Effects of asking questions in conversational user interfaces

Gleb Satyukov

s1592823

Media Technology MSc program, Leiden University

August 2017

Thesis advisors: Bas Haring and Fenna Poletiek

Effects of asking questions in conversational user interfaces

Abstract

Increasing popularity of conversational user interfaces has led to the widespread use of chatbots. With a wide range of practical purposes - from virtual assistants to customer service - these bots are useful in helping users in a natural conversation. Designed to convincingly simulate human behavior in a conversation these bots often succeed in fooling us, thereby passing the "Turing test". Applying various tricks such as arbitrarily delayed replies, elaborate conversational intents and personalized responses they manage to keep up the illusion of sentience. But is that the best they can do?

In this research we focused on the effects of chatbots asking questions in task-oriented conversations. The experiment is disguised as a game where the players can chat with RAM-z - chef of the robot cuisine. Participants are invited to provide step-by-step instructions on how to make a grilled ham and cheese sandwich. Depending on assigned condition the robot chef would alter his responses to include a specific question, a generalized question or no question at all.

Introduction

Recent developments in conversational human-computer interaction show a trend influenced by the alleged "Eliza-effect" - the notion that humans tend to ascribe human-like behaviors to machines (Hofstadter, 1996). Developers take advantage of this effect by making chatbots take on human characteristics in their communication. Whether it's the casual responses or the embodiment of conversation characters - these bots do a good job in tricking us to believe that they understand us (Cassell et al., 1999) (Shawar & Atwell, 2007). Some of the advanced examples include chatbots making small talk and asking questions in task-oriented conversations (Kopp, Gesellensetter, Krämer, & Wachsmuth, 2005). In this research we focused on the effects of questions on user performance when learning how to perform a task. The research question is defined as follows:

'Does asking questions in a conversational user interface affect user performance?'

The remainder of this section is going to provide context for this research in terms of related work. Method section covers the setup of the experiment and is followed by an analysis of the collected data. Subsequent section presents an interpretation of the results and a discussion on the implications for conversational user interfaces.

Early research

Research on conversational user interfaces began as early as 1964 with the ELIZA project. Joseph Weizenbaum set out to demonstrate how superficial the communication between a human and a machine can be and instead turned out to prove the opposite - our predisposition to ascribe human traits to machines (Weizenbaum, 1966).

At the heart of the program is a script that matches user input with a list of terms and responds with a reply based on the matches found. Despite it's naive mechanics ELIZA succeeded in producing the illusion of intelligence, not because it was factually intelligent, but rather because of human willingness to attach meaning to words and surrender to the illusion.

This effect has since been known as the "Eliza effect", which is defined as a special case of anthropomorphism applied exclusively to discourse between humans and machines. And consequently ELIZA became known as one of the first programs capable of passing the Turing test (Turing, 1950). The ELIZA program has been successfully used in supporting doctors with treatment of patients suffering from psychological issues (Colby, Watt, & Gilbert, 1966) and (Weizenbaum, 1976).

Thus following the *four possible goals to pursue in artificial intelligence* diagram as outlined by Russel and Norvig this effect falls in the bottom left category: *systems that act like humans* (Russell, Norvig, & Intelligence, 1995). A.L.I.C.E. (Artificial Linguistic Internet Computer Entity) is the successor of the original ELIZA and supports over 40.000 categories of knowledge (as opposed to 200 of the original ELIZA) (Wallace, 2009). A.L.I.C.E. has won the Loebner Prize (also known as the "The First Turing Test") multiple times[1].

[1] http://www.loebner.net/Prizef/loebner-prize.html

Chatbots

Advancements in the field of Artificial Intelligence and more specifically Machine Learning, combined with a growing popularity of instant messaging, have led to a burst in use-cases for conversational user interfaces (CUI). We will establish a formal definition of the term "chatbot" before diving into related research. Chatbots are defined as: 'a computer program designed to simulate conversation with human users, especially over the Internet'.

In essence - a chatbot is an implementation of a computer program that interacts with users by means of a natural dialogue - taking on the appearance of someone a user could talk to the same way they would with a friend or a family member. Leveraging convenience of existing messaging platforms (e.g. Facebook Messenger, Kik, Slack, Telegram, WeChat, etc.) these bots are integrated into the conversational user interfaces that people access for personal use. Not having to switch between different applications reduces cognitive load for the users and adds the familiarity of the messaging environment. Asynchronous nature of text messaging allows for easy access and non-blocking discourse where the user can leave and return to the bot at a later time that is more convenient.

Use cases for chatbots often vary in scope of activity that they support. From virtual assistants that are able to control the lights in the living room and scheduling appointments directly in your calendar[2] to very narrowly defined chatbots that are in charge of booking airline tickets[3]. Other more specific use cases for chatbots include, but are not limited to, use of conversational user interfaces for expertise search in domain of cultural heritage (Pilato, Vassallo, Augello, Vasile, & Gaglio, 2005) and in e-learning environments to engage students as presented in (Benotti, Martinez, & Schapachnik, 2014) and (Lundqvist, Pursey, & Williams, 2013).

Users expect to be able to use colloquial expressions in a conversation with a chatbot and thus require a certain level of comprehension (McTear, Callejas, & Griol,

[2]https://assistant.google.com/

[3]https://www.30secondstofly.com/

2016). Natural language is heavily dependent on context and subtlety of speech allows for ambiguity that is hard to extract from plain text input (Cohen, Cohen, Giangola, & Balogh, 2004). Increase in computational power in the last decades has allowed machine learning techniques to catch up with complexity of natural language processing (Shawar & Atwell, 2005)(Abdul-Kader & Woods, 2015). Existing logs of conversations can be used as training data for machine learning enabled chatbots (Huang, Zhou, & Yang, 2007). And as we have seen with ELIZA, it is not necessary to process all of the input in order to pass for an "understanding" chatbot (Weizenbaum, 1966).

Asking questions

Asking generalized questions and using small-talk to appear more human-like has shown to add value to conversational agents (Weizenbaum, 1966) (Bickmore & Cassell, 2000). While most examples of asking questions in conversational user interfaces consist of chatbot-based question answering systems (Waltinger, Breuing, & Wachsmuth, 2011) (Higashinaka et al., 2014), in some instances chatbots ask questions to the user instead (Kopp et al., 2005).

Existing research on the effects of asking questions extends the A.L.I.C.E. programming system to implement interrogative debugging - asking *why* questions about a program's behavior to improve user comprehension and ultimately and lead to less errors (Ko & Myers, 2004). This approach of asking questions is based on the non-interactive methods used for debugging, such as the "Rubber duck debugging" (Hunt & Thomas, 2000) or the interrogative debugging (Robertson et al., 2004).

In this research we focused on the effects of questions on user performance by combining the theory of interrogative debugging and asking questions in conversational user interfaces.

Method

In this section we cover the setup of the experiment including selection of the appropriate task, iterative development of the remote user study, off-line survey, initial testing (pilot) and the methods used to recruit participants for the final game.

Task Selection

In order to facilitate context for the task-oriented conversational user interface an appropriate task was selected. As a requirement it had to be relatively well-known, simple and short in duration - giving every participant a fair chance to complete the task successfully. Based on the alleged "Peanut butter and jelly sandwich challenge"[4], a task was chosen related to cooking - making a grilled ham and cheese sandwich.

With respect to the conversational user interface, number of messages sent by the participants is used as a measure of efficiency in order to determine a successful trial. As participants start the experiment they are randomly assigned one of the three conditions, these are numbered as follows:

1. No questions condition (control)
2. Generalized questions condition
3. Specific questions condition

Based on the assigned condition participants are presented with different responses by the robot chef. Reactions would also vary based on the current state of the sandwich as well as what the robot chef is holding in his hands. For examples of these questions and reactions as used in different conditions please refer to the Appendix A.

Development

The challenge on how to process user input presented itself early on in the development process and three options were considered at the time. First option was to setup the experiment where participants had the illusion of talking to a bot but in reality were actually talking to a another participant. Another option was to limit user input to only combinations of certain actions and objects. Finally, we could consider

[4]This challenge requires participants (traditionally freshman year college students) to provide step-by-step instructions on how to make a peanut butter and jelly sandwich. These instructions are executed by another participant (while being blindfolded) as literally as possible - in very much the same way a computer would execute a program.

not processing the input while checking that the users have read the questions in a different way (e.g. clicking on a “continue” button).

Inspired by the simplicity and elegance of the algorithms used in the ELIZA project, yet at the same time considerable level of expressive power, we decided to use a predefined vocabulary mimicking the framework used to make ELIZA scripts.

Cooking task provided the necessary context for game in much the same way a psychotherapy session was used to set the context for the ELIZA project running the DOCTOR script. This allowed for a definition of a compact yet comprehensive vocabulary that could be used to parse user input. Excerpts from the vocabulary can be found in the Appendix A.

Following the procedures outlined in (Weizenbaum, 1966) processing of the instructions is done by matching a set of keywords for actions and objects with plain text user input. For the specific vocabulary used to describe each of the ingredients and available actions, please refer to the Appendix A. When user input was parsed successfully, i.e. a valid instruction was recognized, the robot chef would reply with a positive response and perform the action. When no valid action was recognized in processed user input the robot would reply with a negative response and ask a follow-up question. These questions would differ between the specific and generalized questions conditions. For examples of such questions please refer to the Appendix A.

Visual representation of the robot chef and exuberant animations are used to visualize the processing of user input as well as progression throughout the experiment. Obeying commands provided by the users RAM-z would move around the kitchen to gather ingredients, assemble, toast and ultimately serve the sandwich indicating the end of a trial. Upon completing a trial users are presented with their score as well as the opportunity to share the game on social media. The scores are presented numerically as defined by equation 1 (where “least number of messages is the smallest amount of messages required to complete the trial”):

$$score = \frac{\text{least number of messages required}}{\text{number of messages used}} \times 1000 \tag{1}$$

This resulted in a score of *MAX = 1000 pts.* and is visually represented on a scale of 1 to 5 stars. Visual scale indicates how well participants performed with respect to the maximum score. Multiplication factor of a *1000 points* is used to amplify player scores - encouraging engagement with the game and increasing the likelihood of players sharing the game with others (Malone, 1981) (Pandelaere, Briers, & Lembregts, 2011).

Preliminary testing

Preliminary testing was performed to make sure that the task was easy enough to complete without requiring any specific qualifications. This was done by means of a survey where participants were asked to write down step-by-step instructions on how to make a grilled ham and cheese sandwich. Analyzing results of the survey clearly illustrated the expectations that users have with respect to the level of abstraction that was used in the their instructions.

A second round of testing was conducted using the pilot version of the conversational user interface. This version featured initial designs of the kitchen and included RAM-z - chef of the robot kitchen. In this version user input was structured and processed line by line while responses by RAM-z were presented in a single chat-bubble. Testing revealed ambiguous expectations with respect to playback of previous instructions (e.g. whether or not the sandwich assembly should reset to the playback position). A screen shot of the first version can be found in the Appendix E1.

Other notable observations included requirements for an introduction to the task, a list of available ingredients and a clear goal condition. Each of these observations has been integrated into the final version of the game.

Final version

The final version of the chatbot was developed including all requirements as determined by the initial off-line testing and the pilot version. For a screen shot of this final version (and of a trial in condition 3) please see Appendix E2. Following improvements were made with respect to the interface.

An introduction was added in order to properly introduce participants to the experiment. This consisted out of a welcoming message and a short description of the game on the landing page followed by an introduction by RAM-z the chatbot where he invited participants to help him make a grilled ham and cheese sandwich. Player name is registered during the introduction and is used during the analysis to identify unique trials. Please refer to the Appendix A for the raw transcript of the introduction flow.

Tooltips with a short description are added as hints to allow users to inspect the available ingredients and familiarize with the names used to describe them. This was added with the goal to improve user experience and understanding of the accepted commands, raw text used in the tooltips can be found in the Appendix C.

Based on feedback received during the pilot, the final version was more tolerant in usability with respect to spelling mistakes of impatient participants. In other words all messages were placed in a queue processed one by one, regardless of whether the users waited for the robot chef to finish processing the previous command. In addition to the processing queue, Levenshtein distance of 1 was used for tolerance of spelling mistakes in user input (Levenshtein, 1966).

Last but not least, some time has been devoted on improving the designs and turning rough sketches into something easier on the eyes. Main goal here was to make the game look sufficiently appealing for participants to share the experiment with others.

Collected data

For the measurements used in this research we collected data from of each of the trials, the format used can be found in Appendix D. Imperative elements used for measurements is the number of messages sent by the user, as well as the player name and device id for disambiguation of trials (Eckersley, 2010).

Unique trials are filtered out to account for the carry-over effect. This is accomplished by using player name and unique device id. As a result only the first attempt of a single person is used (provided they use their name on every trial).

Results

In this section we present a summary of the collected data as well as an analysis of the results in terms of variance and effect size.

Participants

Out of 434 registered trials only the unique trials have been selected and in case where multiple trials have been registered with the same user only the first attempt was taken into consideration in order to account for the carry-over effect, thus resulting in 423 unique trials.

Unique trials are subdivided into a group of 103 trials where participants did not enter any messages, 134 trials where participants did not finish the trial and 186 trials where participants successfully finished the trial by making a grilled ham and cheese sandwich. Number of trials and the proportion of complete versus incomplete trials is presented in the following table.

Table 1: Number of total, incomplete and complete trials by condition

Condition	**Total**	**Incomplete**	**Complete**
C1: No questions	133	100	52 (52%)
C2: Generalized questions	134	100	64 (64%)
C3: Specific questions	156	120	70 (58.3%)

Descriptive analysis

Following analysis of the data is performed only on the completed trials, taking the number of messages sent by the players as a measure of their performance in the game. Number of messages is a consistent measure across all trials. Alternatively, the duration of the trials in each condition could be used.

Instead of using calculated scores as presented in the game, the number of messages sent by the user was used directly as a measure of the performance. For the raw data please refer to the Appendix G.

Frequency distribution for each of the conditions is shown in Figures F1 through

F3. These charts show that the data is positively skewed, much more in the first two conditions than the third.

The variance of each of the samples is presented in a box-and-whiskers diagram in Figure F4. Except for a couple outliers in the third condition the means of the samples are relatively close to each other, indicating little variance between the groups.

Analysis of variance and effect size

A between subjects ANOVA test was conducted to compare the effects of asking specific questions, generalized questions and no questions (control) on user performance conditions. Based on the one-way ANOVA ($F(2, 183) = 1.3469$, $p = 0.2626$) no statistically significant differences between the means were determined. Results of the test are presented in the following table.

Table 2: Results of one-way ANOVA significance test for k=3 independent groups

Source	**SS**	**df**	**MS**	**F**	**p**
Between	1.2126	2	0.6063	1.3469	0.2626
Within	82.3768	183	0.4501		
Total	83.5894	185			

It is important to note that raw data was normalized prior to the analysis using the square root function. This is done in order to align the data closer to a normal distribution and in the process mitigate the effect of outliers. Please refer to the histograms of normalized data in Figure F5 through F7 to compare the distributions with those of the raw data (number of user messages) as presented in the previous section.

Known deficiency of significance tests is their dependence on sample size potentially leading to confounded P values (Coe, 2002). Effect size between groups is calculated to verify that sample size is not influencing the results. For between groups analysis, calculation of effect size is defined as shown in equation 2.

$$\eta^2 = \frac{\text{between groups sum of squares}}{\text{total sum of squares}} = \frac{1.2126}{83.5894} = 0.01450662405 \tag{2}$$

Cross referencing this result with relative effect sizes table as presented in (Sullivan & Feinn, 2012) we find there is a very small effect size for $\eta^2 < 0.2$ between the groups.

Discussion

The results of the conducted experiment suggest that there is no significant difference in asking generalized or specific questions. That is to say that effects of asking generalized or specific questions were not found given the context of this research and specific method of execution. In this section we will provide some insights into why no correlation was found and what the possible implications are for the conversational user interfaces.

In addition to the number of messages in each trial, durations can be used as a measure of efficiency. Each user message is timestamped at the time of creation and these timestamps are part of the data collected during the experiment. Trial duration is defined as the difference in time between the creation date of the last message minus the first one. Preliminary variance analysis of the timestamps did not show any significant difference in the means, confirming our previous findings. Referring back to the research question as stated in the Introduction we conclude that there is no effect of asking questions on user performance.

Interpretation of results

Taking into account the fact that the experiment is set up as a remote user study would explain a substantial amount of users (103 trials as shown in Table 1) who started the experiment but did not enter any messages. This is most likely due to the fact that those users were accessing the study from their mobile phone and did not get the “complete” experience of participation (e.g. smaller screen, absence of peripheral keyboard, etc.).

The differences between the number of complete and incomplete trials as shown in Table 1 requires further observation in order to be fully explained. Most likely here is the increased level of interaction with the chatbot as the bot is asking the questions to the users, however this statement would require further research to verify.

There are three theories that aim to explain some of the extreme outliers as observed in the data. The first theory deals with the "questionable" participants that were more likely playing around probing what actions were and weren't possible. Even though these data points can be considered a measurement error they are still taken into account as they resemble the nature of an uncontrolled remote experiment.

Another theory that perhaps more generally explains why some participants performed worse than others deals with the assumptions made in the experiment. As explained in the Method section in full length, assumptions were made in the order in which the task was performed (i.e. assemble the sandwich first and then grill). Assumptions like these would conflict with users that are used to a different method of performing the same task (e.g. grill every ingredient one by one and assemble the sandwich in the end).

Last but not least, other errors in measurement can be found due to the inherently uncontrollable nature of the remote experiment. This would include, but are not limited to, unstable Internet connection during the experiment, failing of the peripheral keyboard, deviations in screen resolutions where text would become harder to read as it overflows the available screen width and or height, and others.

Discoveries

One of the main discoveries made during the experiment is the variety of different approaches possible when it comes to making a grilled ham and cheese sandwich. The experiment was set up in a way that was able to handle unexpected user input, however it is very well possible that unrecognized input would lead to frustration of the users.

Another important discovery is the level of abstraction in user input and its alleged dependence on the appearance of the chatbot. In the context of this research

participants tended to agree on a certain level of abstraction in their messages. For example instead of commanding the robot to “extend left arm, lower left claw, grab the cheese”, users would say “grab the cheese”. This is most likely due to the level of abstraction in responses as well as appearance of the robot in the game. Related to this, a large proportion of the users reported a learning curve when getting to know what kind of commands RAM-z understands.

Future work

Included among the improvements that can be made to the user study is a rerun of the original game for a longer period of time with the intent to gather a larger sample size. Larger sample would allow us to make statements about the results with more certainty and would potentially result in a more normally distributed data.

An extension to the current research setup would be to include a qualitative analysis of the conditions. Preferably this would be realized in an unobtrusive way by leaving participants with the option to complete a questionnaire after the game is completed. Questions would include subjects such as perceived intelligence of the robot, robot comprehension, clear responses and overall satisfaction of the experience. Participants should be able to answer these questions using a 5-point Likert scale (Likert, 1932).

References

Abdul-Kader, S. A. & Woods, J. (2015). Survey on chatbot design techniques in speech conversation systems. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, *6*(7).

Benotti, L., Martinez, M. C., & Schapachnik, F. (2014). Engaging high school students using chatbots. In *Proceedings of the 2014 conference on innovation & technology in computer science education* (pp. 63–68). ACM.

Bickmore, T. & Cassell, J. (2000). How about this weather?" social dialogue with embodied conversational agents. In *Proc. aaai fall symposium on socially intelligent agents*.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., & Yan, H. (1999). Embodiment in conversational interfaces: rea. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 520–527). ACM.

Coe, R. (2002). It's the effect size, stupid: what effect size is and why it is important.

Cohen, M. H., Cohen, M. H., Giangola, J. P., & Balogh, J. (2004). *Voice user interface design*. Addison-Wesley Professional.

Colby, K. M., Watt, J. B., & Gilbert, J. P. (1966). A computer method of psychotherapy: preliminary communication. *The Journal of Nervous and Mental Disease*, *142*(2), 148–152.

Eckersley, P. (2010). How unique is your web browser? In *Privacy enhancing technologies* (Vol. 6205, pp. 1–18). Springer.

Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., ... Matsuo, Y. (2014). Towards an open-domain conversational system fully based on natural language processing. In *Coling* (pp. 928–939).

Hofstadter, D. R. (1996). Fluid concepts and creative analogies.

Huang, J., Zhou, M., & Yang, D. (2007). Extracting chatbot knowledge from online discussion forums. In *Ijcai* (Vol. 7, pp. 423–428).

Hunt, A. & Thomas, D. (2000). *The pragmatic programmer: from journeyman to master*. Addison-Wesley Professional.

Ko, A. J. & Myers, B. A. (2004). Designing the whyline: a debugging interface for asking questions about program behavior. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 151–158). ACM.

Kopp, S., Gesellensetter, L., Krämer, N. C., & Wachsmuth, I. (2005). A conversational agent as museum guide–design and evaluation of a real-world application. In *International workshop on intelligent virtual agents* (pp. 329–343). Springer.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady* (Vol. 10, *8*, pp. 707–710).

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

Lundqvist, K. O., Pursey, G., & Williams, S. (2013). Design and implementation of conversational agents for harvesting feedback in elearning systems. In *European conference on technology enhanced learning* (pp. 617–618). Springer.

Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction. *Cognitive science*, *5*(4), 333–369.

McTear, M., Callejas, Z., & Griol, D. (2016). Creating a conversational interface using chatbot technology. In *The conversational interface* (pp. 125–159). Springer.

Pandelaere, M., Briers, B., & Lembregts, C. (2011). How to make a 29 increase look bigger: the unit effect in option comparisons. *Journal of Consumer Research*, *38*(2), 308–322.

Pilato, G., Vassallo, G., Augello, A., Vasile, M., & Gaglio, S. (2005). Expert chat-bots for cultural heritage. *Intelligenza Artificiale*, *2*(2), 25–31.

Robertson, T., Prabhakararao, S., Burnett, M., Cook, C., Ruthruff, J. R., Beckwith, L., & Phalgune, A. (2004). Impact of interruption style on end-user debugging. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 287–294). ACM.

Russell, S., Norvig, P., & Intelligence, A. (1995). A modern approach. *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*, *25*, 27.

Shawar, B. A. & Atwell, E. (2007). Chatbots: are they really useful? In *Ldv forum* (Vol. 22, *1*, pp. 29–49).

Shawar, B. A. & Atwell, E. S. (2005). Using corpora in machine-learning chatbot systems. *International journal of corpus linguistics*, *10*(4), 489–516.

Sullivan, G. M. & Feinn, R. (2012). Using effect size—or why the p value is not enough. *Journal of graduate medical education*, *4*(3), 279–282.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460.

Wallace, R. S. (2009). The anatomy of alice. *Parsing the Turing Test*, 181–210.

Waltinger, U., Breuing, A., & Wachsmuth, I. (2011). Interfacing virtual agents with collaborative knowledge: open domain question answering using wikipedia-based topic models. In *Ijcai* (pp. 1896–1902).

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.

Weizenbaum, J. (1976). Computer power and human reason: from judgment to calculation.

Appendix A

Excerpts of vocabulary

Actions

```
{
  "actions": {
    "get": ["get", "grab", ..., "retrieve"],
    "put": ["put", "add", ..., "set"],
    "grill": ["grill", "cook", ..., "toast"],
    "drop": ["drop", "dump", ..., "discard"],
    "remove": ["remove", "take off", ..., "withdraw"],
    "finish": ["serve", "eat", ..., "exit"],
  },
}
```

Objects

```
{
  "objects": {
    "pronouns": ["it", "this", "that", "them"],
    "cheese": ["cheese", "gouda"],
    "ham": ["ham", "bacon", "meat"],
    "bread": ["bread", "loaf", "bun"],
    "sandwich": ["sandwich", "toast", "melt"],
  },
}
```

Positive responses

```
{
  "responses": [
    "all right",
    "absolutely",
    ...,
    "your wish is my command"
  ],
}
```

Negative responses

```
{
  "err_responses": [
    "ehh, i don't get it",
    "sorry, did not catch that",
    ...,
    "i'm sorry.. i don't understand"
  ],
}
```

Finished responses

```
{
  "finished_responses": [
    "done",
    "ready",
    "all set",
    "you're on",
    "ready when you are"
  ],
}
```

Generalized questions

```
{
  "generalized_questions": [
    "what's the next step?",
    "anything else i can do?",
    ...,
    "what's next?"
  ],
}
```

Specific questions

```
{
  "specific_questions": [
    "what should i do with the <OBJECT>?",
    "what do you want me to do with the <OBJECT>?",
    "maybe i should go get the <OBJECT> first?",
    ...,
    "maybe i should go get that first?"
  ],
}
```

Appendix B

Introduction flow

RAM-z: hello!

RAM-z: what's your name?

USER: <NAME>

RAM-z: hello <NAME>!

RAM-z: can you please help me out?

RAM-z: my chef asked me to make a grilled ham and cheese sandwich

RAM-z: and i'm still new to this

RAM-z: but if you could give me some instructions..

RAM-z: i'm sure we can work it out!

RAM-z: i think i've got all the ingredients here somewhere

RAM-z: (move your mouse over the table to see what we've got)

RAM-z: uhm, where should i start?

Appendix C

Tooltip texts


```
"tooltips": {
  "sandwich": {
    "title": "Sandwich",
    "description": "Mmmmm, looks delicious!"
  },
  "cheese": {
    "title": "Cheese (Gouda)",
    "description": "This is an essential ingredient of any grilled cheese sandwich"
  },
  "cheese-slices": {
    "title": "Cheese slices",
    "description": "Looks like the cheese is already sliced, awesome!"
  },
  "ham": {
    "title": "Ham",
    "description": "Put ham on your sandwich to make it a ham and cheese sandwich"
  },
  "ham-slices": {
    "title": "Ham slices",
    "description": "Take as many slices of ham as you like"
  },
  "bread": {
    "title": "Bread",
    "description": "Slices of bread form the outer layers of a sandwich"
  },
  "grill": {
    "title": "Grill",
    "description": "Universal tool that can roast, toast and grill your sandwich golden brown"
  }
}
```

Appendix D

Data structure

```
{
  "trial": {
    "player": <NAME>,
    "device": <UNIQUE DEVICE ID>,
    "user_agent": <BROWSER INFORMATION>,
    "grilled": <BOOLEAN>,
    "condition": <NUMBER>,
    "messages": [
      {
        "message": <TEXT>,
        "created_at": <TIMESTAMP>,
      },
      ...,
    ],
    "created_at": <TIMESTAMP>,
    "updated_at": <TIMESTAMP>,
  }
}
```

Appendix E

Visual impressions of the remote user study

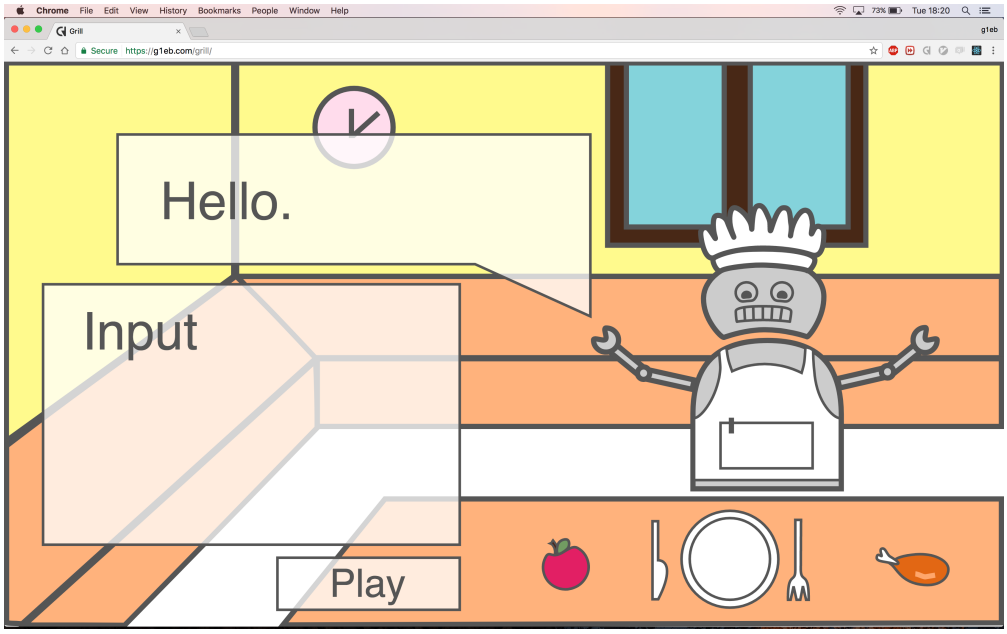

Figure E1. Screen shot of the initial version built for the remote study

Figure E2. Screen shot of the final version (condition with specific questions)

Appendix F

Data analysis

Frequency distributions (number of messages)

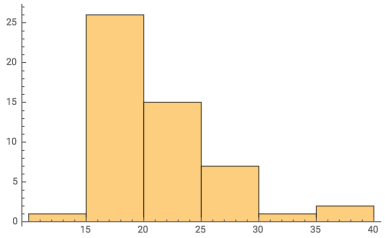

Figure F1. Number of messages in condition 1

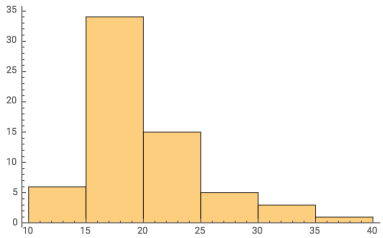

Figure F2. Number of messages in condition 2

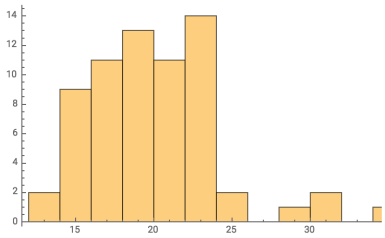

Figure F3. Number of messages in condition 3

Sample variance

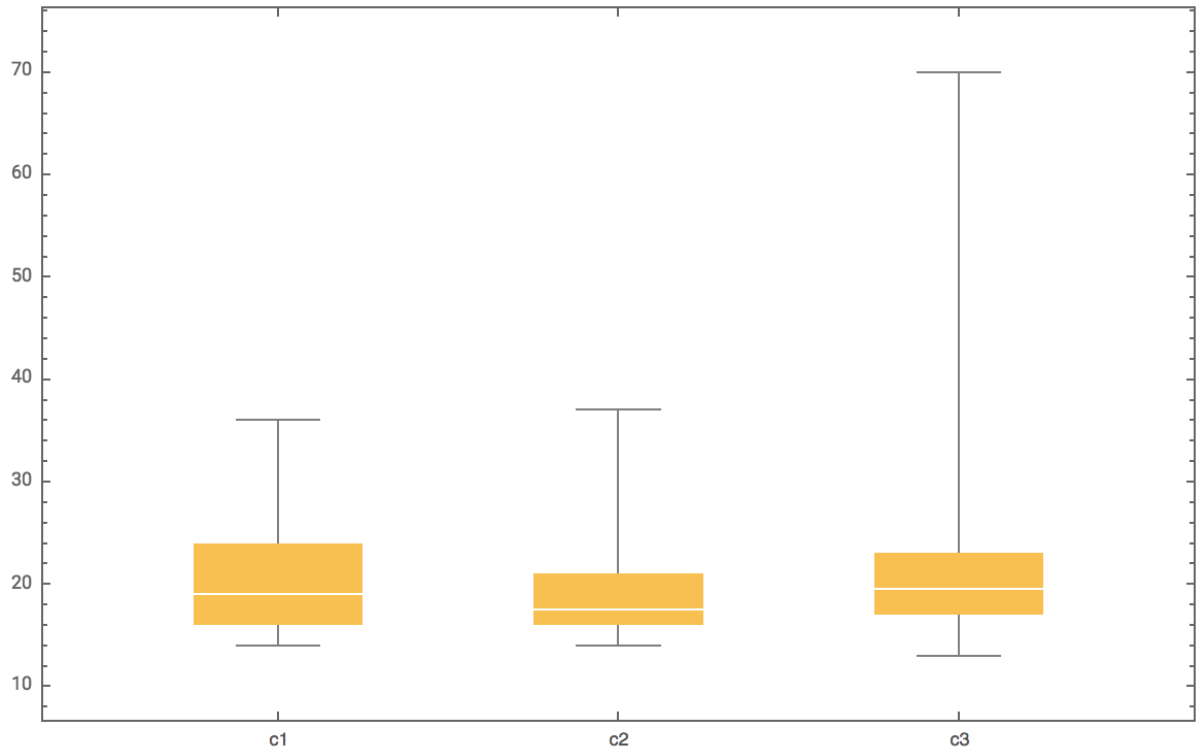

Figure F4. Variance of the samples in conditions 1, 2 and 3

Frequency distributions (normalized data)

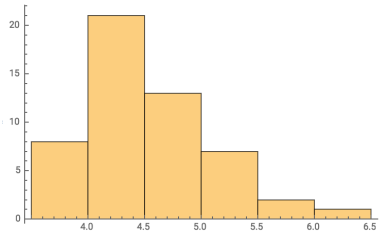

Figure F5. sqrt(number of messages) in condition 1

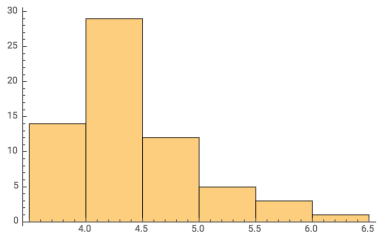

Figure F6. sqrt(number of messages) in condition 2

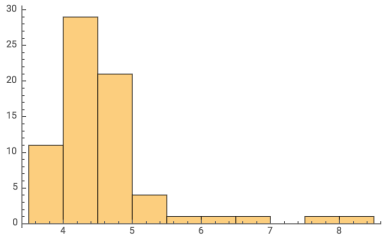

Figure F7. sqrt(number of messages) in condition 3

Appendix G

Raw data (number of user messages by condition)

condition 1	condition 2	condition 3
19	24	14
24	16	18
16	24	25
25	16	15
18	15	13
16	17	21
15	14	23
21	15	14
28	17	37
15	24	24
24	19	29
19	19	16
22	15	16
18	19	23
24	27	23
14	17	22
23	17	15
19	16	59
17	27	16
21	18	18
20	20	16
28	14	19
15	29	20
27	15	23
35	17	35
18	15	18
15	21	21
22	17	23
28	21	20
22	16	18
16	16	19
22	15	18
31	26	23
15	17	15
15	19	23
16	17	22
17	15	22
36	16	44
15	14	17
18	14	18
21	21	21
17	14	17
20	23	70
16	18	21
28	22	19
18	21	21
19	34	16
23	21	15
16	17	21
24	17	22
27	20	23
16	27	23
	20	30
	16	19
	14	19
	34	13
	21	23
	18	19
	16	19
	19	14
	22	20
	15	17
	32	30
	37	20
		20
		17
		17
		15
		15
		17