Social and Contextual Effects on Human Interaction with a Digital Conversational Agent

Hannah Qu
Undergraduate Research Scholar
Department of Psychology, University of Minnesota

Wilma Koutstaal: Undergraduate Research Scholarship Mentor, Department of Psychology
Libby Ferland: Graduate Student Mentor, Department of Computer Science

Abstract

Smart home assistants, or digital conversational agents (CA), can improve people's lives by helping with reminders and schedules, as well as serving a social function. Research shows that users prefer some features of CAs, such as their adaptability to individuals, more simple speech, and human-like voice. This study aims to develop a CA for older people to support their cognitive and social needs. Before conducting research on older participants, in order to gain a basic understanding of our study design and effects of the interactions, we first conducted the study with young college students at the University of Minnesota, by asking them to interact with a prototype CA for multiple times. User experience data show that participants were more satisfied after being familiarized with the process and after the CA adapted to them. Sociolinguistic analyses of the transcribed content of participants' interactions revealed that the interactions appeared to be more natural when the participant was alone during their interaction with the CA than when an outside observer (research assistant) was also present during their dialogue with the CA. These findings have implications for understanding and researching human computer interactions. Obtaining accurate profiles of users' patterns of interactions with smart home assistants requires a longitudinal rather than cross-sectional approach, that also considers the broader social or interpersonal context of the assessment environment itself.

Introduction

Conversational agents (CA), such as Amazon's "Alexa" or Apple's "Siri" improve life quality by helping users keep track of plans and have a social function by having conversations with users. Research shows that CAs can be especially helpful for older people by providing cognitive support and encouraging skill maintenance. CAs sustain the independence and dignity of older people; they would need less human help and would not feel embarrassed about sensory or cognitive deficits. CAs also help older people schedule events and reminders, and reduce stress, so that people do not need to worry about forgetting about things (Ferland et al., 2019).

Previous studies have noticed several features of the CAs that are preferred by users. People are interested in the adaptability and user acceptance of the smart home assistant: they prefer that the system adapt to, or memorize, their unique habits (such as users' "capacities, preferences, and fears") and develop or "learn" through interactions with the user (Ferland et al., 2019; Portet et al., 2013). The usefulness of the technology is also important. Providing memory support and maintaining independence would be key functions for older people

(Ferland et al., 2019). Another feature is social inclusion, which helps decrease loneliness. The smart home assistant acts as an intelligent companion whom older individuals can talk with. The CA can provide "security reassurance by detecting situations of distress" (Portet et al., 2013). One research study on using smart assistants to track people's health showed that both conversational agents and humans in a message-based health coaching are preferred by people, and the overall experience of using the technology to track health status is good (Fadhil, Wong, & Reiterer, 2018).

Studies show that the language of the system should be natural but simple, so that it is easier for older individuals to use and accept the new technology (Ferland et al., 2019). Portet and other researchers showed that seniors tend to pay attention to the key words in audio, indicating that the scripts should be simple and direct (Portet et al., 2013). Also, interacting with the CA via a voice interface is strongly preferred over a tactile system (or touch-screen) that requires being physically available and may require computer-based skills (Portet et al., 2013).

In this study, we aim to develop a smart home assistant, or a CA, that is well-suited for older people. We are interested in the effect of social aspects (human-like language and interaction) and the scheduling functions of the CA. Participants interact with a prototype CA in the study and provide feedback on user experience. The interaction between each participant and the CA happened multiple times, in order to see a long-term effect of the interactions. Following methods developed in initial work (Ferland & Koutstaal, 2020), the audio-recorded interactions of the participants with the CA were first transcribed and then characterized on several sociolinguistic dimensions (e.g., the number of words spoken, and the degree of frankness or "authenticity" of the dialog) based on the "Linguistic Inquiry Word Count" program (Pennebaker et al, 2015). After the first phase of experiment (more details are provided in the Method section), I analyzed the data to answer two research questions:

RQ1. The effects of increased familiarity or experience with the CA beyond the initial encounters.

RQ2. The effects of the presence versus the absence of an observer or another person on the nature of the interactions with the CA.

Method

Sample

In the first phase, we began by testing younger participants to gain a general understanding of the data and revise research procedures for the older participants in phase 2 study. A total of 33 undergraduate students from University of Minnesota Twin Cities were recruited through the university REP points system across three academic terms (Spring and Fall 2019, and Spring 2020).

Procedure

Each participant interacted with the CA in the research lab (the CA asked questions about participants' daily schedules) typically for four sessions, answered a post-session survey at the end of first three sessions and one post-study survey at the end of the fourth session. The first two sessions were conducted in the morning and afternoon of the first day, and the last two sessions were conducted in the morning and afternoon of the third day. A total of 30 sessions, including 8 participants, were conducted online via Zoom after March, 2020, with one

participant completing only the first 2 sessions, and the others completing all 4 sessions.

A research assistant (RA) was present in the lab room when a participant interacted with the CA in Spring 2019, and absent during the interaction in Fall 2019 and Spring 2020 after I suggested that an observer (the RA) might affect the user-CA interactions.

There were two conditions in the study. Condition A refers to refers to interactions in which the CA didn't remember or know anything about the user, and Condition B refers to interactions in which the CA adapted and remembered earlier interactions and details about the user (such as the individual's calendar events). Participants were randomly assigned to conditions "AABB", meaning that the CA did not adapt to or remember participants' information until the third session, or "ABBB", meaning that the CA adapted to or remembered participants' information in the second session; the CA adapted to all participants in the last two sessions.

Soundboard

The study used a "Wizard of Oz" system to gather data on user experience. A "Wizard of Oz refers to a person (usually the experimenter, or a confederate) remotely operating a robot, controlling any of a number of things, such as its movement, navigation, speech, gestures, etc." (Riek, 2012). This technique allows researchers to simulate a human to robot conversation when autonomy cannot be realized in the stage of development. In this study, a researcher controlled the speech of a communication robot – and we told participants that it was a "prototype CA" – to collect data of users' possible responses and needs for a future AI. The voice of the CA ("Joanna") was provided by Amazon Web Service's "Polly" service. To make the CA respond accurately, promptly, and as naturally as a human, a researcher developed a sound board (see Figure 1). The sound board contained most questions and answers in a general form that the CA would say to participants, and the researcher typed responses that were more specific to each participant in the middle area of the sound board.

Figure 1. Sound board before modification

I helped modify the human-computer interaction interface of the sound board (see Figure 2). Main changes included the layout of the soundboard and development of a tutorial for other researchers to use. The phrase tabs are organized in the order of being used. During the in-

person study sessions, the CA has to follow a general script to ask questions and at the same time respond accordingly to different participants without there being a long lag. Thus, it will be easier for the researcher to navigate through during a stressful period as she needs to respond promptly and individualize answers to different participants. Additionally, we added the "Generic slots" with future and past tenses. By separating these generic terms from other tabs, users can find them easily, which is especially useful when the CA needs to follow up with an event the participant indicates. For the first phase of this study, we added a portion of questions dedicated to the student participants. Having these tabs reduces the amount of questions the user needs to type, which makes it easier to control the CA.

Color coding is another enhancement to the user interface. During a stressful time where the researcher needs to react fast, colors help with navigation and avoiding mistakes. For example, on the lower half of the sound board, "F" and "P" choices are color-coded to explicitly show which tense (future or past) the CA will use. The tabs for events turn gray (or dim) after being used; it ensures that the CA does not miss or repeat a question.

Furthermore, a tutorial of the sound board was developed for new users. In the tutorial mode, by hovering the cursor to a tab on the soundboard, a full description appears to explain what the tab means. For instance, if the cursor lands on "Intro", the new user is able to understand that by clicking the tab, the CA will introduce itself to the user. This tutorial mode increases the usability of the sound board for a broader range of users.

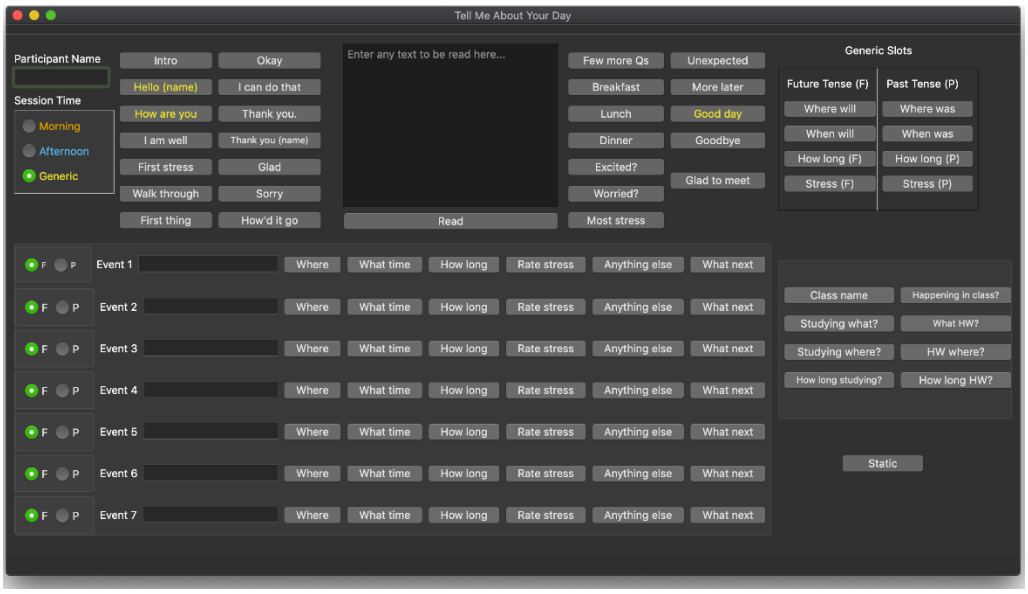

Figure 2. Sound board after modification

Measures

System Usability

Originally created by John Brooke in 1986, the 10-item "System Usability Scale" (SUS) was given to participants at the end of each of the four CA-participant calendar-related experimental sessions (usability.gov/how-to-and-tools/methods/system-usability-scale.html; Appendix A). The 10 questions, adapted for the experimental setup, provided a global measure of system satisfaction, usability, and learnability. Each question was answered on a 5-point scale (from

strongly disagree to strongly agree); reverse-coded items are rescaled, and the final composite score is normalized to a value between 0 and 100. A higher score suggests a better system usability.

Linguistic Inquiry and Word Count (LIWC) psycholinguistic variables (Pennebaker et al, 2015). The LIWC application provides both simple counts of the frequency of words in different semantic and syntactic categories, and standardized measures assessing particular semantic and sociolinguistic properties. We focused on the ***average Word Count*** (the number of words spoken by both the participant and the CA during their interactions) and four summary variables (standardized measures based on a 100-point scale ranging from 0 to 100) for participants and the CA:

(a) ***Analytical thinking***. A high score means a more formal and logical response, and a low score means a less structured or less reserved response. Participants getting a lower score in Analytical thinking suggests that they may feel more comfortable or relaxed with the CA.
(b) ***Clout***. A high score means a more authoritative and confident response, and a low score means a more humble and tentative tone.
(c) ***Authenticity***. A higher score means a more self-disclosing response, and a lower score means a more guarded response.
(d) ***Emotional tone***. A high score means a more positive tone, and a low score indicates anxiety or hostility.

We also considered a few other variables, for raw word counts (without standardized scales): (e) Use of the personal pronoun ***"I"***, (f) use of ***Informal speech***, (g) words relating to ***Motivational drives*** (revealing what participants care or worry about), and (h) ***Social-related*** *words*. A higher frequency of all four of these latter variables indicates more self-disclosing, and a lower frequency indicates a more guarded state.

Results

System Usability

Figure 3 presents a box plot of the system usability (SUS) scores, separated by session (session 1 through session 4) and by whether the CA adapted quickly to the user ("fast" condition, or "ABBB") or the CA adapted somewhat more slowly to the user ("slow" condition, or "AABB"). Results are summarized from 19 participants who completed all four lab-based sessions gathered in Fall 2019 and Spring 2020 (11 in the fast adaptation condition and 8 in the slow adaptation condition).

Based on a large number of studies using the SUS, scores above 68 are considered above average; scores above 80.3 reach the top 10% of scores and this is also "the point where users are more likely to be recommending the product to a friend" (Sauro, 2011). As shown in Figure 2, in general, the median scores reached the top 10% of the SUS, and most ranges were above 68. Additionally, the median SUS of session 3 and session 4 (when the CA has adapted to all participants) was higher than that of the first two sessions. This indicates that users were more satisfied with their interaction with the CA after its adaption. Although the sample size is small, a similar pattern was observed in the mean (rather than median) SUS ratings, such that there was a numerical increase in SUS ratings from session 1 to session 4 for both the fast-adapted

system (M = 78.41 and M = 83.64 for session 1 and session 4, respectively) and the slow-adapted system (M = 80.00 and M = 83.44, respectively). Thus, although user satisfaction was generally high even for the unadopted assistant, the individualization of a smart home assistant improves the usability and user satisfaction of the device.

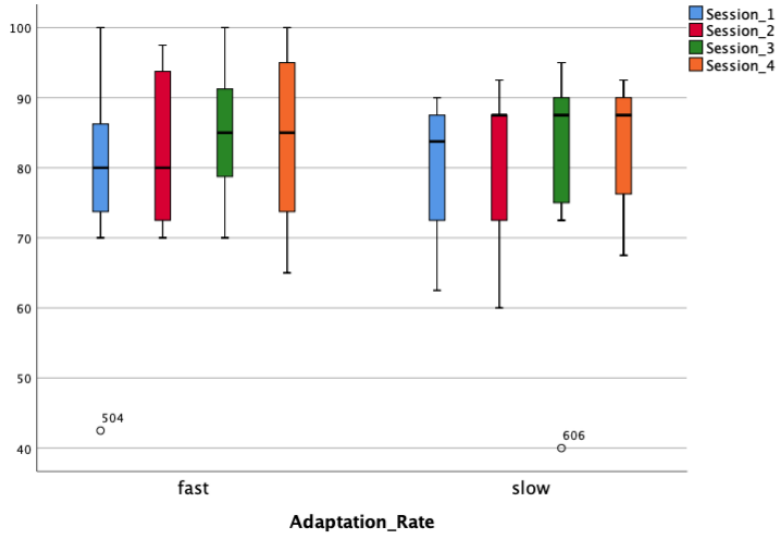

Figure 3. System Usability Scale vs. CA Adaptation Rate across four sessions

Effects of increased familiarity or experience with a user-adapted CA

The first research question is about the effects of increased familiarity after the CA was adapted to all users. Data were gathered from session 3 (N = 28) and session 4 (N = 26). According to LIWC variable analysis, as participants became more familiar with the CA, and the CA was adapted, or somewhat individualized to have learned about participants' daily schedules, they felt more comfortable and were more willing to reveal themselves. Specifically, participants used fewer analytical words in session 4 (median = 30) than in session 3 (median = 64), clout increased slightly from session 3 (median = 22) to session 4 (median = 29), and positive tone (Figure 4) increased markedly from session 3 (median = 72) to session 4 (median = 99, with a very condensed distribution at the high score). Participants' use of the personal pronoun "I" also increased slightly from session 3 (median = 5.5) to session 4 (median = 7.3), social-related words increased (session 3 median = 4 and session 4 median = 7), and mentioning of motivational drives also increased (session 3 median = 3, session 4 median = 5.6), with a higher degree of variability in the later session (session 3, interquartile range of 2.06, session 4, interquartile range of 6.21).

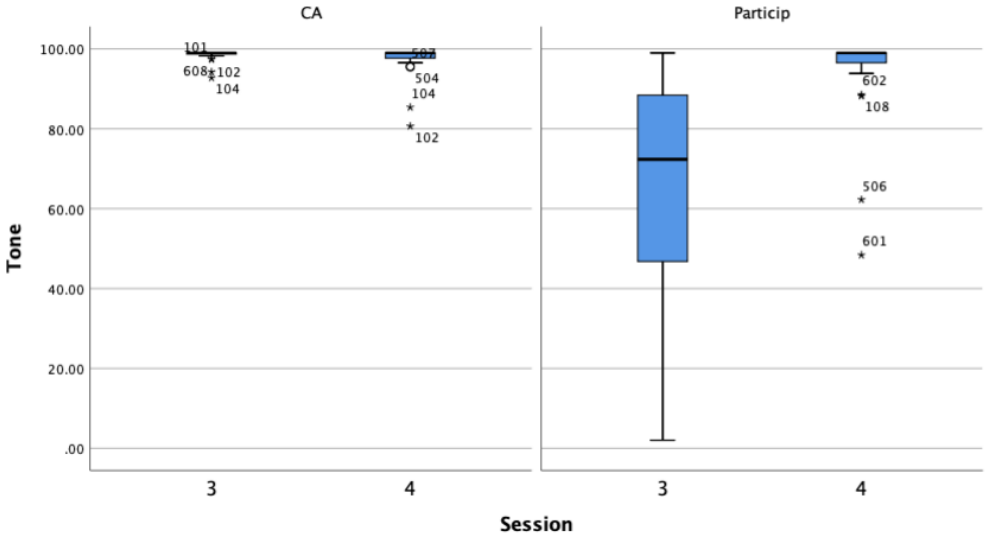

Figure 4. Positive Tone in session 3 and session 4 shown separately for the CA and participants

Effects of an observer on CA interactions

The second research question addresses the effects of the presence versus the absence of an observer or another person during the CA interactions. Data was gathered from 60 observer (the RA) present sessions and 24 observer-absent sessions; data from online testing was excluded because we did not know if participants were by themselves at home during the interaction. The number of words participants spoke (word count) and the LIWC psycholinguistic variables (Pennebaker et al, 2015) show that the presence of an observer makes participants more guarded and less likely to reveal themselves, and the absence of an observer reduces the possible side-effect of differences in participants' personality.

Specifically, participants spoke fewer words on average when the RA was in the lab room during the interaction with the CA, showing that they felt less comfortable with another person nearby. This "social context" effect on the number of words spoken was statistically significant when considering only the first session (Figure 5), when all participants were equally unfamiliar with the CA and were interacting with the unadapted CA, with the mean number of words spoken when the participants were alone (M = 404, SD = 220, n = 15) nearly double that spoken when the RA was present (M = 223, SD = 92, n= 16), independent samples Mann-Whitney U Test, p = .015. This effect on number of words spoken was also significant during the final session, when all participants were acclimated to the CA and the CA was adapted to participants, M_{alone} = 155, SD = 125, n = 12, $M_{RA\ present}$ = 87, SD = 78, n = 14, independent samples Mann-Whitney U Test, p = .031. Similarly, participants tended to use the word "I" slightly more often when alone; this difference, although not large, was statistically significant in the first session (M_{alone} = 7.9, $M_{RA\ present}$ = 5.52), Mann-Whitney U Test, p = .002, indicating that participants talked more naturally and confidently when alone, especially early on in their interactions with the CA. In contrast, the median for positive emotional tone tended to be somewhat higher when the RA was present in the lab room (median across all sessions = 93.32) than when the participant was alone (median across all sessions = 73.00). A possible explanation is that people tend to hide their negative side in front of others, especially strangers. The authenticity and clout scores were similar in both conditions. The median frequencies of

motivational drives were similar in both conditions too, but the variability appeared to be somewhat greater when the RA was present in the room (interquartile range = 3.86, SD = 2.78) than when participants were alone (interquartile range = 1.42, SD = 1.94). This may be due to participants with different personality reacting differently when there is an observer around: a more open and extraverted person may talk more, and a more introverted person may talk less in front of another person.

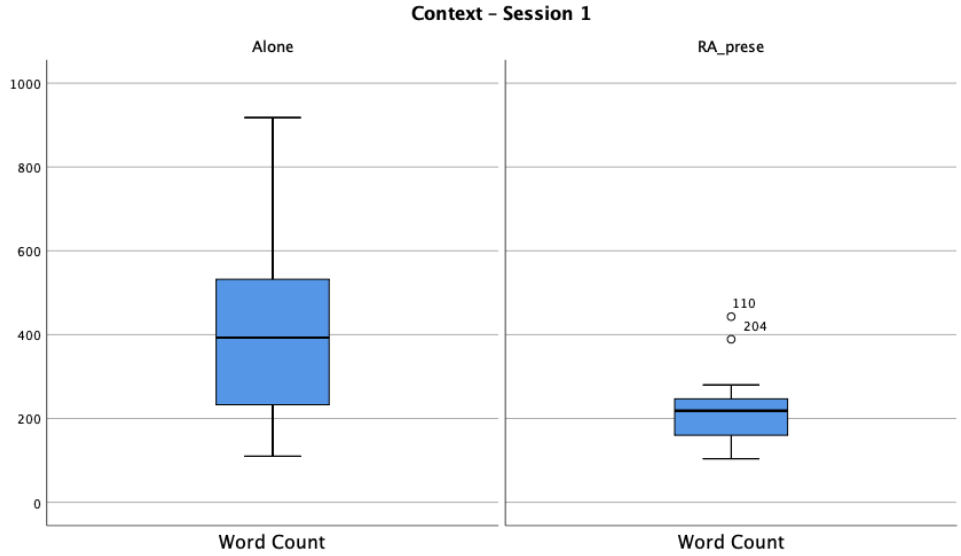

Figure 5. Participant Word Count without an observer present ("Alone") and with an observer present ("RA present"), in the first session

Conclusion

The generally high System Usability scores indicate that the CA we used for the study has a good communication system and is easy to use. The sociolinguistic analysis suggests that participants were more satisfied with user experience after the continued familiarization with the CA and after the CA adapted to participants. Additionally, participants tended to be more relaxed and willing to reveal themselves when interacting with the CA without an outside observer being present. These findings show that users prefer the individualized CAs, and are more comfortable after becoming more extensively familiarized with the system. The results are encouraging and inspiring for the next phase of experiment on older participants in that we would continue to use this experimental CA, have the participants interact with the CA alone, and ask participants to interact with the CA for multiple times to give time for extended user familiarization.

References

Fadhil, A., Wang, Y., & Reiterer, H. (2019). Assistive Conversational Agent for Health Coaching: A Validation Study. Methods of Information in Medicine, 58(1), 9–23. Doi: 10.1055/s-0039-1688757

Ferland, L., Huffstutler, T., Rice, J., Zheng, J., Ni, S., & Gini, M.L. (2019). Evaluating Older Users' Experiences with Commercial Dialogue Systems: Implications for Future Design and Development. *ArXiv, abs/1902.04393*.

Ferland, L. & Koutstaal, W. (2020). How's Your Day Look? The (Un)Expected Sociolinguistic Effects of User Modeling in a Conversational Agent. CHI'20 Extended Abstracts, April 25–30, 2020, Honolulu, HI, USA ACM 978-1-4503-6819-3/20/04. Doi: 10.1145/3334480.3375227

Pennebaker, James W., Booth, Roger J., Boyd, Ryan L. & Francis, Martha E. (2015). *Linguistic Inquiry and Word Count: LIWC 2015 - Operator's Manual*. Retrieved February 19, 2020, https://s3-us-west-2.amazonaws.com/downloads.liwc.net/LIWC2015_OperatorManual.pdf

Portet, F., Vacher, M., Golanski, C., Roux, C., & Meillon, B. (2013). Design and evaluation of a smart home voice interface for the elderly: Acceptability and objection aspects. *Personal and Ubiquitous Computing, 17*(1), 127–144. Doi: 10.1007/s00779-011-0470-5

Riek, L. D. (2012). Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines. *Journal of Human-Robot Interaction, 1*(1), 119–136. https://doi.org/10.5898/JHRI.1.1.Riek

System Usability Scale (SUS). Retrieved August 1, 2020, from https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html.

Sauro, Jeff. (2011, February 2). *Measuring Usability with the System Usability Scale (SUS)*. https://measuringu.com/sus/.

Appendix A

System Usability Scale Questions

(1) I think that I would like to use the Scheduling Assistant frequently.
(2) I found the Scheduling Assistant unnecessarily complex.
(3) I thought the Scheduling Assistant was easy to use.
(4) I think that I would need the support of a technical person to be able to use the Scheduling Assistant.
(5) I found the calendar and stress-related functions in the Scheduling Assistant were well integrated.
(6) I thought there was too much inconsistency in the Scheduling Assistant.
(7) I imagine that most people would learn to use the Scheduling Assistant very quickly.
(8) I found the Scheduling Assistant very awkward to use.
(9) I felt very confident using the Scheduling Assistant.
(10) I needed to learn a lot of things before I could get going with the Scheduling Assistant