Assessing conversational agent interaction style preference in children for information retrieval tasks

Suyash Sharma
Bachelor Advanced Technology
University of Twente
Enschede, The Netherlands

Abstract

This bachelor assignment aimed to assess the preference that children aged 9 to 12 had in the interaction style of a conversational agent while they were attempting to complete information retrieval tasks. This research used a Furhat robot[1] as a conversational agent to interact with the children participants to conduct experiments. The participants were given search tasks and asked to talk to the conversational agent to complete them. The conversational agent would talk back to them by asking questions and would print out relevant search results to a screen next to it, if it recognized a keyword that was said by the participant. The participants interacted with two different versions of the conversational agent. Both the versions had a different style of asking clarification questions to the participants. After every interaction, the participants were asked to assess their interaction by filling out a questionnaire. The questionnaire asked questions related to the robot's likability(which were taken from the popular Godspeed questionnaire) and also asked them to give an overall rating to the each version of the robot they talked to. After the participant had completed interacting with both versions of the robot, a short interview was conducted by the principal researcher where they were asked to describe their experience with the robot qualitatively in a series of questions. The results of the questionnaires and interviews were then used by the researcher to draw conclusions about the interaction style preference that the children had.

1 Introduction

1.1 Motivation

Despite being digital natives, most children today face difficulties trying to access information on the internet. This is in large due to the fact that current search engines were mainly designed keeping adults in mind - adults who can formulate and communicate queries effectively. There exist clear differences in how adults and children approach searching for information. Research shows that children tend to "make more web moves, loop searches and hyperlinks, backtrack and deviate more often from their target" do not follow the more linear approach that adults do[1]. Children also tend to have less knowledge to base recall than adults. In addition to this, their spelling, vocabulary, punctuation, and grammar skills are still developing. All these factors affect their abilities to form queries that are coherent enough [1][2].

Interviews conducted with children aged 9 to 12 conducted by Druin et al [3], mentioned issues such as "not being good at writing" and that it is difficult to "find the right words to put in the box" when writing queries on search engines[3].From research, it is seen that children tend to resort to using a full natural language query when using search interfaces for query needs, rather than using specific keywords like adults[1]. Presently, methods to assist with query formulation for information retrieval for children include query expansion and query suggestions. But in current search interfaces, these methods are still geared toward adults. While there exist search environments such as Kidrex, Kiddle, and Kidzsearch [2], which are made specifically for children, they do not assist greatly with query formulation but are rather focused on curating information and content aimed at children. Upon inspection, it is evident that the sites Kidrex and Kiddle provide little to no query suggestions and Kidzsearch provides only query completion tools similar to Google. But all these sites do not provide solutions for some of the more common difficulties that children face while formulating queries like misspelled query terms and long natural language queries, which has been reported by the work of Fails et al[2].

---

[1] url to Furhat Robotics website https://furhatrobotics.com/

A lot of the issues that children face can be explained by child development psychology. The work of Jean Piaget in this field has been of great importance. In his research, Jean Piaget talks about the concrete-operational stage (from ages 7 to 11 or 12) and formal-operational stage(ages 11 or 12 and beyond) in children[4]. In the concrete-operational stage, children tend to show a trial and error approach to solving problems and start to think more logically[5]. It is only after the children reach the formal-operational stage that they "have the ability to formulate, test and discard a whole range of possible solutions to a problem until a possible solution is found"[6]. Before this stage, they are still developing their problem solving capabilities. This points to the need for them to be assisted with their problem solving. This is also supported by research by Gossen et al. [7] which shows that children pay more attention to tools such as suggested queries than adults.

Using a conversational agent could prove to be a good interface that children could use to retrieve information as proposed by the work of Landoni et al.[8] and further recommended by Beelen et al[9]. This would help children overcome a lot of the issues that are also mentioned above like, grammar spelling errors, punctuation etc. They would also be communicating their needs in natural language, which from research is shown to be their preference. But most present-day conversational or voice agents currently have not been able to improve the experience of children greatly. Beelen et al. describes how these agents provide "simple question-answer style" interactions, that are not conversational, and may actually be hinder the children's' experience, not serving their appropriate needs[10]. As mentioned above before, children tend to pay more attention to suggested queries which points to the fact that they do not feel that their initial queries are adequate enough to retrieve the information that they are seeking. Thus, a more useful interaction style with a conversational agent posing clarification questions to the child might be helpful to form a query and retrieve information.

1.2 Problem Statement

As mentioned before, modern-day search engines are not catered to help young children formulate their queries to complete information retrieval tasks. Children tend to have difficulties expressing their query needs and thus an interface that by posing clarification questions may prove to be helpful to them. However it is not yet certain what type of interaction-style children would prefer. The interaction-style could be of the following two types- 1) a system where the interface that they use prompts them to explain their query needs more, requiring them to be more proactive, making them lead the interaction or 2) an interface that is more proactive in making suggestions and so, leads the interaction while having the user react more to the system. Thus, based on the setup, the type of clarification questions that each system would ask will have to be different. One system will have to ask more open-ended clarification questions to engage the user more and be more user-driven and the other system will have to ask more close-ended questions and make query suggestions to be more system-driven.

1.3 Research question

This problem statement above gives rise to the following question-
What interaction style do children aged 9 to 12 prefer while talking to a conversational agent to formulate a query to complete their information retrieval needs-

- *A system that uses open-ended clarification questions to create a more user-driven interaction? Or*
- *A system that uses close-ended clarification questions by asking fixed query suggestions to create a more system-driven interaction?*

2 Related work

While there is not a lot of work that has been carried out on studying which style of clarification questions are preferred by children for retrieving information, there is well cited literature available on the topic of query formulation assistance for children [2]. The topic of generating clarification questions for general users (no-specific age group) has been reported on extensively by Zamani et al. where models for generating clarification questions have been studied and developed[11]. In addition to this, the Qulac database has been developed for generating relevant clarification questions [12][2]. But, as mentioned earlier, no research has been done into investigating generating clarification questions for children specifically. The idea of using conversational agents for information retrieval amongst children has been an idea that has been introduced in several authors such as Ladoni et al. and further recommended by Beelen et al.[9][8].

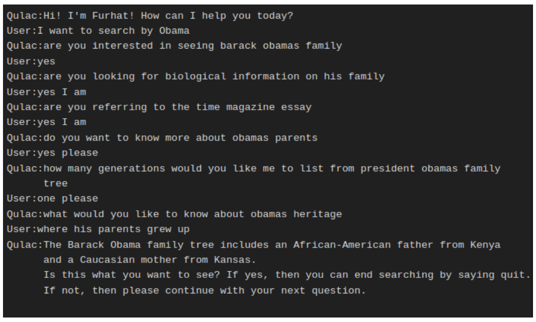
3 Methodology

3.1 Design

The research question was answered through the means of an experiment. The experiment was conducted the following way. The children were told to carry out two search tasks(finding factual information about two different topics) with two separate systems of the conversational agent(CA)- one system where the clarification questions were more open-ended and relied more on user input to generate the next questions and the other system where the clarification questions were fixed and close-ended and did not take in user input to generate the next set of questions. For this, a within-subject study was set up. Under this method, the participant would test all conditions of the CA, but with different topics. This was done to eliminate any bias that may have come up by having the participant test with just one topic. The two different conditions were created using two different algorithms, generating the different types of clarification questions. These two different systems were called the Qulac-based system version and the ELIZA-based system version. The experiment will take place in a closed knowledge-domain system for the sake of simplicity.

3.1.1 Qulac-based system

The system version that used a fixed question and close-ended approach made use of the Qulac database, “a database on asking Questions for a Lack of Clarity in a lack of clarity” [12]. The Qulac database contains data-sets with clarification questions for several of the topics. The questions in the data-set had been collected and stored through crowd-sourcing with the help of the TREC Web Track 2009-2012 collections. Most of the topics in the Qulac database had about 13-14 pre-existing clarification questions, stored in a JSON file with a unique topic ID and a unique question ID for every question. These questions were used to pose clarification questions to the participant during the interaction with the help of a Python script.



```
Qulac:Hi! I'm Furhat! How can I help you today?
User:I want to search by Obama
Qulac:are you interested in seeing barack obamas family
User:yes
Qulac:are you looking for biological information on his family
User:yes I am
Qulac:are you referring to the time magazine essay
User:yes I am
Qulac:do you want to know more about obamas parents
User:yes please
Qulac:how many generations would you like me to list from president obamas family
      tree
User:one please
Qulac:what would you like to know about obamas heritage
User:where his parents grew up
Qulac:The Barack Obama family tree includes an African-American father from Kenya
      and a Caucasian mother from Kansas.
      Is this what you want to see? If yes, then you can end searching by saying quit.
      If not, then please continue with your next question.
```

Figure 1: Example of user interaction with Qulac version of the conversational agent from the research

3.1.2 ELIZA-based system

The system version that asked open-ended questions and relied on user-input was based on the popular ELIZA chatbot algorithm. The ELIZA algorithm was a psychotherapist chatbot developed in the 1960s based on the Rogerian school of psychotherapy. The goal of this program was to hold a natural language conservation and act as a therapist to the person it was talking to. The algorithm generated rule-based responses that it drew from a “doctor” script. This script had pre-written responses written for specific keywords. So the ELIZA algorithm would take user input, identify the keywords, assign them a numeric weight, and then match those keywords to those in the doctor script and return the pre-written responses in the doctor script. The algorithm for the purpose of this study was also designed in a similar way. An ELIZA algorithm written by GitHub user *wadetb* was modified to develop the ELIZA based algorithm.[3]There were pre-written responses written for specific keywords relating to the topics. The responses were rewritten to pose clarification questions that were more open-ended (For example- if the user says “I want to search about dinosaurs”, the algorithm would reply with a response like “what exactly do you want to search about dinosaurs?”, prompting the user to explain their query more.)

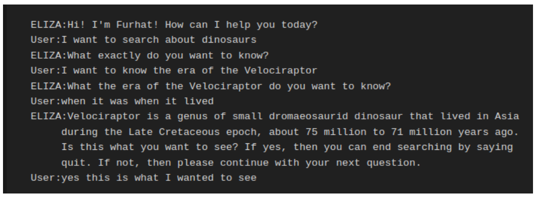
3.1.3 Topics

As mentioned before, two different topics were picked for the participants to work with the different test conditions. The two topics thus chosen

---

[2] URL to GitHub page of Qulac database https://github.com/aliannejadi/qulac

[3] URL of Github page of python implementation ELIZA chtabot by user *wadetb* https://github.com/wadetb/ELIZA



```
ELIZA:Hi! I'm Furhat! How can I help you today?
User:I want to search about dinosaurs
ELIZA:What exactly do you want to know?
User:I want to know the era of the Velociraptor
ELIZA:What the era of the Velociraptor do you want to know?
User:when it was when it lived
ELIZA:Velociraptor is a genus of small dromaeosaurid dinosaur that lived in Asia
     during the Late Cretaceous epoch, about 75 million to 71 million years ago.
     Is this what you want to see? If yes, then you can end searching by saying
     quit. If not, then please continue with your next question.
User:yes this is what I wanted to see
```

Figure 2: Example of user interaction with ELIZA version of the conversational agent from the research

were- Barack Obama and Dinosaurs. The reason these were used was because these topics had existing clarification questions in the Qulac database and were also considered to be child-friendly topics. The search tasks that the children were given to complete were the following-

- Obama task
  Please find some information about Barack Obama and his personal family history.
- Dinosaur task
  Please find information about what era the velociraptor dinosaur lived in and what it ate.

Both the system algorithms were designed to generate the clarification questions and also recognize a few keywords related to the tasks and print out search results onto a screen that was placed next to the user. For example, keywords like “parents”, “family tree”, and “ancestry” were programmed to be recognized for the Obama topic and keywords like “age”, “habitat”, and “food” were programmed to be recognized for the Dinosaur topic. Once one of these keywords were spoken by the user and recognized by the algorithm, a corresponding snippet of information, relevant to the keyword, would be displayed on the screen, as a search result. This was done to provide the users some positive feedback during the interaction. These information snippets were all found using the website *resultstoexcel.com* [4], which displays search results for a query and allows you to download the first 10 search results to an excel file. The downloaded excel file included a lot of columns containing information like the first few lines of text on the website of the search result, the url of the search result and the name of the website. The column containing the first few lines of text from the website’s search results were printed out onto a screen next to the user by the systems’ algorithms when they were recognized.

As mentioned before, the system was completely closed-knowledge domain. Thus, it was kept offline.

3.2 Participants

A total of 35 participants (23 boys and 12 girls) were recruited for the purpose of this experiment. The ages of the participants ranged from 9 to 12 years. Several schools across the Netherlands were reached out to via email or phone to establish contact and recruit participants. Schools that were contacted were either international schools (where the language of instruction was English) or bilingual schools (where the language of instruction was both Dutch and English), as the researcher was proficient only in English and needed to recruit participants who were also proficient in English. In the end, two schools, one in The Hague and one in Amsterdam, agreed to participate in the research. There were 35 participants recruited in total, 17 being from the school at The Hague and 18 from the school in Amsterdam. Prior to participating in the research, all the candidates had to get consent from their parents and/or guardians. The consent was granted through signing consent forms that were prepared by the researcher. The parents and/or guardians were sufficiently informed about the scope of the research and what their children would be doing in the research through the means of an information brochure, which was also prepared by the researcher. Through the consent form, permission to record data and audio was also taken. The information brochure and consent forms had been reviewed and approved by the Ethics Board of the University of Twente, prior to their distribution.

3.3 Conversational agent

The conversational agent used to carry out the study was a Furhat robot. [5] The Furhat robot is a social robot that possesses the ability to display “human-like expressions” and can carry out advanced conversational capabilities. The algorithms for both system versions were written in Python and then made to run on the Furhat robot using its RemoteAPI tool[6]. The Furhat robot would take input by listening to the user,

---

[4] URL to website https://resultstoexcel.com/

[5] See footnote 1

[6] URL to Furhat RemoteAPI documentation https://docs.furhat.io/remote-api/

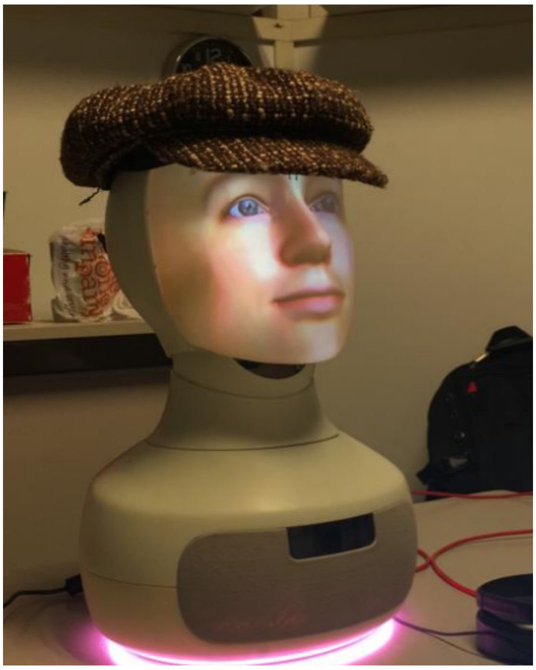


Figure 3: Furhat Robot

convert the speech to text, run the ELIZA or Qulac algorithm, generate a response based on the algortihm's program and then convert that response back from text to speech for the user to hear.

3.4 Pre-experiment procedure

Prior to participating, the participants and their legal guardians were informed about the contents of the research through an information brochure and their parents/legal guardians were requested to sign consent forms if they wanted their child to participate in the research. They were also asked if they consented to having their child's data and audio recording collected during the research. They were also made aware that there was a STOP procedure in place and that their child was able to stop or end the participation whenever they wanted. All the steps of the research were presented to the Ethics committee of the University of Twente and were conducted with their approval.

3.5 Experiment procedure

The experiment procedure was conducted in the following way. The participant was invited into the room and requested to sit on the seat facing the Furhat. The primary investigator (PI) explained to the participants that the goal of the session was to talk to the robot to complete a search task that would be given to them. The participant was informed that the robot was not "very smart" and so they should try to be as clear and expressive as possible when talking and asking questions to the robot. The PI also informed the participant that if the robot found a search result, it would display it on the screen next to the robot. The PI then asked the participants if they had any questions or doubts. The PI also informed them that the participants were free to end the session at any point, using the STOP procedure. If the participants had no question, the interaction session would begin. The PI provided them with the search task and instructed them that they had five minutes to talk to the robot to complete the task. The PI requested the participant to start with a specific statement to begin the interaction with the robot- if the task was about Obama, this sentence was “I want to search about Obama” and if the task was about dinosaurs, the sentence was “I want to search about dinosaurs”. Once the participant was ready, the PI started the timer and allowed the participant to begin speaking with the robot. The PI would also start recording the audio of the interaction if the parents/guardians of the participant had consented to it and if the participant was also comfortable with it. If the participant was not comfortable with being recorded, no audio recording would be taken, despite the consent that the parents/guardians had granted. After the participant was done with the interaction session, they were asked to fill in a questionnaire to evaluate their experience with the robot. The form contained questions in the form of a Smiley Face Likert scale. The choice to use a Smiley Face Likert scale was made after referring to well cited literature supporting it as a good way of having children make evaluations.[13] [14]. Once they had completed the form, they had a short interview conducted where they were asked to answer questions qualitatively about their experience with the robot. This marked the end of the first session. After a 2-3 minute break, this procedure was repeated. This time, the participant would interact with a different version of the robot and would be given a different search task to work on. At the end of the second session, they would be asked to fill out another questionnaire and would have another qualitative interview conducted. At the end of the second session, they would be asked which version of the robot they preferred. This marked the end of the experiment. This entire procedure was repeated for every participant. Fig.4 shows a step-by-step flowchart of the experiment procedure. The versions and search tasks were determined in advance by the PI. There were a total of four combinations of topics and systems decided on the for the experiment It was as follows-

- Combination 1- ELIZA system with the Obama topic first and Qulac system with the Dinosaur topic second
- Combination 2- ELIZA system with the Dinosaur topic first and Qulac system with the Obama topic second
- Combination 3- Qulac system with the Obama topic first and ELIZA system with the Dinosaur topic second
- Combination 4- ELIZA system with the Obama topic and Qulac system with the Dinosaur topic

This order was repeated after every four participants.

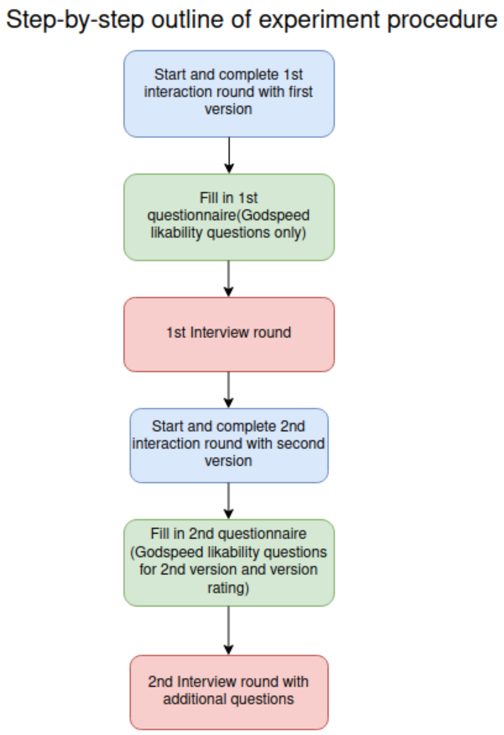

Step-by-step outline of experiment procedure
Start and complete 1st interaction round with first version
Fill in 1st questionnaire(Godspeed likability questions only)
1st Interview round
Start and complete 2nd interaction round with second version
Fill in 2nd questionnaire (Godspeed likability questions for 2nd version and version rating)
2nd Interview round with additional questions

Figure 4: Step-by-step proceducre of the experiment

3.6 Post-experiment procedure

After the experiment concluded, the participants were asked if they had any questions. After that, they were de-briefed about some deception that had been used for the sake of the research. They were made aware of the fact that the CA was made to appear “smart”, by having them ask questions on the specific topics of Barack Obama and dinosaurs. They were informed that if the robot had asked them questions about any other topics, it would have not been able to respond the same way and would not fetch results.

3.7 Measurements

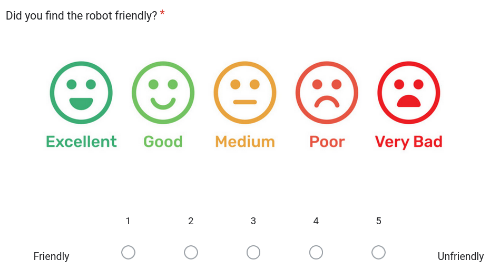
The participant responses of two questionnaires were used to carry out the quantitative analysis of the research. The first questionnaire included questions about the first version the participant tested. These questions had been taken from the likeability section of the Godspeed questionnaire, a questionnaire which was developed to measure users’ perception of robots [15].The questions could be answered in the form of a 5-point smiley face Likert scale [14].An example of the type of question from the first questionnaire can be seen in Fig. 5. The second questionnaire asked the participants to fill out the likability section from the Godspeed questionnaire for the second version of the CA they interacted with. The second questionnaire had two additional questions, which asked participants to rate each version they interacted overall on a scale of 1 to 10 with on a smiley face Likert scale. Fig. 6 shows the way the participants were asked to fill in the overall rating for each version. After the participant completed filling in the questionnaires, they had a short interview conducted. The interview part of the session was done to collect qualitative results. The following questions were asked in the interviews-
Interview questions- After each robot

- What did you think of this version of the robot? Why?
- Would you like to use this robot to search for information online? Why or why not?
- What did you not like about the robot?
- Would you want to change something about the robot? If so, what?
- Did you not like the topic or the version of the robot that you used?

Interview questions- After seeing both robots

- Which version of the robot did you prefer? Why?
- Did you notice a difference in the way the robot talked to you and asked questions in both the versions?

Audio recordings of the session were taken if the participant and their legal guardian had given consent. Other than that, the algorithm running logged all the things that were said in the sessions in a text file and recorded the rounds of interactions as well.



Did you find the robot friendly? *

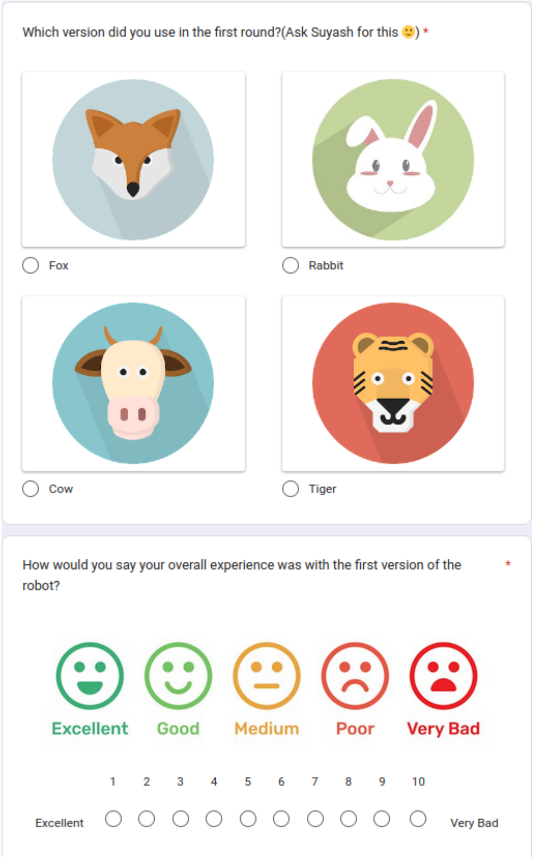
Excellent Good Medium Poor Very Bad

	1	2	3	4	5	
Friendly	○	○	○	○	○	Unfriendly

Figure 5: Example of question from the questionnaire assessing likability using smiley likert scale

3.8 Analysis

As mentioned above, the data collected during the research was both qualitative (interview questions) and quantitative (questionnaire responses). The quantitative results were processed by using the responses from the questionnaires and turning them into numeric scores. There were two main data-sets prepared- 1) a data-set containing an averaged out scores from the likability test for each version tested by the participant and 2) a data-set containing the overall version scores that the participants filled in the second questionnaire. These scores were grouped by version type(ELIZA or Qulac).Then a Wilcoxon Signed Rank Test was performed on both these data-sets to determine whether there was a statistically significant preference for either version[16]. There were 70 total scores with 35 scores for the Qulac and ELIZA version each.
The qualitative data collected was analyzed by performing a thematic analysis. The responses of the interviews were also used to calculate a percentage statistic for the preference in versions by the participants.



Which version did you use in the first round?(Ask Suyash for this 🙂) *

○ Fox
○ Rabbit
○ Cow
○ Tiger

How would you say your overall experience was with the first version of the robot? *

Excellent Good Medium Poor Very Bad

	1	2	3	4	5	6	7	8	9	10	
Excellent	○	○	○	○	○	○	○	○	○	○	Very Bad

Figure 6: Example of question asking user the overall score for a system version (Fox(Eliza-Obama), Cow(Eliza-Dinosaur), Tiger(Qulac-Obama), Rabbit(Qulac-Dinosaur)

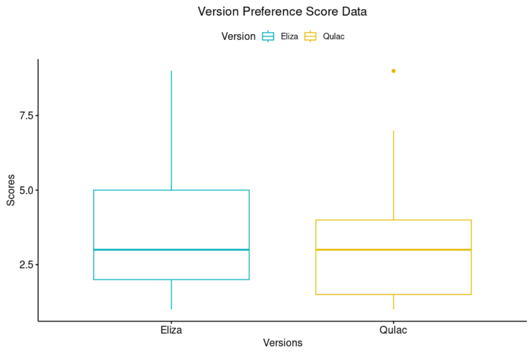


Figure 7: Boxplot data visualization for Preference of versions

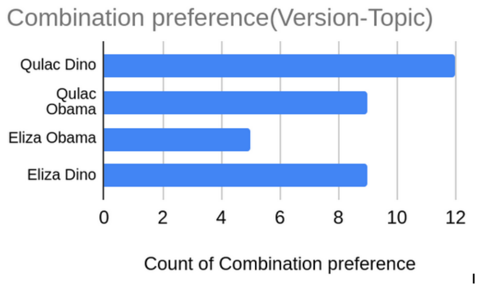


Figure 8: Version preference (based on responses in the interview)

4 Results

4.1 Results for overall version rating scores

The final rating scores that participants(N=35) gave to both versions of the CA was used to analyze the version preference. The data distribution of scores for both versions can be seen in the box plot in Fig7, which was obtained using the *ggpubr* package in R. The median score value for the ELIZA version(*median*=3) was the same as that of the Qulac version(*median*=3). Since the data collected from the participants is paired, a Wilcoxon Signed Rank Test[16] was carried out in the R software to see if there was a statistically significant preference in the version that the participants used. Using the *wilcox.test()* function with the data gave a result of $T = 254.5, z = -1.168, p = 0.2427$. Since the p-value is more the $\alpha = 0.05$ value, there is no statistically significant difference in the preference scores of the versions

In addition to the questionnaire, the participants were also asked a few questions in an interview after every interaction. After they had interacted with both versions, they were asked which version they preferred talking to and which topic they preferred talking about. The results from the interviews show that 60%(21 out of 35 participants) preferred the Qulac version and 40%(14 out of 35 participants) preferred the ELIZA version. The combination preference between the four possible combinations was highest for the Qulac-Dinosaur version (with 34.2%), as can be seen in the Fig. 8 above.

4.2 Results for likability scores from the Godspeed Questionnaire

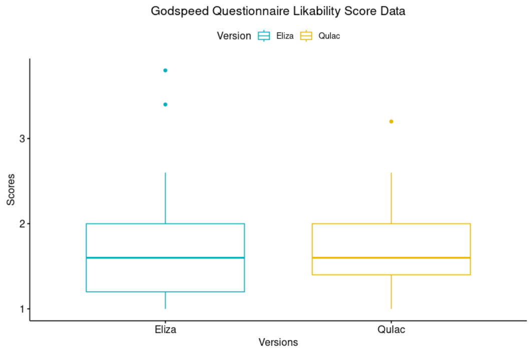


Figure 9: Boxplot data visualization for Likability Score of versions

Scores from the likability section of the questionnaire were used to determine if there was a difference in likability between the two versions. The likability data for both versions can be seen in the boxplots in Fig.9 which was also obtained using the *ggpubr* package in R. The median score value for likability for the ELIZA version(*median*=1.6) was again the same as the Qulac version(*median*=1.6). Since the data for likability was also paired, the Wilcoxon Signed Rank Test[16] was carried out in the R software to see if there was a statistically significant preference in the likability between the versions. Using the *wilcox.test()* function with the data gave a result of $T = 244, z = -0.2269, p = 0.8205$. Since the p-value is much higher than the $\alpha = 0.05$ value, there is no statistically significant difference in the likability scores of the versions

4.3 Thematic Analysis

A thematic analysis was performed the results of the participants(N=35) interviews to identify recurring themes and patterns related to their interaction experience with the CA. The following themes were the most common from the responses.

4.3.1 Ease of use

Most of the participants (97%; 34 out of 35 participants) mentioned the fact that they would prefer talking to at least one version of the conversational agent like Furhat, over typing out information using an iPad or a laptop. Among the 97% of participants, 71%(25 out of 35 participants) said they would prefer talking to both versions of the conversational agent over typing. They gave several reasons for it, including- it being “quicker than typing” and that it asked them questions back. A few participants(11.4%; 4 out of 35 participants) also mentioned that they felt as though they had a friend they were talking to, who was helping them out. There were participants who had been diagnosed with ADHD who mentioned that searching for information online was harder using a computer, and if the CA was made smarter, it would be “easy” for someone with ADHD to find information by speaking to a CA. A couple of participants (5.7%; 2 out of 35 participants), whose first language was not English, said that they would prefer speaking to the CA over typing as they were not fluent in English yet. While most participants said they preferred using the CA to searching for information through a computer, there were a few exceptions. A few of the older participants said they preferred the CA less than using computers. Their reasons for this included them saying they could find images and search through different websites if they just used a computer.

4.3.2 Difficult words

A few participants (11.4%; 4 out of 35 participants) reported back that the CA sometimes used words like “query” that were a little complicated for them to understand. This influenced the way they felt about the CA and gave lower scores in the questionnaires to the versions of the CA where they encountered more of these words.

4.3.3 Redundant questions(ELIZA and Qulac)

In the interview, a few participants(8.57% ; 3 out of 35 participants) reported that the versions of the CA sometimes kept asking them questions, despite them having stated what they wanted. They sometimes found this bothersome. Participants said that they would have preferred it if the CA had "listened better".

4.3.4 Didn’t understand what I wanted to search(Qulac)

An issue that a few participants (17.1%; 6 out of 35 participants) reported about the Qulac version of the CA was that it was difficult for it to understand what they were saying. They reported that instead of responding to what they wanted, the version would “keep jumping to other questions” and “go further away”. Some participants stated that they felt the system was “unresponsive” to what they were saying.

4.3.5 Understood what I wanted to say(ELIZA)

A few participants(5.71%; 2 out of 35 participants) stated a preference for the ELIZA version of the CA as it asked simpler questions and its “replies were clearer”. Some participants said during the interview that they were able to “make more of a connection” with the ELIZA version of the CA.

4.3.6 Random/Vague questions (ELIZA)

While there were participants that enjoyed talking more to the ELIZA for its simplicity and clarity, a lot of other participants (14.2%; 5 out of 35 participants) directly mentioned that they found it asking a lot of vague and confusing questions. Participants reported that the ELIZA system sometimes kept repeating itself. They also talked about how the system can “distract” and go off-topic often.

4.3.7 Prompts good questions(Qulac)

Participants that liked the Qulac version more said that they liked how the Qulac system prompted them more questions(5.71%; 2 out of 35 participants). They said that they liked how it “asked more specific questions” than the ELIZA version, did not go “off-track” and asked questions about the topics they wanted to talk about. They found the clarification questions that the system suggested helpful as it would ask different kinds of questions around the same topic.

4.3.8 Other

Other comments from the interviews included how the participants preferred that the system printed out just one search result instead of having to scan through thousands of search results like Google. When participants were asked in the interviews whether they noticed a difference in the two versions of the CA, only a few(2 out of 35 partcipants) could correctly notice the difference.One of the participants who did notice a difference said that the Qulac version of the CA *"went further"* with its questions, and the ELIZA version *"came back"* to the participants question. The partcipant who noticed this difference also said that they would like to use the Qulac version to find out more about a broad topic(they gave the example of US presidents) as the Qulac system asks several relevant questions. And they would prefer using the ELIZA version for specific information (they mentioned “in-depth searching”) as it would ask more open-ended questions that would prompt the participant further.

5 Discussions and Conclusion

5.1 Discussion

The results from the study show that participants in this study have a preference for the Qulac version of the CA, as 60% of the participants said that they preferred it over the ELIZA version. But these results were not statistically significant, as shown by the results of both of the statistical tests performed for this study. However, it can be seen that there were a few reasons that the participants preferred the Qulac version over the ELIZA version, which included- *prompting good query suggestions relevant to the topic in the form of clarification questions, staying on topic of the search task* and *not repeating itself.* But there were reasons why not all participants preferred the Qulac version. These included reasons such as *them not feeling heard* by the Qulac version and the Qulac version asking the participants *long, complexly-worded questions*. Participants that did state a preference for the ELIZA version more reported feeling more *"heard"* and found its questions *easier and simpler*. Thus, it can be seen that both versions have their strengths and flaws. Both versions proved to be useful in different ways.

5.2 Limitations

The preference for the versions was influenced directly by how the algorithms for the different versions were programmed. The algorithm for the Qulac version was much simpler than the ELIZA version, as the Qulac version had a set number of clarifications questions that it asked every time, regardless of what the user said. The ELIZA version was more intricate in its programming. Its responses and overall level of interaction could have been improved if more time was spent working on the algorithm and if more pre-testing had been done, before conducting the sessions with the participants.

Another point of improvement would be the Obama search task. Several participants did not know who Barack Obama was. Although they were given enough information about Obama before starting the search task, many participants stated that they would have preferred another topic to search information on. This is reflected in the data too, where the Obama versions are the versions with the lower preference,as is shown in Fig.8. Topics that participants would have preferred instead of the Obama topic included geography, outer space, climate change and the environment. As one participant put it, a more “universal” topic may have been better suited as there are “a lot of people who do not know who Obama is outside the United States.” Other than the topic, the search task could have also been framed to be something simpler. Few participants struggled with understanding the term “personal family history” from the Obama task description.

All the participants were asked to start off their queries by either saying*"I want to search about Obama"* or *"I want to search about dinosaurs"*. This was done to start off the interactions in the right direction, with the focus on the topic of the search task. But doing this did influenced the naturalness of the conversation that the participants had with the CA.

Another limitation was the search results displayed influencing the final preference participants had with the CA versions. There was one instance where a participant preferred one version of the CA over the other because it returned more search results. This however was not always the case. There were cases where participants got search results with one version of the CA and no search results with the other and still ended but preferring the version of the CA with no search results, due to the interaction style it had.

5.3 Further Research

Improving the ELIZA algorithm by getting rid of some of the issues mentioned before could lead to a different outcome if this study was repeated again. Also,developing an algorithm that combines the best parts of both the versions might be interesting to explore. The algorithm for example, could alternate between the ELIZA style of questions and the Qulac style of questions. The performance of this algorithm could then be compared against the performance of Qulac and ELIZA versions in assisting children in query formulation and information retrieval.

As discussed in the thematic analysis, a participant suggested that they would prefer using the Qulac version for exploring more broad topics that they are unfamiliar with (giving the example of American Presidents) and the ELIZA version for more specific information to do more "in-depth searching". A study could be set-up to research this- whether children have a specific preference for the version of CA(Qulac vs ELIZA) for different types of information retrieval tasks (broad vs specific).

5.4 Conclusion

The goal of this bachelor assignment was to investigate the preference in interaction-style of a conversational agent that helps the children complete information retrieval tasks. Through the means of an experiment, it was determined that the participants of this research study showed a preference for the Qulac version, a system-driven version that posed closed-ended clarification questions in the form of query suggestions relevant to the search task, over the ELIZA version, a user-driven version that posed open-ended clarification questions. Although these results were not statistically significant, the findings from the research provide many insights into the different aspects that children find important when they interact with a conversational agent to complete information retrieval tasks.

Acknowledgements

First and foremost, I would like to thank my loving parents, ***Shruti*** and ***Yatendra***, for everything they have done for me. They work incredibly hard to give me the opportunities I have.

To my loving sister, ***Yasshita***, thank you for the joy you bring in my life.

To my aunt and uncle, ***Smriti*** and ***Suraj***, thank you for your unconditional love and support.

To my friends ***Jorik*** and ***Luukas***, thank you for all the support and words of encouragement that you provided me during the period of my bachelor thesis. To my friends ***Pablo***, thank you for helping me out with all the technical issues I had at the start of my thesis. You saved me a lot of time and trouble.

To my friends ***Anshul, Prakash*** and ***Marco***, thank you for letting me stay at your house in Rotterdam when I needed to travel to perform my user studies. I will always remember your hospitality.

To all my friends at ***DesignLab***, my home away from home, I am forever grateful for your friendship and support.

To ***Miss Flora Dullea, Miss Sara Vermeulen, Miss Cara Howe*** at the HSV International Primary School (NSL) and ***Mr. Michael McGlade*** at the International School of Amsterdam, thank you so much for giving me the opportunity to work with your schools. This research would not have been possible without you.

To all the ***student participants*** from the HSV International Primary School and the International School of Amsterdam thank you for helping me out with this research.

To my old professors, ***Prof. Aveek Bid, Prof. Vivek Venkatachalam*** and ***Prof. Toyoko Orimoto***, thank you for all the guidance and opportunities you gave at the start of my university career. I will forever be grateful for it.

To ***Prof. Khiet*** and ***Thomas***, thank you for all the support, guidance and encouragement and patience you have shown me throughout the course of my thesis. Working on this thesis under your supervision has been the most rewarding experience of my university career.

References

[1] Hanna Jochmann-Mannak, Theo Huibers, and Ted Sanders. Children's information retrieval: Beyond examining search strategies and interfaces. In *Proceedings of the 2nd BCS IRSG Conference on Future Directions in Information Access*, FDIA'08, page 8, Swindon, GBR, 2008. BCS Learning amp; Development Ltd.

[2] Jerry Alan Fails, Maria Soledad Pera, Oghenemaro Anuyah, Casey Kennington, Katherine Landau Wright, and William Bigirimana. Query formulation assistance for kids: What is available, when to help amp; what kids want. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*, IDC '19, page 109–120, New York, NY, USA, 2019. Association for Computing Machinery.

[3] Allison Druin, Elizabeth Foss, Leshell Hatley, Evan Golub, Mona Leigh Guha, Jerry Fails, and Hilary Hutchinson. How children search the internet with keyword interfaces. In *Proceedings of the 8th International Conference on Interaction Design and Children*, IDC '09, page 89–96, New York, NY, USA, 2009. Association for Computing Machinery.

[4] Nicholas Vanderschantz, Annika Hinze, and Sally Jo Cunningham. "sometimes the internet reads the question wrong": Children's search strategies & difficulties. *Proceedings of the American Society for Information Science and Technology*, 51(1):1–10, 2014.

[5] Tatiana Gossen and Andreas NüRnberger. Specifics of information retrieval for young users: A survey. *Inf. Process. Manage.*, 49(4):739–756, jul 2013.

[6] Ion Madrazo Azpiazu, Nevena Dragovic, Oghenemaro Anuyah, and Maria Soledad Pera. Looking for the movie seven or sven from the movie frozen? a multi-perspective strategy for recommending queries for children. In *Proceedings of the 2018 Conference on Human Information Interaction amp; Retrieval*, CHIIR '18, page 92–101, New York, NY, USA, 2018. Association for Computing Machinery.

[7] Tatiana Gossen, Juliane Höbel, and Andreas Nürnberger. A comparative study about children's and adults' perception of targeted web search engines.

In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 1821–1824, New York, NY, USA, 2014. Association for Computing Machinery.

[8] Monica Landoni, Davide Matteri, Emiliana Murgia, Theo Huibers, and Maria Soledad Pera. Sonny, cerca! evaluating the impact of using a vocal assistant to search at school. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings*, page 101–113, Berlin, Heidelberg, 2019. Springer-Verlag.

[9] Thomas Beelen, Ella Velner, Roeland Ordelman, Khiet P. Truong, Vanessa Evers, and Theo Huibers. Does your robot know? enhancing children's information retrieval through spoken conversation with responsible robots. *CoRR*, abs/2106.07931, 2021.

[10] Thomas Beelen, Ella Velner, Roeland Ordelman, Khiet P Truong, Vanessa Evers, and Theo Huibers. Designing conversational robots with children during the pandemic. *arXiv preprint arXiv:2205.11300*, 2022.

[11] Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. Generating clarifying questions for information retrieval. In *Proceedings of the web conference 2020*, pages 418–428, 2020.

[12] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 475–484, 2019.

[13] Janet C Read and Stuart MacFarlane. Using the fun toolkit and other survey methods to gather opinions in child computer interaction. In *Proceedings of the 2006 conference on Interaction design and children*, pages 81–88, 2006.

[14] Lynne Hall, Colette Hume, and Sarah Tazzyman. Five degrees of happiness: Effective smiley face likert scales for evaluating with children. In *Proceedings of the the 15th international conference on interaction design and children*, pages 311–321, 2016.

[15] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1(1):71–81, 2009.

[16] Robert F Woolson. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*, pages 1–3, 2007.