Evaluating Conversational Question Generation: CoffeeBot

Jelte van Waterschoot
Human Media Interaction,
University of Twente
Enschede, The Netherlands
j.b.vanwaterschoot@utwente.nl

Mariët Theune
Human Media Interaction,
University of Twente
Enschede, The Netherlands
m.theune@utwente.nl

Figure 1: Set-up of the CoffeeBot near the coffee machine at location L2.

ABSTRACT

Asking questions is an important way of showing interest in people and learning more about them. We deployed a social robot, CoffeeBot, capable of asking personalized conversational questions based on topics mentioned by the user over time. We describe our evaluation setup and results of our pilot study. Our initial results, based on interactions with seven participants, indicate a high level of enjoyment among participants and provided us with useful feedback on the CoffeeBot experience for a future long-term study.

CCS CONCEPTS

• **Human-centered computing** → **Field studies**; *Usability testing*; *Natural language interfaces*; Sound-based input / output.

KEYWORDS

evaluating long-term interaction, real-world pilot, question generation, personalization

ACM Reference Format:
Jelte van Waterschoot and Mariët Theune. 2021. Evaluating Conversational Question Generation: CoffeeBot. In *Proceedings of Lifelong Learning and Personalization in Long-Term Human-Robot Interaction (LEAP-HRI 2021).* ACM, New York, NY, USA, 6 pages.

LEAP-HRI 2021, March 8th, 2021, Boulder, CO, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00

1 INTRODUCTION

Conversations near coffee machines might appear superficial, but actually reveal people's world views and beliefs [6]. These conversations happen frequently at public places where classmates and colleagues run into each other at the coffee machine. Colleagues ask each other how they are doing and follow up on past interactions and topics. Asking (personal) questions is an important part of making conversation and showing interest [9]. Some work has already been done on follow-up question asking by conversational agents. In task-oriented conversation, Furhat Robotics [7] and Inoue et al. [10] developed agents capable of asking follow-up questions for job interviews. Mattar and Wachsmuth [18] built an agent that asked follow-up questions in open domain conversation to improve the user's attitude towards the agent. However, none of these agents were tested autonomously for long-term interaction, despite a need for long-term studies [20, 22]. We present the results of our pilot with *CoffeeBot*, a social robot capable of asking personalized questions in real-world situations.

2 PERSONALIZATION

The focus of this paper is on the pilot test of the CoffeeBot, but here we briefly introduce its personalization strategy, which involves asking conversation starters, follow-up questions, and memory-based questions, as well as self-disclosure. Conversation starters are questions such as *Where was your last holiday?*. Follow-up questions are meant to create the impression that the agent is interested in what the user is talking about, for which we use Mandasari [17]'s method of generating follow-up questions. Mandasari used semantic role labelling (SRL) for extracting predicate arguments such as agent, patient, and location. These arguments are then matched with WH-question templates based on SRL patterns in the dataset of Huang et al. [9]. The templates contain question words such as "why" and "where". An example is shown in Table 1, in which the

Table 1: Part of a dialogue with the CoffeeBot. Line 4 is an example of starter question, lines 6 and 8 are examples of self-disclosure and lines 10, 12 and 14 are examples of follow-up questions.

#	S	Turn
4	A	If you opened a business, what kind of business would it be?
5	U	big business
6	A	What is your favorite interest? My favorite has to be to play games.
7	U	play basketball
8	A	My favourite book is The Lord of the Rings. What do you think about The Lord of the Rings?
9	U	I love it
10	A	Interesting! Why do you love it?
11	U	because I like fantasy
12	A	I see. Why do you like fantasy?
13	U	it just imagination like I like I think that's that are not now out of this world
14	A	Okay. Why do you like you think that is that are not now?

agent asks about fantasy in line 12. Memory-based questions are based on recalling "positive" topics (predicate arguments) of past interactions. If sentiment analysis assigns a high average valence to a topic phrase that has been mentioned at least 5 times by the user, the CoffeeBot will ask about this topic, for example *Last week we talked about "big business". How is it going with that?*. Topic phrases are extracted through part-of-speech (PoS) tagging and lexicon-based sentiment analysis is done with Pattern [25]. Finally, self-disclosure is based on manually crafted templates from [16], based on the dataset of [9]. A sentiment memory (list of nouns with positive or negative values) served as persona for the agent for self-disclosure and contained preferences about for example books or games (Table 1 lines 6 and 8). Each interaction with the CoffeeBot starts with small-talk, after which the CoffeeBot asks questions or self-discloses until either it or user ends the conversation.

3 METHOD

The main aim of our setup was to learn about the usability of the CoffeeBot and to learn about personalization through question generation in a spoken interaction. Instead of using a virtual agent for embodiment and placement [27], we opted for a physical robot, which would draw more attention in a public space [23]. Additionally, a life-like human avatar might set expectations higher than a robot with less humanoid characteristics. Therefore we designed a low-cost low-fidelity prototype that could be deployed for a longer period of time in a real public place, such as a coffeeshop or near a coffee machine.

3.1 System setup

The CoffeeBot was embodied as a silver-colored cardboard cut-out robot, shown in Figure 1. The head of the robot contained a Bluetooth speaker, used for recognition of the user speech and the production of speech synthesis for the CoffeeBot. The lower body contained an Arduino Uno, with an MFRC522 connected for scanning RFID (radio-frequency identification) cards, similar to the identification method used by Davison et al. [4]. Both components were connected to a laptop, hidden out of sight, that ran the CoffeeBot's autonomous core system. A remote server ran NLU and NLG tools to reduce the strain on the laptop. The agent used Google Cloud Speech for Automatic Speech Recognition (ASR) and ReadSpeaker's British English voice James for Text-to-Speech Synthesis (TTS).

3.2 Participants

There was no specific target group for the CoffeeBot, except that participants had to be 1) 18 years or older and 2) had to be relatively fluent in English. We deployed the CoffeeBot at two different locations in late 2020, one at a university of applied sciences (L1) and one at a university college (L2). All participants in the study were students attending either of the universities. Before interacting with the CoffeeBot located at L1, students filled in an informed consent form. At L2 the informed consent could only be given by scanning a QR code that had a digitalized version of the informed consent form, which was stored on our university's server. The Ethics Committee of our faculty approved the forms and checked that all data collection was in compliance with university policy and General Data Protection Regulation (GDPR). At both locations, participants were either recruited personally by us or filled in the online forms that were near the CoffeeBot.

Recruitment was less than ideal with many of the staff and students working from home (most of the time) during the COVID19 pandemic. During recruitment and deployment on-site, we wore masks and kept 1,5m distance. We also put a note near the CoffeeBot that only one person at a time could talk to it.We asked participants to talk to the CoffeeBot at least three times, preferably on different days, but it was up to participants when and if they would interact with the CoffeeBot.

3.3 Procedure

At L1, the CoffeeBot was placed on the side of the room, where a participant could sit down. For L2, the CoffeeBot was placed next to an automated coffee machine in a central location. Participants initialized the conversation by holding their card or key with RFID in front of the CoffeeBot's body.

In the first interaction, the CoffeeBot would introduce itself and its goal of wanting to get to know the participant. It also gave the instruction that the participant could end the conversation by saying "goodbye". After the introduction, the CoffeeBot started a round of questions. The user might answer these or not, but at the end of the user's turn, the CoffeeBot asked another question, either as a follow-up on what the user said or introducing a new topic. The conversation could go on for up to 6 minutes, after which the CoffeeBot ended the conversation with a message to hopefully see the user again soon, or the user could end the conversation earlier by saying "goodbye". In subsequent conversations with the same participant, the CoffeeBot would start by saying it would ask questions and continue to do so until either the CoffeeBot or participant ended the conversation.

3.4 Data collection & Measurements

In the pilot experiment we collected data from three different sources: from the CoffeeBot itself, which made speech recordings, transcriptions and logs from its interactions with the user, two questionnaires filled in by users and semi-structured interviews held at the end of the experiment by the researcher.

3.4.1 Recordings, transcriptions and logs. The speech recordings were saved in 16-bit PCM wav format named by date and user ID. Each recording started when a user scanned their radio-frequency identification (RFID) card and stopped when the user or CoffeeBot ended the conversation. Additionally, the CoffeeBot saved the transcriptions of the speech and logs of the interaction in a database in JSON format. The logs contain the meta-data of the conversations, such as the user ID, the number of interactions, the sentiment levels of user utterances and the dialogue history with timestamps. We also calculated how many topics were not recalled and how many follow-up questions and self-disclosures were incorrectly used.

3.4.2 Questionnaires. We compiled two questionnaires, which took about 4-5 minutes each to complete. One was sent to participants after their first interaction and the second one was sent after the experiment ended. The second questionnaire contained additional questions about how often people had been at the university during the run of the experiment and how often they had left their working place for a break (see Appendix A.1). Both questionnaires contained the same items, which were taken from four different questionnaires as described below. All questionnaires were standardized to a 7-point Likert scale.

The first two questionnaire components focused on the *personalization* aspect of the CoffeeBot in the open-domain small-talk. The first was based on the Interpersonal Communication Satisfaction (ICS) measure by Hecht [8], which has been specifically designed for dyadic conversations for either friends, strangers or acquaintances and can measure closeness in computer-mediated communication [26] (Appendix A.1.1). ICS has been applied before in Human Robot Interaction (HRI) research, in a study about conversational memory [2]. The second component was based on the McGill Friendship questionnaire [19], which was used by Leite et al. [14] to evaluate how helpful and encouraging the participants found their social robot. Four out of the six categories of the questionnaire did not apply to the CoffeeBot's type of casual conversation: help, reliable alliance, self-validation and emotional security, so we did not include these in our questionnaire (Appendix A.1.2). The two categories we did include were intimacy and stimulating companionship. Intimacy is about being honest, expressing yourself and how comfortable you are with sharing personal information. The items for stimulating companionship measure how enjoyable the conversation is. They served as an indicator for engagement and if people wanted to talk again to the CoffeeBot.

The purpose of the other two questionnaire components was to measure the quality of the CoffeeBot in general: Robotic Social Attribute Scale (RoSAS) for the perception users had of the CoffeeBot [3] and a social presence questionnaire for measuring their engagement [12]. We used the RoSAS questionnaire because it has been used commonly as an evaluation tool for social robots and agents and is well-known to HRI researchers [13] (Appendix A.1.4). We are aware of Werner [28]'s statement that the validation evidence for RoSAS has been rather limited [13, 21]. Biocca et al. [1] reviewed the concept of social presence. There are many different interpretations of this concept in a human-human context [24], but here we use the perceived social presence of the robot as a measure of engagement, similarly to Jung and Lee [12]. We measured the social presence of the CoffeeBot with Jung and Lee [12]'s questionnaire to gauge the feeling of users being socially present with the CoffeeBot (Appendix A.1.3).

3.4.3 Interview. At the end of the experiment, together with the second questionnaire, an email invitation for a semi-structured interview was sent to all of the participants to informally talk about their experience with the CoffeeBot. See Appendix A.2 for the leading interview questions.

4 COFFEEBOT PILOT RESULTS

Thirteen people gave informed consent to participate in the experiment. Of these participants, 2 never talked to the CoffeeBot. Out of the 11 people who did interact with the CoffeeBot, 6 people interacted with it more than once and 5 interacted with it only once. Of these 5 participants, only one interacted with the CoffeeBot with more than 2 (user) turns. If we filter out participants who had no more than one turn per interaction, 7 participants (3 male, 2 female, 2 non-binary) remain (see Table 2). The results we describe relate to these final 7 participants.

The CoffeeBot was deployed for two weeks at location L1 and three weeks at location L2, with one week in between. In the last week at location L2, a strict lockdown was enforced. At location L1, only one participant interacted with the CoffeeBot more than once, but this participant did not have more than one turn per interaction, and no questionnaires were filled in by any of the participants. Therefore the results of the questionnaire and interviews are all from participants who talked to the CoffeeBot at location L2.

4.1 Interaction metadata

We are interested if the CoffeeBot can mimic short coffee talk conversations, for multiple interactions and for around 5 minutes of conversation. The average number of turns per interaction was 10,6 (σ=6,68), the average duration was 3 minutes and 7 seconds. The average number of interactions was 2,42 (σ=0,98). The average topic recall was 0.607: around 3 out of 5 topics were recalled. About half the follow-up questions were inappropriate (0.46). Self-disclosure error rate was similar to the reverse topic recall rate with 0.4. See Table 2 for the results per participant, and for more details on how the errors were computed.

4.2 Questionnaire

In total 6 participants filled in the first questionnaire and 2 participants filled in the second questionnaire. Only one of the participants filling in the second questionnaire talked to the CoffeeBot on separate days. Though this is limited data, we do want to mention the results of the first questionnaire. For the ICS part, most participants were favorable about having another conversation (median=5), enjoyed the conversation (5) and were satisfied (5.5). However, they did not feel the conversation went smoothly (2.5). Additionally, the agent performed poorly in terms of invoking user laughter (3.5),

Table 2: Quantitative results of the CoffeeBot pilot. The user ID indicates the location (L) and participant (U) number. The interactions indicate total (different days) interactions. The topic recall is calculated by taking the number of recalled topics (noun and verb phrases) divided by the total number of topics mentioned. For the self-disclosure (sd) and follow-up errors, the error was calculated by manually annotating coherence and dividing incoherent self-disclosure and follow-up questions by the total number of self-disclosure and follow-up questions by the agent, respectively.

ID	total turns	inter-actions	mean length	topic recall	sd error	follow up error
L1U4	8	1	3:53	1/2	0/0	0/0
L2U1	30	3(2)	2:39	11/21	0/0	3/7
L2U2	12	2(1)	2:06	7/9	1/1	0/2
L2U4	24	4(1)	2:08	12/17	1/4	3/6
L2U5	35	3(2)	2:48	9/19	0/0	4/6
L2U6	15	2(1)	2:17	5/12	0/0	3/4
L2U7	50	2(2)	5:59	32/47	0/0	6/16

letting the user know they were communicating effectively (3.5) and talking about interesting things (3). On all points of the intimacy part of the McGill friendship questionnaire, the CoffeeBot scored extremely poor to poor (≤3). Only for the statement "CoffeeBot would listen if I talked about my problems" the ratings were average (4). The CoffeeBot scored much better on all items of the stimulating companionship scales (≥4), where only the statement "CoffeeBot has good ideas about entertaining things" was rated as poor (3). Social presence scored mostly mediocre, with the CoffeeBot not being very life-like (3), but it was considered quite sociable (5) and people felt it communicated with them (4). On the RoSAS scale, the interaction with the CoffeeBot was found to be interactive (5), strange (4.5), competent (4) and responsive (4.5), whereas it was not found to be emotional (2), compassionate (2), aggressive (1), dangerous (1) or feeling (2.5).

4.3 Interview

Of the 7 participants, 2 consented to a separate short interview. The other participants were sent reminders via email, but did not respond. The length of each interview was about 15 minutes. Both interviewees said that it was easy to interact with the bot, though the signing up process was a bit difficult to understand and they had to repeat themselves. One of the interviewees said that the robot looked cute as is, though the other one would have liked to see something like movement of the arms to make it more engaging. Neither of them found the questions the CoffeeBot asked to be too personal, and both thought the topics were appropriate for the setting. Annoyances mentioned by the interviewees were that the CoffeeBot repeated itself often, made some weird topic switches and the voice was sometimes hard to understand. When asked about other practical uses for a CoffeeBot-like robot, one interviewee mentioned it could help train interviewing skills or could be positioned at waiting rooms.

5 DISCUSSION

The foremost limitation of this research was the small number of participants and interactions for each participant. The COVID19 pandemic and a lockdown severely restricted the number of interactions possible over the course of 2 or 3 weeks. Our analysis is therefore impacted by the novelty effect, because users did not get familiar with the CoffeeBot.

Even though the Google ASR worked relatively well in a not-so crowded coffee place, the ASR had some issues with speaker diarization, separating background voices from the participant's. About 40% of the topics were not recalled, mostly due to ASR errors. Irfan et al. [11] found the same issues with their social robot as barista in a coffee shop. They recommend constraining the grammar of the ASR model and adapting it to non-native English as well. Follow-up questions were misclassified or inappropriate due to ungrammatical sentences that could not be correctly parsed with the current SRL implementation. Inoue et al. [10]'s work on follow-up questions might be more robust to ungrammaticalities and can be adapted for open-domain chat. Additionally, some of the sentences spoken by the TTS were not comprehensible for participants, as also mentioned by one of our interviewees.

Despite these errors, participants did still enjoy the conversations and company of the CoffeeBot, though they felt the flow of the interaction was insufficient. Unfortunately, due to the limited interactions and no topic being discussed at least 5 times, none of the memory-based questions that should trigger a sense of personalization were asked in any of the interactions.

The design of the CoffeeBot is simple, to set expectations low. According to Werner [28] this usually has a negative impact on the interaction. As mentioned by one of the interviewees, the CoffeeBot could have benefited from more engaging behaviour such as movement without raising expectations too much. Another issue was the turn-taking in the interaction, where both the robot and user had trouble adapting to each other's timing. For the robot this could be solved by having better incremental speech recognition, whereas for the user it could be resolved by adding thinking behaviours in the robot so that the user would know not to talk too soon.

6 FUTURE WORK

In future work, we want to deploy the CoffeeBot for at least two months at the same location in a public space. Two months should be sufficient for the robot to be socially accepted and reduce the novelty effect [5, 15]. We plan to set up a between-subject study where the experiment group gets memory-based questions and the control group does not. We will also finalize the prototype based on the feedback from the participants, by either improving the current prototype (e.g. moving arms, adding lights to guide turn-taking) or using an existing social robot that has these features.

ACKNOWLEDGMENTS

We thank the University College Roosevelt in Middelburg and the University of Applied Sciences in Vlissingen for their collaboration on this project. This work is part of the research programme Data2Person with project number 628.011.029, which is (partly) financed by the Dutch Research Council (NWO).

REFERENCES

[1] Frank Biocca, Chad Harms, and Judee K. Burgoon. 2003. Toward a More Robust Theory and Measure of Social Presence: Review and Suggested Criteria. *Presence: Teleoperators and Virtual Environments* 12, 5 (Oct. 2003), 456–480. https://doi.org/10.1162/105474603322761270 Publisher: MIT Press.

[2] Joana Campos, James Kennedy, and Jill F Lehman. 2018. Challenges in Exploiting Conversational Memory in Human-Agent Interaction. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, M. Dastani, G. Sukthankar, Elisabeth André, and S. Koenig (Eds.). IFAAMAS, Stockholm, Sweden, 1649–1657. https://dl.acm.org/doi/10.5555/3237383.3237945 Series Title: AAMAS 2018.

[3] Colleen M. Carpinella, Alisa B. Wyman, Michael A. Perez, and Steven J. Stroessner. 2017. The Robotic Social Attributes Scale (RoSAS). In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17*, Vol. Part F1271. ACM Press, New York, New York, USA, 254–262. https://doi.org/10.1145/2909824.3020208 ISSN: 21672148.

[4] Daniel P. Davison, Frances M. Wijnen, Vicky Charisi, Jan van der Meij, Vanessa Evers, and Dennis Reidsma. 2020. Working with a Social Robot in School: A Long-Term Real-World Unsupervised Deployment. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20)*. Association for Computing Machinery, Cambridge, United Kingdom, 63–72. https://doi.org/10.1145/3319502.3374803

[5] M. M. A. de Graaf, Soumaya Ben Allouch, and Johannes A. G. M. van Dijk. 2016. Long-term acceptance of social robots in domestic environments: Insights from a user's perspective. In *The 2016 AAAI Spring Symposium Series*. AAAI, Palo Alto, CA, USA, 96–103. https://research.utwente.nl/en/publications/long-term-acceptance-of-social-robots-in-domestic-environments-in

[6] Suzanne Eggins and Diana Slade. 2001. *Analysing Casual Conversation* (1st ed.). Continuum, London, UK.

[7] Thomas Grose. 2019. FAIR IMPRESSIONS. *American Society for Engineering Education* 28, 8 (2019), 14–14. https://search.proquest.com/openview/99eaa6c96dfd687e7801d7f440e5ff8c/1?pq-origsite=gscholar&cbl=33050

[8] Michael L. Hecht. 1978. The Conceptualization and Measurement of Interpersonal Communication Satisfaction. *Human Communication Research* 4, 3 (1978), 253–264. https://doi.org/10.1111/j.1468-2958.1978.tb00614.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-2958.1978.tb00614.x.

[9] Karen Huang, Michael Yeomans, Alison Wood Brooks, Julia Minson, Francesca Gino, Karen Huang, Michael Yeomans, Alison Wood Brooks, Julia Minson, and Francesca Gino. 2017. It Doesn' t Hurt to Ask : Question-Asking Increases Liking. *Journal of Personality and Social Psychology* 113, 3 (2017), 430–452.

[10] Koji Inoue, Kohei Hara, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2020. Job Interviewer Android with Elaborate Follow-up Question Generation. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20)*. Association for Computing Machinery, New York, NY, USA, 324–332. https://doi.org/10.1145/3382507.3418839

[11] Bahar Irfan, Mehdi Hellou, Alexandre Mazel, and Tony Belpaeme. 2020. Challenges of a Real-World HRI Study with Non-Native English Speakers: Can Personalisation Save the Day?. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20)*. Association for Computing Machinery, New York, NY, USA, 272–274. https://doi.org/10.1145/3371382.3378278

[12] Younbo Jung and Kwan Min Lee. 2004. Effects of Physical Embodiment on Social Presence of Social Robots. In *Proceedings of the 7th Annual International Workshop on Presence*. International Society for Presence Research, Valencia, Spain, 80–87.

[13] Christian U. Krägeloh, Jaishankar Bharatharaj, Senthil Kumar Sasthan Kutty, Praveen Regunathan Nirmala, and Loulin Huang. 2019. Questionnaires to Measure Acceptability of Social Robots: A Critical Review. *Robotics* 8, 4 (Dec. 2019), 88. https://doi.org/10.3390/robotics8040088 Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[14] Iolanda Leite, Ginevra Castellano, Andr\'{e} Pereira, Carlos Martinho, and Ana Paiva. 2014. Empathic Robots for Long-term Interaction: Evaluating Social Presence, Engagement and Perceived Support in Children. *International Journal of Social Robotics* 6, 3 (2014), 329–341. https://doi.org/10.1007/s12369-014-0227-1 ISBN: 1236901402271.

[15] Iolanda Leite, Carlos Martinho, and Ana Paiva. 2013. Social Robots for Long-Term Interaction: A Survey. *International Journal of Social Robotics* 5, 2 (April 2013), 291–308. https://doi.org/10.1007/s12369-013-0178-y tex.ids: leite2013social.

[16] Mikael Lundell Vinkler and Peilin Yu. 2020. *Conversational Chatbots with Memory-based Question and Answer Generation*. Master's thesis. Linköping University, Linköping. http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-171927

[17] Yani Mandasari. 2019. *Follow-up Question Generation*. Master's thesis. Universtiy of Twente. http://purl.utwente.nl/essays/79491

[18] Nikita Mattar and Ipke Wachsmuth. 2014. Let's Get Personal. In *Human-Computer Interaction. Advanced Interaction Modalities and Techniques (Lecture Notes in Computer Science)*, Masaaki Kurosu (Ed.). Springer International Publishing, Cham, 450–461. https://doi.org/10.1007/978-3-319-07230-2_43

[19] Morton J. Mendelson and Frances E. Aboud. 1999. Measuring friendship quality in late adolescents and young adults: McGill Friendship Questionnaires. *Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement* 31, 2 (1999), 130–132. https://doi.org/10.1037/h0087080

[20] Nahal Norouzi, Kangsoo Kim, Jason Hochreiter, Myungho Lee, Salam Daher, Gerd Bruder, and Greg Welch. 2018. A Systematic Survey of 15 Years of User Studies Published in the Intelligent Virtual Agents Conference. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. ACM, Sydney, NSW, Australia, 17–22. https://doi.org/10.1145/3267851.3267901

[21] Matthew K. X. J. Pan, Elizabeth A. Croft, and Günther Niemeyer. 2017. Validation of the Robot Social Attributes Scale (RoSAS) for Human-Robot Interaction through a Human-to-Robot Handover Use Case. In *IROS 2017 Workshop Human-Robot Interaction in Collaborative Manufacturing Environments (HRI-CME)*. IEEE, Vancouver, Canada, 2.

[22] Astrid Marieke Rosenthal-von der Pütten, Astrid Weiss, and Selma Šabanović. 2016. The challenge (not) to go wild! Challenges and Best Practices to Study HRI in Natural Interaction Settings. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction (HRI '16)*. IEEE Press, Christchurch, New Zealand, 583–584.

[23] Elena Márquez Segura, Michael Kriegel, Ruth Aylett, Amol Deshmukh, and Henriette Cramer. 2012. How Do You Like Me in This: User Embodiment Preferences for Companion Agents. In *Intelligent Virtual Agents (Lecture Notes in Computer Science)*, Yukiko Nakano, Michael Neff, Ana Paiva, and Marilyn Walker (Eds.). Springer, Berlin, Heidelberg, 112–125. https://doi.org/10.1007/978-3-642-33197-8_12

[24] John Short, Ederyn Williams, and Bruce Christie. 1976. *The Social Psychology of Telecommunications*. Wiley, London, UK. Google-Books-ID: Ze63AAAAIAAJ.

[25] Tom De Smedt and Walter Daelemans. 2012. Pattern for python. *Journal of Machine Learning Research* 13, 66 (2012), 2063–2067. http://jmlr.org/papers/v13/desmedt12a.html

[26] Lisa Collins Tidwell and Joseph B. Walther. 2002. Computer-Mediated Communication Effects on Disclosure, Impressions, and Interpersonal Evaluations: Getting to Know One Another a Bit at a Time. *Human Communication Research* 28, 3 (2002), 317–348. https://doi.org/10.1111/j.1468-2958.2002.tb00811.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-2958.2002.tb00811.x.

[27] Michel Valstar, Soumia Dermouche, Catherine Pelachaud, Eduardo Coutinho, Björn Schuller, Yue Zhang, Dirk Heylen, Mariët Theune, Jelte van Waterschoot, Tobias Baur, Angelo Cafaro, Alexandru Ghitulescu, Blaise Potard, Johannes Wagner, Elisabeth André, Laurent Durieu, and Matthew Aylett. 2016. Ask Alice: an artificial retrieval of information agent. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016*. ACM Press, Tokyo, Japan, 419–420. https://doi.org/10.1145/2993148.2998535 Series Title: ICMI '16 Issue: November.

[28] Franz Werner. 2020. A Survey on Current Practices in User Evaluation of Companion Robots. In *Human-Robot Interaction: Evaluation Methods and Their Standardization*, Céline Jost, Brigitte Le Pévédic, Tony Belpaeme, Cindy Bethel, Dimitrios Chrysostomou, Nigel Crook, Marine Grandgeorge, and Nicole Mirnig (Eds.). Springer International Publishing, Cham, 65–88. https://doi.org/10.1007/978-3-030-42307-0_3

A RESEARCH METHODS

A.1 Questionnaire

Here we list the questions of the questionnaire for each item we were interested in. All items are rated on a 7-point Likert scale.

A.1.1 Interpersonal Communication Satisfaction [8].

- CoffeeBot let me know I was communicating effectively.
- I would like to have another conversation like this one.
- CoffeeBot genuinely wanted to get to know me.
- I was NOT satisfied with the conversation.
- I actually had something else to do.
- I felt that during the conversation I was able to present myself as I wanted CoffeeBot to view me.
- CoffeeBot understood what I said.
- I was very satisfied with the conversation.
- CoffeeBot expressed a lot of interest in what I had to say.
- I did NOT enjoy the conversation.
- CoffeeBot did NOT provide support for what he was saying.
- I felt I could talk about anything with CoffeeBot.
- We each got to say what we wanted.
- I felt that we could laugh easily together.

- The conversation flowed smoothly.
- CoffeeBot changed the topic when his feelings were brought in the conversation.
- CoffeeBot frequently said things which added little to the conversation.
- We talked about something I was NOT interested in.

A.1.2 McGill Friendship Questionnaire [19].

- Intimacy:
  - CoffeeBot is someone I can tell private things to.
  - CoffeeBot knows when I'm upset.
  - CoffeeBot is someone I can tell secrets to.
  - CoffeeBot knows when something bothers me.
  - CoffeeBot would listen if I talked about my problems
  - CoffeeBot would understand me if I told him my problems.
  - CoffeeBot is easy to talk to about private things.
  - CoffeeBot understands my feelings.
- Stimulating companionship:
  - CoffeeBot is fun to do things with.
  - CoffeeBot tells me interesting things.
  - CoffeeBot has good ideas about entertaining things to do.
  - CoffeeBot makes me laugh.
  - CoffeeBot is exciting to talk to.
  - CoffeeBot is enjoyable to be with.
  - CoffeeBot is exciting to be with.
  - CoffeeBot is fun to stand and talk with.

A.1.3 Social Presence [12].

- How sociable was the CoffeeBot?
- How personal was the CoffeeBot?
- How life-like was the CoffeeBot?
- How sensitive was the CoffeeBot?
- While you were interacting with the CoffeeBot, how much did you feel as if he was a social being?
- While you were interacting with the CoffeeBot, how much did you feel as if he was communicating with you?

A.1.4 RoSAS [3]. Instead of statements, for RoSAS participants had to rate whether the word described the CoffeeBot very well (7) or not at all (1).

Scary	Emotional	Dangerous
Knowledgeable	Compassionate	Awkward
Reliable	Organic	Aggressive
Interactive	Social	Awful
Responsive	Feeling	Strange
Capable	Happy	Competent

A.1.5 Second questionnaire. We asked these two questions to put into context how often people had spontaneous opportunities to talk to the CoffeeBot.

- How often were you in your own standard workplace over the course of two (three) weeks?
- How often did you walk away from your desk on a full work day at your standard workplace over the last two (three) weeks?
- Did you talk to the CoffeeBot more than once?

A.2 Interview questions

The semi-structured interview contained five questions:

- How easy was it for you to interact with the CoffeeBot?
- What would you change about the appearance of the CoffeeBot?
- What kind of topics would you like to discuss with the CoffeeBot?
- What were some annoying things about the CoffeeBot?
- What practical uses do you see for a CoffeeBot?