motion processing, speech, and vision. NAO has support with its application programming interface (API) for natural user interaction.

The onboard sensors include an inertial measurement unit (IMU), touch and force sensors, and sonars. The force-sensitive sensors are on the hands and feet and are used for detecting contact with objects. Light-emitting diodes (LEDs) are located on the eyes and the body. There are four microphones to identify the source of sounds and two loudspeakers for communication. NAO has two hands with self-adapting gripping abilities, but a single engine controls the three fingers of each hand, so they cannot be moved independently. It has 25 degrees of freedom in the joints, allowing independent movement of the head, shoulders, elbows, wrists, waist, legs, knees, and ankles.

The robot also has two built-in cameras, one at the mouth level and the other on its forehead. Both are 920p cameras able to run at 30 images per second with (up to) 1280×720-pixel images. NAO's head can move 68 degrees vertically and 239 degrees horizontally. NAO can see 61° horizontally and 47° vertically with its cameras located in the forehead. Hence, it has a good overview of its environment. Its forehead camera records videos of the person in front of NAO during the interaction.

The NAO robot has been used in social robotics research as a tutor, a therapist, and a peer learner (Amirova et al., 2021; Namlisesli et al., 2024). In educational settings it has been used with children (Tanaka and Matsuzoe, 2012; Lemaignan et al., 2016) and university students (Banaeian and Gilanlioglu, 2021). As a therapist or assistant in rehabilitation, research with NAO has been done with patients that have cognitive impairments (Pino et al., 2020), as well as physically impaired patients (Pulido et al., 2017). In this study, the NAO robot will transmit what the user says to the DAISY dialogue manager and verbally report the information from DAISY back to the user.

2.2 Dialogue Management

Currently, most research in conversational AI is devoted to chatbots based on large language models (LLMs), such as, for example, GPT-4 (OpenAI, 2023), LLAMA-2 (Touvron et al., 2023), and so on. These models are highly capable in the sense that they can and will respond to any questions posed to them (except for certain questions that, for example, involve offensive language). Such models generate responses probabilistically, token by token. While this process often results in sensible answers, LLM-based chatbots also quite often (and unpredictably) generate output that is factually incorrect or even completely nonsensical, sometimes simply making things up (a phenomenon sometimes called *hallu-*

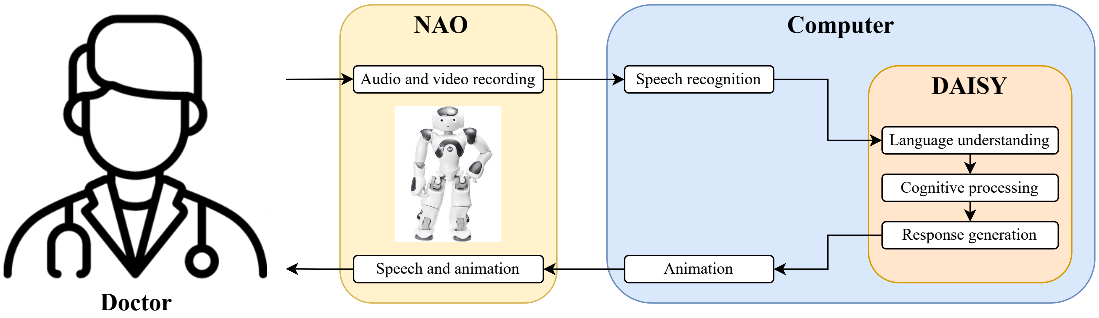


Figure 1: System Architecture.

cination, although one should be careful to avoid drawing detailed parallels with hallucination in humans). This is a persistent problem, despite various measures that have been taken to counteract it, such as fine-tuning and using retrieval-augmented generation (Lewis et al., 2020) or step-by-step reasoning. Moreover, sometimes chatbots may even respond in a strongly aggressive manner (Si et al., 2022), a highly undesirable feature in an AI system operating in, say, a healthcare setting or a teaching situation. As mentioned above, LLMs are, to all intents and purposes, black-box models whose inner workings are not accessible to a human observer. Moreover they are trained over huge data sets that may (and often do) contain offensive language and unwanted biases.

Taken together, these aspects of LLMs (and, indeed, black-box models in general) imply that such models are not suitable in high-stakes situations (Rudin, 2019) such as, for example, in mental health care (Coghlan et al., 2023) and elderly care: A model that lacks transparency also lacks safety and accountability.

It is precisely for those situations that the DAISY dialogue manager (Wahde and Virgolin, 2023) has been developed. This dialogue manager allows for a very precise, reliable, accountable, and fully transparent interaction between a human user and a computer. Moreover, DAISY has a built-in procedure for providing correct-by-construction explanations of its reasoning[1].

On the negative side, DAISY's capabilities are generally limited to the task at hand and are, at present, encoded by hand (a procedure for allowing automated learning is under development, however). Furthermore, unlike LLM-based chatbots, DAISY is neither probabilistic nor generative. Thus, while DAISY is by no means a direct competitor to LLM-based chatbots over their full range of applicability, in its present state the model is useful in clearly defined tasks, where precision and reliability are more important than flexibility and human likeness.

A detailed description of DAISY can be found in (Wahde and Virgolin, 2023). Here, only a brief overview will be given. It should be noted that, recently, DAISY has undergone important modifications, such that (i) its declarative memory (facts) are now instead stored as a knowledge graph (KG) and (ii) it includes another KG that contains linguistic and grammatical information, such as explicit similarity relationships between words and phrases. The basic flow of information in DAISY is shown in Figure 2.

The input to DAISY is in the form of a text string, obtained either by typing on a keyboard or via speech recognition (that, in itself, is not a part of DAISY). The text string is then preprocessed, by carrying out the steps of tokenization, spelling correction, spelling normalization (for example, converting between British and American spelling), and punctuation handling (splitting text into sentences, if needed). Then, the text is passed through a linguistic KG that handles paraphrasing, meaning that (when possible) it generates alternative texts that are semantically identical to the given input text.

After that, the text (with all the variants just mentioned) is passed to the procedural memory, consisting of three different types of entities: *Input items* that match the input text in order to determine what processing needs to be carried out, *cognitive items* that do the actual processing, retrieving relevant data (if any) from working memory and (long-term) declarative memory, and *output items* that formulate the output text, using (when needed) variables that were generated in the cognitive processing step.

The input matching is based on patterns with varying levels of complexity: The simplest possible pattern specification consists of a single string such as *I want to make a schedule for a patient*. However, since the user input is matched *exactly* against the stored patterns, typically the patterns must be quite a bit more dynamic to handle the natural variability in user

[1]It should be noted that, if prompted, an LLM-based chatbot will also provide an explanation, but there is no guarantee that the explanation (or the original statement, for that matter) is correct.

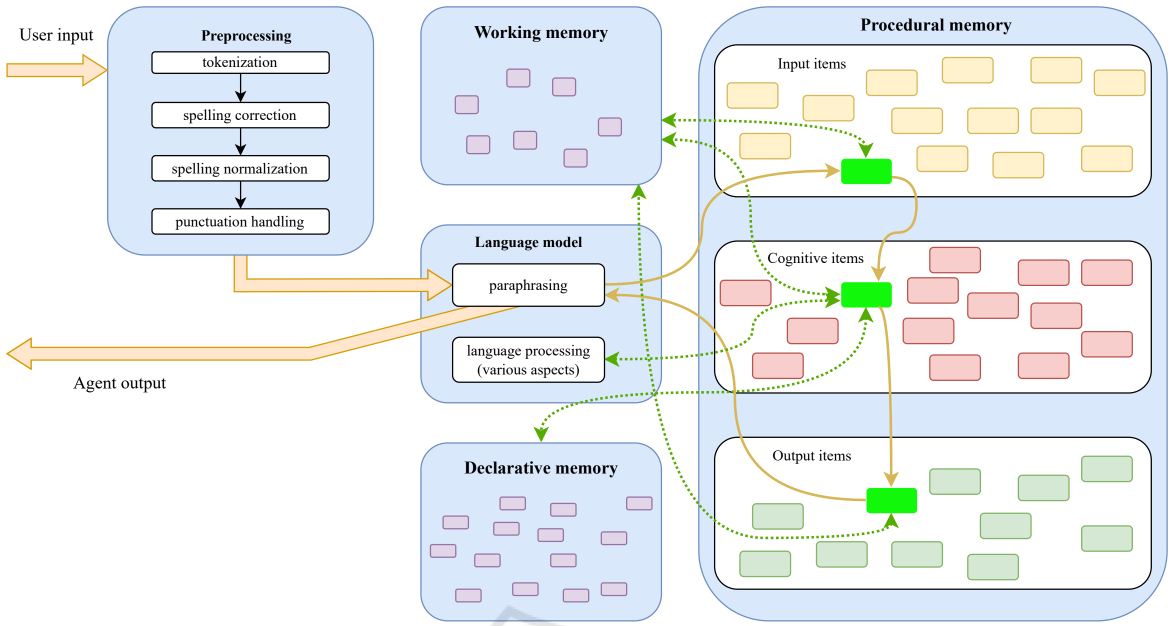


Figure 2: A schematic view of the DAISY dialogue manager.

input. A more sophisticated pattern for the case just mentioned could take the form *I (WANT) to (MAKE) a (daily;-), (PLAN) for a patient*. Here, the specification *(daily;-)* means that the word *daily* is optional. The capitalized specifications, e.g. *(MAKE)*, represent semantic grammars (SGs), i.e., a set of words or phrases with identical meaning. For example, in the case of *(MAKE)*, the semantic grammar is defined as *(MAKE) = (make; create; generate; establish; set; organize; arrange; define; plan)*, and so on. Thus, the sample pattern just described would be matched by user inputs such as *I want to make a plan for a patient*, *I want to generate a schedule for a patient*, *I wish to define a daily plan for a patient*, and so on. Note that patterns can also define variables that are placed in working memory. For example, the pattern *I want to make a schedule for (TITLE) $name*, where the SG *(TITLE)* is defined as *(TITLE) = (Mr.,Mrs.,Ms.)*, would be matched by, say, *I want to make a schedule for Mr. Duck*, in which case the variable *$name* (in working memory) would hold the name *Duck*, making it available during cognitive processing (which may, for example, involve extracting a room number for the patient in question; see also the dialogue examples below).

After input matching, cognitive processing and output generation ensues, during which the agent carries out a sequence of steps necessary for determining which output to return to the user. This procedure will not be described here, but the interested reader can find a thorough description in (Wahde and Virgolin, 2023). The generated output text is then again passed through the linguistic KG, where it can be paraphrased to allow for a more human-like output, rather than just robotically answering similar questions in exactly the same way.

Specifying a DAISY agent for a particular task involves defining input, cognitive, and output items using a scripting language, a process that is further described in (Wahde and Virgolin, 2023). The specification of input items, which is our focus here, involves the definition of a set of patterns that together handle the natural variation in the human-agent interaction. This would be a near-impossible task for general interaction over an unlimited range of topics (where instead LLMs excel). However, in a task-oriented setting, where the user is trying to achieve a specific outcome, it may be possible to provide a sufficiently varied set of patterns with a reasonable effort. Determining whether this is actually the case is indeed one of the tasks in this paper, as outlined at the end of Section 1. The analysis related to this issue is presented in Section 4 below.

2.3 NAO-DAISY Interaction

The user can interact with the DAISY dialogue manager either directly (via textual input on a keyboard) or via NAO (using speech). A client-server configuration is used, where the computer (running DAISY) is the server and NAO is the client.

Robots that use speech as a primary mode of interaction need to understand the user's utterances and select appropriate responses given the context: A social robot with conversation ability acquires sound signals, processes these signals to recognize sequences of words in the user's input, and then formulates an appropriate reply (using DAISY, in our case). The robot synthesizes the sound signal corresponding to the reply, and then emits this signal using a loudspeaker.

The external computer (See Figure 1) runs a program, here called the *main script*, which is responsible for handling the interaction between the DAISY dialogue manager (that also runs on the same computer) and the NAO robot. Initially, the main script sends a command to NAO to calibrate its microphones. The calibration process includes 4 seconds of sampling the energy received by each microphone at 10Hz, and the baseline is extracted by taking the average of this value. Later on, the baseline is used for determining whether a person has completed an utterance.

When the calibration is completed, NAO greets the user with predetermined greeting phrases and speaking animations. Then, NAO starts listening to the participant and simultaneously recording the participant's audio and video. The duration of the recordings is dynamic and based on the average microphone energy of the environment, as previously mentioned. Considering the computation time for emotion recognition, the video frame rate can also be changed. When the participant has completed an utterance, the audio and video recordings are saved onto the NAO computer. The main script in the computer running DAISY then requests the saved audio and video using a secure copy protocol (SCP). After the computer has retrieved the file, noise reduction is carried out on the audio file to enhance the speech signal (Sainburg, 2019); (Sainburg et al., 2020). Next, the participant's speech is converted into text using the speech-to-text model Whisper (Radford et al., 2022). The text is fed into the DAISY dialogue manager, which then (after processing) returns a response. NAO reads the message from DAISY to the user using the Naoqi library, which includes a text-to-speech function.

Interactions with robots are not purely speech-based but can also use multimodal cues like gestures, gaze, or facial expressions. Thus, NAO uses a related tag based on the conversation stage. Tags are groups of predetermined animations on NAO that indicate certain gestures for specific circumstances, such as comparison, confirmation, disappointment, etc. The activity-specific behaviors of NAO used during the interaction are given in Figure 3. The communication between the user and DAISY via the NAO robot continues until the task at hand has been completed.

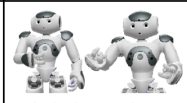
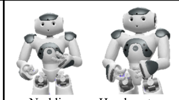
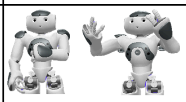
Stage	Non-verbal	Animated Say tags	Verbal
Greeting	 Waving hands, Greeting user	Self, Joy	Hello, Hi, Greetings
Confirmation	 Nodding, Hand gestures	Confirmation, Affirmative	Understood, Confirmed
Activities	 Exercise movements	Custom animations	Breakfast, Shower, Rest, Wake up, Physical exercises, Sleep
Appreciation	 Thanking gestures	Joy, Global	Thank you

Figure 3: Activity-specific gestures and behaviors.

3 EXPERIMENTS USING THE DIALOGUE SYSTEM

The primary task of the study involved creating a daily schedule for an elderly person, by a person who is a non-expert using the NAO robot (and the dialogue manager). The participants were asked to complete an experiment of this kind for two different configurations: (1) With DAISY only and (2) with DAISY and the NAO robot.

Five different dialogues were generated. Each dialogue includes greeting, making a plan (one action for morning, one action for noon, one action for afternoon, and one action for evening) and, finally, saying goodbye. Every participant carried out one randomly assigned dialogue. An example dialogue is presented below where P is the participant, and R is the NAO robot.

- P: Hello.
- R: Hello doctor.
- P: I want to make a daily schedule for a patient.
- R: Yes certainly what is the name of the patient?
- P: Mister Duck.
- R: Ok schedule for Mister Duck in room 101 please provide the activities one by one.
- P: Wake-up at 6:00.
- R: Ok wake-up will take place at 6:00.
- P: Shower at 6:15.
- R: Ok shower at 6:15.

- P: He should have breakfast at 7.
- R: Understood changed the time for breakfast from 7:30 to 7:00.
- P: After that he should have some light exercise at 8.
- R: Ok exercise will take place at 8.
- P: Nap at 10:00.
- R: Noted, nap will take place at 10:00.
- P: Lunch at 12:30.
- R: Ok lunch at 12.30.
- P: Then walk at 13.
- R: Confirmed walk is scheduled for 13.
- P: Dinner at 18:00.
- R: Confirmed dinner will take place at 18:00.
- P: Then he should go to bed at 21:00.
- R: Noted bedtime is scheduled for 21:00.
- P: Please repeat his entire schedule.
- R:... (Repeats the entire schedule)
- P: Done.
- R: Ok noted.

Note that, before the interaction starts, the robot has been given information about the names of patients and the room number in which they reside. It should also be noted that, in this example, a default time for breakfast was set, at 07:30, but it is changed by the user to 07:00, as can be seen in the dialogue above. All the other events lacked a default value and were thus added through the interaction above.

25 participants who were not included in the initial dialogue generation participated in the experiments. The participants were university students aged between 21 and 33 years. A cross-over evaluation was done, meaning that half the attendees had initially communicated only with DAISY and generated their schedule (Figure 4). Then, they were asked to perform the exact same conversation with NAO and DAISY together (Figure 5). The other half of the attendees had this conversation with NAO first and then with DAISY only.

In the literature, different subjective and objective methods have been used to evaluate the performance of dialogue managers. Detailed surveys can be found in (Deriu et al., 2021; Reimann et al., 2024). User studies have the advantage of allowing a review of the performance of the dialogue manager and the robot in a real interaction. Users can also provide their opinions in questionnaires like the Godspeed questionnaire (Bartneck et al., 2009). In that way, user frustration and other subjective measurements can also be

Figure 4: Experiment with DAISY.

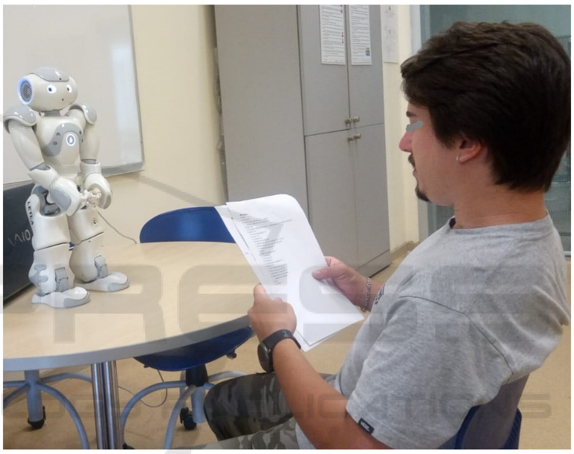


Figure 5: Experiment with NAO.

included in the evaluation. The participant numbers reported for user studies differ significantly, ranging from 2 to 97, with an average of 21 participants (SD 18.4) (Reimann et al., 2024). Some papers report the number of turns (Campos et al., 2018; Chai et al., 2014), while some focus on the length of the turns (Dino et al., 2019) or the total amount of time (Johansson and Skantze, 2015). In this study, the participants were asked to provide an evaluation using the System usability scale (SUS; see below) at the end of the experiment with NAO and DAISY. SUS has been widely used to evaluate subjective usability (Lewis, 2012) and has been shown to detect differences at smaller sample sizes than other questionnaires (Brooke, 2013).

SUS is a 10-item questionnaire (each question with a Likert scale ranging from strongly agree to strongly disagree) designed to measure the usability of a system (Brooke, 1996). The possible ratings were Strongly Disagree (1), Disagree (2), Neutral (3), Agree (4), and Strongly Agree (5). SUS is formulated as a balanced survey consisting of 5 questions with positive statements and 5 with negative statements, with scores ranging from 0 to 100. The process for computing a SUS score is as follows: i) subtract 1 from the user's Likert ratings for odd-numbered items or questions, ii) subtract the user's Likert ratings from 5 for even-numbered items, so that each item score will range from 0 to 4, and iii) sum the numbers and multiply the total by 2.5, so that the calculation will provide a range of possible SUS scores from 0 to 100.

4 RESULTS AND DISCUSSION

Starting with the primary task (generating a schedule; see Section 3), the average completion time for experiments involving DAISY only was 6.3 minutes, whereas, for the case involving DAISY with NAO it was 8.8 minutes. The longer duration of the experiments involving NAO can be explained by the fact that it takes some time for NAO to complete its animations. Furthermore, speech recognition in human-robot interaction may be impacted by noise from the robot's motors and the environment in which the robot is placed. The age, accent, and way of speaking can vary. All of those factors can affect the performance of the speech recognition module and may cause some delay.

The SUS scores were 73.1 (for NAO + DAISY) and 68.8 (DAISY only), indicating a higher usability for the combined system. In the literature, a threshold of around 68 is typically used (Sauro and Lewis, 2016), putting the combined system comfortably above the threshold, whereas DAISY alone is essentially at the threshold.

The gender and age of the participants may also play a role in the responses. As opposed to male users, female users have been observed to prefer some non-verbal behaviors in embodied agents (Krämer et al., 2010). In this case, however, we found that there was no big difference between female and male participants (9 and 16 participants, respectively), with an average SUS score of 72.2 (NAO) and 68.8 (DAISY) for female respondents, and 73.6 (NAO) and 68.9 (DAISY) for male respondents. However, inspecting the average SUS score of two different age groups (16 participants aged 21-24, 9 participants aged 25-33) showed that the older age group (25-33) had a larger preference for the robot, average scores being 76.4 (NAO) and 66.7 (DAISY), while the scores of the younger age group (21-24) did not exhibit a big difference, 71.2 (NAO) and 70.0 (DAISY).

The participants were also asked if they preferred to complete the task with only DAISY or with NAO and DAISY combined. Additionally, they were asked to provide reasons for their preferences. 18 participants (out of 25) preferred NAO. The robot was preferred because the participants found interaction with a robot enjoyable and exciting. Furthermore, some participants mentioned that they do not like writing and thus appreciated that they were able to complete the task just by talking. Some participants said that it was enjoyable to talk to a robot. One participant found the NAO robot to be cute and one participant found NAO easy to communicate with. On the other hand, seven participants preferred DAISY. These participants mostly did not like interacting via speech and were frustrated when NAO did not understand their speech. The participants were not happy to repeat their statements a couple of times.

While the results of our investigation indicate a preference for the setup involving the NAO robot, it is important not to draw too far-reaching conclusions based on a fairly small pool of participants solving a single task. The results obtained are to be seen as a first indication, rather than a definitive answer. An obvious step for future work would be to expand the number of participants and to study user preferences in other tasks as well. Another possible avenue for further work would be to compare user preferences between, on the one hand, the DAISY and NAO setup and, on the other hand, a setup involving DAISY combined with an on-screen 3D-animated face (rather than just the textual interface).

Regarding the secondary task mentioned in Sections 1 and 2.2, the aim was to see how much variation there is in a task-oriented setting, such as the scenario presented in this paper. As mentioned earlier, in DAISY, the user's input is matched against a set of predefined patterns that redirects the system to do cognitive processing. By determining the degree of variability, one can estimate the number of patterns needed in order to handle the natural variations in the user's inputs.

For this task, 24 participants (different from the participants in the NAO experiments) were asked to manually create five variations of the dialogue scenario (described in Section 3). From these dialogues, each individual statement was inspected manually. That is, the investigation involved statements that describe the same action but expressed with different words. For example, the sentences *I'm making a schedule for Buzz* and *I am currently arranging Buzz's timetable* use mostly different words to express the same intent. In measuring the similarity between statements, one can ignore variables such as, for instance, *Buzz* in the previous example, which can be replaced by another individual and still maintain the same meaning from the perspective of DAISY. Differences in lexical choices, such as the choice of word for an action (e.g., *awaken* as *wake up, rise, get up, etc.*), can be handled by the linguistic KG in DAISY; see Section 2.2, assuming that all the relevant entries in this KG have been defined. Actions may also include placeholders to handle subjects within a phrasal verb, for example, *wake* ***her*** *up*. Statements involving corrections or multiple answers are excluded from this evaluation.

Statements can contain utterances that are not useful from a task-oriented perspective (*Oh, I almost forgot to tell you, ...* or, *She prefers to sleep a little longer, so ...*). One challenge for DAISY is to detect and ignore these parts of the statement completely. Another challenge may be ambiguous statements, where it is not clear what the speaker means, such as the statement *he needs to start the day quite early, 7:15*, where it is unclear whether the given time refers to wake-up or some other undefined first task of the day. For some statements, the dialogue manager needs to do some additional cognitive processing to determine the correct action, for example, in the sentence *he should eat at 12*, the system needs to set *lunch* in the schedule, even if it is not explicitly stated. Additionally, statements where the action or time is referred to with a pronoun may pose challenges (e.g., *At 12:00 he should be getting hungry so serve lunch at* ***that time***).

While sentences of the kind just described exhibit quite a bit of variation, we can conclude that, for the remaining sentences (that form a majority), a single pattern can indeed often cover most of the variation in user statements. For example, for the statement that sets the wake-up time, most participants appear to use the structure of subject-verb-prepositional phrase (*he wakes up at 6*) in their dialogues, where the verb specifies the action (*wake up*) and the prepositional phrase describes the time (*at 6*). Variations in this structure include, for example, dropping the subject (*wake up at 6*), using auxiliaries (*he will wake up at 6*), or adding adverbial phrases (*he usually wakes up at 6*). Still, these variations can be handled by DAISY using a single pattern with optional structures included, as described in Section 2.2.

Some variation requires more than just optional parameters within the pattern, though. In certain cases, the common structure of the statements is modified by moving the prepositional phrase before the verb (*at 6 he wakes up*). This can be either handled by adding a new pattern (essentially a modified version of the previously mentioned pattern) or modifying the existing pattern by allowing a flexible position of the prepositional phrase within the pattern. In addition, three more patterns were identified for this statement, to handle (1) a structure with the action as the subject and *is* as the verb (*wake up is at 6*), (2) using imperative requests (*set wake up at 6*), and (3) the action specified in a subordinate clause (*let's start by waking her up at 6:00*). In all of these cases, patterns can be defined, as discussed above, with optional slots to handle the variability in the user's statements.

The same patterns for one statement can be reused for other statements. The most common pattern for each statement varies; for example, for the statement *lunch is at 12*, the most common pattern is still the structure of subject-verb-prepositional phrase, but with the action (in this case lunch) as the subject instead of the verb of the sentence.

5 CONCLUSIONS

We have carried out user studies involving either a dialogue manager (DAISY) alone, or a NAO robot operating together with the dialogue manager, applied in a task that involved the definition of a daily schedule for a patient. The main conclusion is that most participants have a preference for the setup that involves the robot, possibly due to its pleasant visual appearance and the fact that the interaction between the robot and a user is verbal rather than textual.

We also studied the linguistic variability of human-machine dialogue in the task described above. Our findings show that, for the most part, the dialogue can be handled with a rather limited set of patterns. Thus, while the specification of patterns would be a prohibitively complex task for general, unconstrained human-robot interaction, we conclude that it is manageable for task-oriented dialogue where a single pattern (that allows for some variability) or a few such patterns suffice to handle most user inputs. For future work, this conclusion paves the way for applying our DAISY dialogue manager in other task-oriented settings.

ACKNOWLEDGEMENTS

This research was funded in part by the GENIE programme at Chalmers University of Technology.

REFERENCES

Amirova, A., Rakhymbayeva, N., Yadollahi, E., Sandygulova, A., and Johal, W. (2021). 10 years of human-nao interaction research: A scoping review. *Frontiers In Robotics And AI*, 8:744526.

Banaeian, H. and Gilanlioglu, I. (2021). Influence of the nao robot as a teaching assistant on university students' vocabulary learning and attitudes. *Australasian Journal of Educational Technology*, 37(3):71–87.

Baraka, K., Alves-Oliveira, P., and Ribeiro, T. (2020). An extended framework for characterizing social robots. *Human-Robot Interaction: Evaluation Methods and Their Standardization*, pages 21–64.

Bartneck, C., Kulić, D., Croft, E., and Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1:71–81.

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., and Tanaka, F. (2018). Social robots for education: A review. *Science robotics*, 3(21):eaat5954.

Breazeal, C., Dautenhahn, K., and Kanda, T. (2016). Social robotics. *Springer handbook of robotics*, pages 1935–1972.

Broekens, J., Heerink, M., Rosendal, H., et al. (2009). Assistive social robots in elderly care: a review. *Gerontechnology*, 8(2):94–103.

Brooke, J. (1996). SUS: A "quick and dirty" usability scale. *Usability Evaluation in INdustry/Taylor and Francis*.

Brooke, J. (2013). SUS: a retrospective. *Journal of usability studies*, 8(2).

Campos, J., Kennedy, J., and Lehman, J. F. (2018). Challenges in exploiting conversational memory in human-agent interaction. In *Proceedings of the 17th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 1649–1657.

Chai, J. Y., She, L., Fang, R., Ottarson, S., Littley, C., Liu, C., and Hanson, K. (2014). Collaborative effort towards common ground in situated human-robot dialogue. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 33–40.

Coghlan, S., Leins, K., Sheldrick, S., Cheong, M., Gooding, P., and D'Alfonso, S. (2023). To chat or bot to chat: Ethical issues with using chatbots in mental health. *Digital health*, 9:20552076231183542.

Deriu, J., Rodrigo, A., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E., and Cieliebak, M. (2021). Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54:755–810.

Dino, F., Zandie, R., Abdollahi, H., Schoeder, S., and Mahoor, M. H. (2019). Delivering cognitive behavioral therapy using a conversational social robot. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2089–2095. IEEE.

Fong, T., Nourbakhsh, I., and Dautenhahn, K. (2003). A survey of socially interactive robots: Concepts, design, and applications. *Robotics and autonomous systems*, 42:3–4.

Johansson, M. and Skantze, G. (2015). Opportunities and obligations to take turns in collaborative multi-party human-robot interaction. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 305–314.

Krämer, N. C., Hoffmann, L., and Kopp, S. (2010). Know your users! empirical results for tailoring an agent s nonverbal behavior to different user groups. In *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings 10*, pages 468–474. Springer.

Lemaignan, S., Jacq, A., Hood, D., Garcia, F., Paiva, A., and Dillenbourg, P. (2016). Learning by teaching a robot: The case of handwriting. *IEEE Robotics & Automation Magazine*, 23(2):56–66.

Lewis, J. R. (2012). Usability testing. *Handbook of human factors and ergonomics*, pages 1267–1312.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Namlisesli, D., Baş, H. N., Bostancı, H., Coşkun, B., Barkana, D. E., and Tarakçı, D. (2024). The effect of use of social robot nao on children's motivation and emotional states in special education. In *2024 21st International Conference on Ubiquitous Robots (UR)*, pages 7–12. IEEE.

OpenAI (2023). GPT-4 technical report (arxiv: 2303.08774).

Pino, O., Palestra, G., Trevino, R., and De Carolis, B. (2020). The humanoid robot nao as trainer in a memory program for elderly people with mild cognitive impairment. *International Journal of Social Robotics*, 12:21–33.

Pulido, J. C., González, J. C., Suárez-Mejías, C., Bandera, A., Bustos, P., and Fernández, F. (2017). Evaluating the child–robot interaction of the naotherapist platform in pediatric rehabilitation. *International Journal of Social Robotics*, 9:343–358.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision.

Reimann, M. M., Kunneman, F. A., Oertel, C., and Hindriks, K. V. (2024). A survey on dialogue management in human-robot interaction. *ACM Transactions on Human-Robot Interaction*.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215.

Sainburg, T. (2019). timsainb/noisereduce: v1.0.

Sainburg, T., Thielk, M., and Gentner, T. Q. (2020). Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS computational biology*, 16(10):e1008228.

Sauro, J. and Lewis, J. R. (2016). *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann.

Si, W. M., Backes, M., Blackburn, J., De Cristofaro, E., Stringhini, G., Zannettou, S., and Zhang, Y. (2022).

Why so toxic? measuring and triggering toxic behavior in open-domain chatbots. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2659–2673.

Tanaka, F. and Matsuzoe, S. (2012). Children teach a care-receiving robot to promote their learning: field experiments in a classroom for vocabulary learning. *J. Hum.-Robot Interact.*, 1(1):78–95.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models.

Wahde, M. (2024). Models with verbally enunciated explanations: Towards safe, accountable, and trustworthy artificial intelligence. In *ICAART*, pages 101–108.

Wahde, M. and Virgolin, M. (2023). DAISY: An implementation of five core principles for transparent and accountable conversational ai. *International Journal of Human–Computer Interaction*, 39(9):1856–1873.

Youssef, K., Said, S., Alkork, S., and Beyrouthy, T. (2022). A survey on recent advances in social robotics. *Robotics*, 11(4):75.