User Preference in Conditional and Unconditional Grounding in Turn-based Spoken Dialogue Systems

Denise de Waard
University of Twente
PO Box 217, 7500 AE Enschede
the Netherlands
d.dewaard@student.utwente.nl

ABSTRACT

Spoken dialogue systems are used more and more in everyday life. Unfortunately, there are often misunderstandings between the dialogue system and the user. To give the user the opportunity to repair these misunderstandings, grounding is implemented in spoken dialogue systems. Grounding is the process of reaching mutual belief that every person partaking in the conversation understands what has been said. This paper attempts to find which implementation of grounding users prefer. To do so, two types of grounding have been defined, conditional and unconditional grounding. Conditional grounding implements speech reparation by means of explicit verification through confidence levels. Unconditional grounding implements speech reparation by means of implicit verification, repetition of the answer of the user in the next question.

In this paper, users (n=20) have been asked to interact with both a conditional and an unconditional prototype, to determine which type of grounding in spoken dialogue systems is preferred by users.

KEYWORDS

Grounding, spoken dialogue systems, Turn-based dialogue systems, speech repair, misunderstandings

1. INTRODUCTION

In recent years almost everyone has heard of virtual assistants, such as Google Assistant and Alexa. These assistants give users the opportunity to give commands by speaking to the application. The application in turn, replies to the user with a digital voice. Google Assistant and Alexa are examples of spoken dialogue systems, systems that communicate with humans by means of spoken dialogue.

These systems are not always flawless. Oftentimes, Alexa or Google Assistant does not understand what the user says or makes wrong suggestions because of misunderstandings. To understand how misunderstandings in spoken dialogue systems can be handled, it is important to first understand how misunderstandings are dealt with in natural conversation.

When talking to each other, it often happens that people do not understand each other properly. Take for example this short conversation:

A. Do you own a j – cat?
B. A rat?
A. No a cat!
B. Yes, I own a cat named Mouse.

Example 1. Grounding in conversation

In example 1, person B does not understand the question asked by person A, a misunderstanding, the incorrect interpretation of the question [4], happens between person A and person B [3]. Person B asks a further question to give person A the opportunity to clarify the earlier utterance. This is an example of grounding. Grounding is the collection of mutual knowledge, mutual beliefs, and mutual assumptions that is essential for communication [6]. In this example, person A and B interact with each other to ensure they both understand correctly what person A meant by the first question.

Another example of grounding occurs when person B does not understand the question at all. In that case, there would be a non-understanding between the two parties [3]. A non-understanding can be solved by person B asking A to repeat the previous utterance. Misunderstandings and non-understandings are both errors that commonly occur in everyday conversation.

In short, grounding is the process of reaching mutual belief that every person partaking in the conversation understands what has been said. In human-to-human conversation, the process of grounding is very natural. Humans often go back and change or repeat something they just said. These occurrences are called speech repairs [8]. While it is easy for humans to detect which words are changed by means of this repair, this process is much more difficult for spoken dialogue systems. Most spoken dialogue systems are not able to handle self-repair within a single utterance, which leads to misunderstanding.

Another difference between human-to-human conversation and human-computer interaction is the way in which the conversations take place. In natural conversation, people can interrupt sentences or talk simultaneously. In turn-based spoken dialogue systems however, this is not possible. A turn-based system is a type of spoken dialogue system where the different parties speak in turns [18]. Generally seen, this means that a question-answer structure exists within the spoken dialogue.

For turn-based conversational agents several ways to implement grounding exist. The most common approaches to deal with misunderstandings are based on implicit and explicit verification [4]. I will further analyse these approaches to misunderstandings by defining conditional and unconditional grounding.

Conditional grounding is a type of grounding depending on how well the virtual agent understands the user. To determine the level of understanding confidence values, values that determine the level of confidence the system has that user utterances have been recognised [9, 16], are used. If the confidence value falls below a certain threshold, the system will take steps to gain more information. In conditional grounding, this will be done through explicit verification. Explicit verification can be defined as asking questions solely aimed at verifying the systems assumptions [11]. In this case, explicit verification is done by explicitly asking the user to repeat themselves: “Sorry I did not understand you; can you repeat your answer?”

36thTwente Student Conference on IT, Febr. 4th, 2022, Enschede, The Netherlands. Copyright 2022, University of Twente, Faculty of Electrical Engineering, Mathematics and Computer Science.

Unconditional grounding on the other hand, does not depend on how well the virtual agent understands the user. Instead, unconditional grounding is based on implicit verification. Implicit verification has two goals. Firstly, it repeats the user utterance in an attempt to verify whether it was correctly understood. Secondly, it proceeds with the next question in the conversation [11]. By means of the repetition of the user utterance, the user can recognise if the virtual agent has understood them correctly.

These two approaches to grounding have been implemented in many different spoken dialogue systems [2, 4, 9, 11, 15]. Very little research has however been done on which of these grounding methods is preferred by users. Because of that, it is difficult for spoken dialogue systems that have not yet implemented a form of grounding to decide on how to implement it.

An example of a project without implemented grounding is the BLISS project. BLISS is a Dutch project at the University of Twente attempting to measure the level of happiness of elderly people by means of a turn-based spoken dialogue system. Currently, there is no grounding system in place. Consequently, the question arises which type of grounding can best be implemented in the bliss project.

To aid the decision-making process of spoken dialogue systems such as BLISS, the following research question has been defined:

> *"Do users prefer conditional or unconditional grounding in turn-based spoken dialogue systems, and why is that the case?"*

In this paper, the research question will be answered through a user study based on prototypes. In the next section, related work to the research will be given. In section 3, the prototypes will be taken a closer look at, and the user study will be defined. In section 4, the results of the user study will be given. In section 5, the results will be further analysed and critiqued against the literature described in section 2. Besides that, some limitations of the research will be considered. Lastly, the conclusions and future directions will be discussed in section 6.

2. RELATED WORK

In this section the theoretical framework and similar studies will be discussed.

2.1 Theoretical Framework

In the correction of speech there is a distinction to be found between self- and other-correction of speech [14]. Self-repair in conversation is defined as the correction of their own speech by the speaker. Other-repair is defined as the correction of the speaker's speech by someone else. After taking an even closer look on self- and other-repair, the difference between self- and other-initiated repairs was found, where self-initiated repairs are repairs started without influence of the other party, and other-initiated repairs are defined as self-repairs which start through a signal of the other party. Signals initiating repair for example include short terms such as "what?" or "huh?", repetition of words, or asking for clarification.

According to research, people prefer self-repair in conversation [14]. Besides that, self-initiated repairs are more common than other-initiated repairs, because the opportunity to initiate repair for the speaker comes first and is easier.

2.1.1 Theory of Least Collaborative Effort

The preferred method of self-repair can be related to the theory of least collaborative effort [6]. According to the theory of least collaborative effort "participants try to minimise the total work that both do from the initiation of each contribution to its mutual acceptance." In short, this means that participants want to do the least amount of effort possible to reach a mutual understanding of the conversation.

Another take on the least collaborative effort is provided by Davies: "Participants in a conversation try to minimise the total effort spent in that interactional encounter." [7]. According to this definition, the principle of least collaborative effort is not only based on the joint goal of participants to reduce the effort in conversation, but also stems from individual motivation to limit the amount of effort they invest. His research suggests that while conversation is a collaborative activity, decisions about the effort put into a conversation are made on an individual basis.

2.1.2 Existing Repair Strategies

The least collaborative effort theory can also be found in relation to research in the field of speech repair strategies in conversational agents. Different papers focus on reducing the effort of reparation in speech by trying to detect errors more locally [10, 16, 17]. The goal of these papers is to ask more specific repair questions, as described in this example 2.

Speaker:	I have XXX plans.
General repair:	Can you repeat that?
Local repair:	What kind of plans?

Example 2. Local error detection [17].

The local repair in Example 2 is different to implicit verification, as the local repair is used to ask for clarification on a misunderstanding, while implicit verification is used to determine if the utterance has been correctly understood.

Besides the minimization of effort by creating local error detection, research is also done in using more context to detect misunderstandings. In natural conversations, communication between all parties of conversation happens both turn by turn and simultaneously. Examples of simultaneous communication include facial expressions and back-channel responses, such as "uh huh" and "yeah" [6]. Several studies research the inclusion of these conversational cues to reduce effort on error detection and reparation in other aspects of conversation, and thus reducing the effort of conversation [12, 19].

2.2 Similar User Studies

As mentioned in the introduction, very little research has been done on the preference of users on grounding in spoken dialogue systems. Ashktorab et al, however, have conducted a similar experiment [2]. In this research, user preferences on repair strategies relating to chatbots, text-based virtual agents, are researched by means of a scenario-based user study. Each participant is presented two different scenarios with the same context (shopping/banking/travel) and outcome (successful/unsuccessful). After the participant has taken a closer look at both scenarios, they are asked to select which scenario they preferred, with an explanation why.

In this study, eight different repair strategies were analysed, out of which "confirmation" can be compared to conditional grounding. In the strategy confirmation, confidence levels are used. If the confidence level falls below the threshold, the chatbot will repeat the phrase and asks the user to confirm whether the phrase has been understood correctly. The strategy "keyword confirmation explanation" makes use of repetition of user utterances to confirm the understanding of the chatbot. This strategy can be compared to unconditional grounding.

The results of the study indicate that users prefer a combination of three levels of contribution from the agent: acknowledging potential breakdowns, providing resources to assist user repair, and proactively suggesting solutions. In terms of the “confirmation” and “keyword confirmation explanation” this results in” keyword confirmation explanation” to be preferred over “confirmation” strategies.

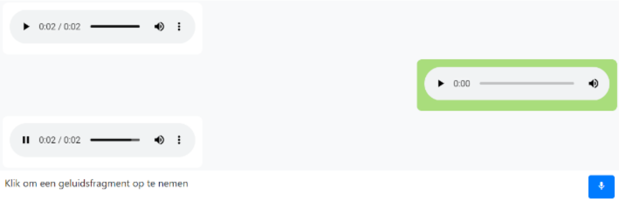
Other user studies in the field of virtual agents do not focus on speech repairs and grounding, but on changes in linguistic behaviour and acceptance of agents based on performance of the agent instead [1, 13].

3. METHODOLOGY

In this section, the methodology to answer the research question will be discussed. To answer the research question, descriptive research has been done by conducting user studies to determine the best type of grounding for the BLISS project. These user studies consist of the participants having a conversation with a prototype implementing conditional grounding and a prototype implementing unconditional grounding. These prototypes wi;l be called the conditional- and unconditional prototype for clarity. After each conversation, further questions based on the interactions were asked. The complete user study is in Dutch, since the BLISS project is in Dutch as well. Further information on both the prototypes and the user studies will be given in the next sections.

3.1 Prototypes

Both the conditional and the unconditional prototypes have been created in the cocos.whappbot environment provided by BLISS. This is a browser-based virtual agent, which was used to program a turn-based conversational agent. In Figure 1, an example of the webpage for the participant is shown.



Klik om een geluidsfragment op te nemen

Figure 1. Participant View of Prototype

On the left-hand side of the image, in grey, the recorded voice messages from by the virtual agent can be seen. On the right-hand side of the images the recorded messages of the participant are to be found. By showing these recordings, the user can replay the questions and replies when needed.

The blue microphone button in the right corner is used by the participants to record a reply to the agent. This is also explained by the text next to the button, “Click to record your voice”.

In the next sections, the design process of the prototypes will be discussed.

3.1.1 Wizard of Oz Prototyping

The first and most important decision in the design process of the prototype was the choice for a wizard of oz implementation. A wizard of oz implementation entails that part of the prototype is substituted by a human without the knowledge of the participant. The decision for a wizard of oz prototype was based on several aspects of the research.

Firstly, the length of the project was considered. Since the research only lasts a total of 10 weeks, there was not enough time to completely implement the implicit verification and the confidence levels in the prototypes. A wizard takes less time to implement and is therefore considered a better fit for the prototypes.

Secondly, using a wizard of oz prototype reduces the chance of the prototype malfunctioning, since a wizard of oz prototype is less complicated in terms of programming. Therefore, using a wizard prototype reduces the chance of the prototype tainting the results of the user study.

Thirdly, having a wizard implement the implicit verification gives control over the repeated answers. By doing so, it can be ensured that the same amount and type of understanding errors are found across all user studies.

3.1.2 Prototype versions

Multiple prototype versions have been created. For both the conditional and unconditional prototype two different dialogues were designed. By doing so, the influence of the dialogue on the participant is reduced.

On top of that, the order in which the participants interact with the prototypes is differentiated. 50% of the participants interact with an unconditional prototype first, the other 50% interacts with a conditional prototype first. This has been done to reduce order bias in the research.

In total, this means there are 4 different versions of the user study:

- Unconditional 1 and Conditional 1
- Conditional 1 and Unconditional 1
- Unconditional 2 and Conditional 2
- Conditional 2 and Unconditional 2

3.1.3 Dialogue

An important aspect of the prototypes is the dialogue. In this section, some important aspects of the dialogue will be discussed. The dialogue of the user studies is in Dutch, since the user studies are conducted in Dutch as explained before.

In compliance with the BLISS project, all dialogue in the prototypes has the theme happiness and wellbeing. This theme is translated into the following subjects:

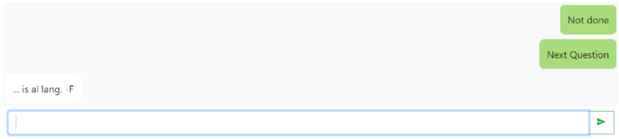
- Hobbies
- Holidays
- Sports
- Pets

Each conversation with a prototype should take about 10 minutes or less. During this conversation, a participant would be confronted with 10 cases of grounding, out of which 5 times the agent misunderstands the participant. The implementation of the dialogue for both the unconditional and conditional dialogue will be further discussed in the next sections.

3.1.4 Unconditional Prototype

Within the unconditional prototype, the process of implicit verification is implemented. This means that the answer to a question is repeated in the follow up question. The repetition of the answer in the follow up question is implemented by means of wizard of oz prototyping as discussed before. After each answer given by the participant, the wizard is tasked with inserting the answer in the follow up question. For that purpose, the wizard interface (figure 2) has been implemented.

The unconditional prototype has been based on implicit verification, the process of repeating answers in the next question. As discussed before, implicit verification was implemented by means of wizard of oz prototyping. For this implementation, a wizard interface was created. This interface is shown in Figure 2.



... is al lang. -F

Figure 2. Wizard interface

After a question has been asked by the agent and the participant answered the question, the follow up question will be posted on this interface, as shown in white on the left-hand side of Figure 2. The follow up question has several important characteristics:

- "…" The ellipsis marks the spot in the sentence where the implicit verification is implemented. The wizard does this by typing the answer of the participant into the line at the bottom of Figure 2.
- -F or -G at the end of the follow up question shows the wizard whether the participant should be correctly or incorrectly repeated. Each prototype has 5 cases of correct repetition and 5 cases of incorrect repetition of the participant. The -F and -G cases have been pre-defined.

False repetition of the participant has been done by means of misunderstanding, understanding different words than the participant has said. It is of importance that the phrases repeated are real words, to give the idea that the agent is intelligent but did not correctly hear the answer. After the wizard has inserted the answer of the participant into the sentence, the complete follow up question will be asked to the participant.

In the introduction of the user study, participants have been told to make sure that the wizard correctly understands what they are trying to say. They have been given the "go back" command, which they can use to return to the previous question to repeat their answer. This command has been implemented by means of wizarding as well. When the participant says "go back" the wizard will press the blue button "go back" as shown in Figure 3. If the participant does not ask to go back to a previous question, the button "next question" is pressed instead.

Figure 3. Go back button

3.1.4.1 Unconditional Dialogue

Since it is not known beforehand what the participant will answer to question 1, it is oftentimes difficult to design the follow up question. By asking specific questions with only a few possibilities for the answers, this process is simplified:

> Question 1: Where would you like to go to on holiday in the future?
> Question 2: I have never heard of that before, where is … located?

Example 3. Follow up question

Example 3 shows how asking very specific questions will result in the ability to ask a follow up question without knowing the answer to the first question. Besides questions with specific answers, closed questions, with only two choices as possible answers, are also included in the dialogue:

> Question: Do you go for a walk sometimes?

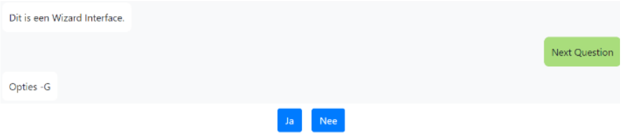
Example 4. Closed question with yes or no option

> Question: Do you read the news online or do you watch the news on television?

Example 5. Closed question with two options

Examples 4 and 5 show two different types of closed questions that have been implemented in the dialogue. Example 4 is a question which can be answered by either yes or no. This is, however, not specifically stated in the question. Participants are informed of closed questions at the start of the study and are asked to keep their answers as short as possible. Example 5, on the other hand, does specifically state the answer options for the participants. By offering a choice in this way, the dialogue can ask more specific questions about the option chosen.

The choice presented in these questions is implemented in the wizard interface by means of two buttons. Figure 4 shows the options "Yes" and "No". The participant does not press these buttons, they reply to the question by speech.



Dit is een Wizard Interface.

Opties -G

Figure 4. Wizard options

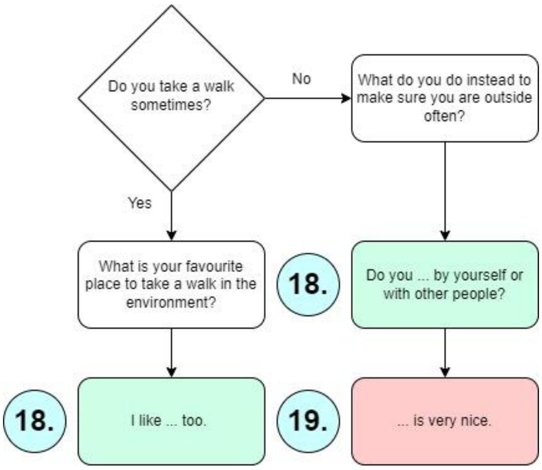

Do you take a walk sometimes?
No
What do you do instead to make sure you are outside often?
Yes
What is your favourite place to take a walk in the environment?
18.
Do you ... by yourself or with other people?
18.
I like ... too.
19.
... is very nice.

Figure 5. Unconditional dialogue scheme

Figure 5 shows a decision tree. In the figure, a closed question, the diamond shaped question, is shown. Both possible answers, yes and no, have different follow up questions in the square boxes. These square boxes have three different colours: white, green, and red. When the box is white, no grounding is present in the question or statement. When the box is green, grounding by means of correctly repeating the answer of the participant is present. In case the box is red, grounding is present by means of incorrectly repeating the participant. The circled, blue numbers next to the red and green boxes keep count of the number of grounding cases in the dialogue. As Figure 5 shows, there are two circled numbers "18", because the participant is located to one of the tracks after the closed question.

3.1.5. Conditional Prototype

The role of the wizard is much smaller in the conditional prototype than in the unconditional prototype. Since the prototype asks the participant to repeat their answer due to a misunderstanding, "Sorry I did not understand you, can you repeat your answer?", it is not needed for the wizard to insert the repetition of answers of the participant. Instead, the misunderstandings have already been programmed in the prototype. The only role for the wizard in this prototype is to select an option when the participant is asked a closed question.

Since the prototype asks the participant to repeat themselves, the "go back" command is not used in the conditional prototype.

3.1.5.1 Conditional Dialogue

For the implementation of conditional grounding, follow up questions are not necessary since the answer of the user is not repeated. However, it has been decided to create a similar dialogue structure to the unconditional grounding dialogue. By doing so, the influence of different types of dialogue during the user study is minimized. Instead of repeating the answer of the participant, the answer will therefore be referenced to:

> Question 1: Where would you like to go to on holiday in the future?
> Question 2: I have never heard of that place before, where is it located?

Example 6. Follow up questions in conditional dialogue

Example 6 shows that instead of user input, the word "it" is used to reference the answer given by the participant.

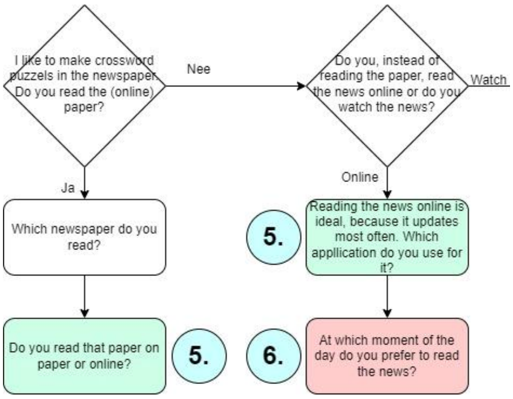

I like to make crossword puzzels in the newspaper. Do you read the (online) paper?
Nee
Do you, instead of reading the paper, read the news online or do you watch the news?
Watch
Ja
Online
Which newspaper do you read?
5.
Reading the news online is ideal, because it updates most often. Which appllication do you use for it?
Do you read that paper on paper or online?
5.
6.
At which moment of the day do you prefer to read the news?

Figure 6. Conditional dialogue scheme (in Dutch)

Figure 6 shows a scheme similar to the unconditional dialogue. From this diagram it can be derived that closed questions are still present in conditional dialogue. While it should be noted that the introduction of closed questions does create a form of implicit verification, it has been chosen to use the closed questions to ensure equivalence between the types of dialogue.

3.2 Experiment

In this section, the way in which the user studies were conducted will be explained.

3.2.1 Participants

The research has been conducted on a total of 20 participants. These participants have been gathered by asking friends and family of the researcher. To represent the population, diversity has been created by means of age. Figure 7 shows the age distribution of the participants.

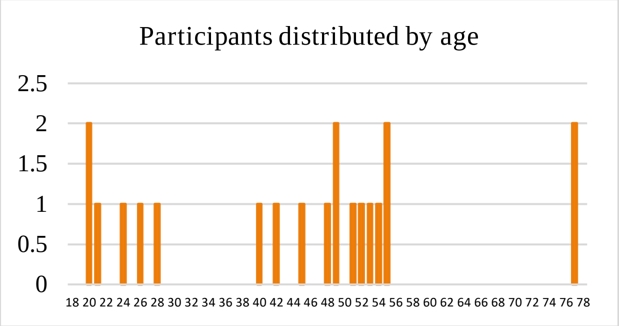


Figure 7. Participants distributed by age

From Figure 7 it can be concluded that there is a large concentration of participants belonging to the age group 40 till 55. Besides that, there are several participants between 20 and 28, and two participants with the age of 77. This graph shows that the diversity could have been better, but diversity in age still exists.

As mentioned before, there are a total of 4 different versions of the user study, based on 2 unconditional prototypes and 2 conditional prototypes. The participants are divided evenly over these different versions, meaning that each version of the user study is experienced by 5 participants. Each (un)conditional prototype is used by 10 participants.

3.2.2 Before the experiment

Before the start of the experiment, the participants are asked to fill out a consent form. In this form, they give permission to have their data recorded. The form gives the participants two options:

A. Share your data only for this project. In this case, participants give permission to create recordings of the conversation with the agent. Besides that, their age is recorded for diversity within the participant group.
B. Share your data with BLISS. In this case, participants give BLISS permission for further use of their conversation with the agent. Besides that, their age, gender, and region in which they grew up are asked.

3.2.3 During the experiment

The experiment consists out of several phases:

1. Introduction
2. Interaction with prototypes
3. Debriefing

3.2.3.1 Introduction

At the start of the experiment, the participants were given a short introduction. In the introduction, the participants were informed in regard of two major themes.

Firstly, the participants were told how the experiment would be performed. They are introduced to the prototype as an agent who talks to you, and who you will reply to. After that, they are told about the planning of the experiment, namely talking to the first prototype, answering some questions for evaluation on the interaction, talking to the second prototype, answering some questions for evaluation on the second interaction, and finally answering some general questions about both prototypes. It is important to note that in this introduction to the experiment, they were not notified about the true purpose of the study, only that a difference between the prototypes would be researched.

Secondly, some further information regarding the prototype was shared with the participants. They were told about the following characteristics:

- The speed of the prototypes. The prototypes were very slow, and the participants were informed to try and not pay attention to that.
- The understanding of the prototypes. The participants were told that the agent often misunderstands them and is still improving. Therefore, they were asked to make sure that the agent understands them correctly, to improve the agent in the future. They were given the "go back" command to do so.
- Be concise. The participants were told to try and give short answers, to help the agent in understanding them.
- Always try to answer the questions. The participants were told to try and always give an answer to the question, even if you for example do not have a favourite book. They are told that they did not need to

speak the truth, as the agent would not notice the difference either way.

Some of these characteristics are created by design, but the user was not made aware of that. The prototype not always understanding the participants for example is by design to make sure the participants run into the different ways grounding is implemented. Similarly, the prototype is slow because the wizard is writing down the answers given by the participants.

3.2.3.2 Interaction with prototypes

During the interaction with the prototypes, the participants were asked questions regarding what makes them happy, as discussed before. After each interaction with the prototype, the following questions were asked regarding the interaction:

1. How do you rate this interaction on a scale from 1 to 10?
2. What would you improve regarding this prototype?
3. What did you like about this interaction?
4. What did you think about the questions asked in this interaction?
5. Sometimes the prototype was unable to understand your answer. What did you think about the way this was handled?

After these questions are answered by the participant about each prototype, some further questions about both prototypes were asked:

6. Did you notice a difference between the two prototypes?
7. <<explain the difference>> Which type of handling errors did you like better and why?

3.2.3.3 Debriefing

Once the experiment was finished, the participants were debriefed. During the debriefing, they were first made aware of the goal of the study. After that, they were told about the wizard of oz prototype, and shown how the conversation they just had was not with a real functioning prototype, but with the experimenter instead.

Lastly, the participants were once again provided with the contact details of the researcher. These details were provided to give the participants the opportunity to withdraw their results from the results.

4. RESULTS

In this section, the results of the experiment will be discussed. These are split in two parts, quantitative and qualitative results.

4.1 Quantitative results

4.1.1 User preference Conditional vs Unconditional

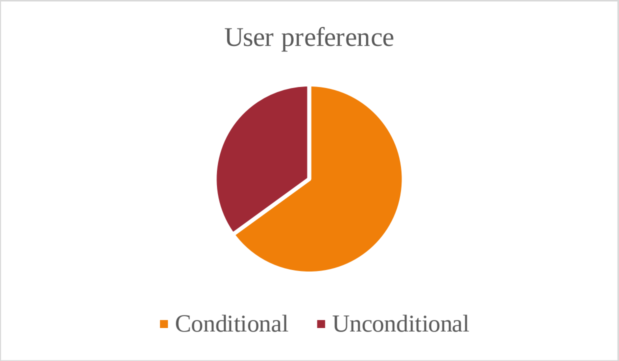


Figure 8. User preference

Figure 8 shows the result of the user study. Out of 20 participants, 13 people considered the conditional grounding prototype better, while 7 people preferred the unconditional grounding prototype.

4.1.2 Order bias analysed

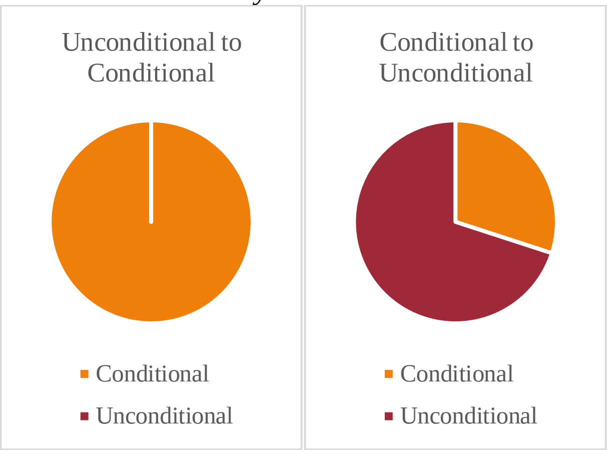


Figure 9 and 10. Order Bias

Figures 9 and 10 show very interesting results regarding order bias. All participants who started with the unconditional prototype and ended with the conditional prototype reported the conditional prototype as preferable. Within the group Conditional to Unconditional on the other hand, 70% reported the Unconditional prototype to be preferable. These results strongly suggest order bias to be present in the research.

4.1.3. Interaction score

On average, the unconditional prototype scores a 6.15. The conditional prototype has an average score of 6.59.

These scores reflect the earlier results, namely a higher mark for the conditional prototype. The difference is, however, not as large as the earlier data suggests. This is mostly because the grades have been given by the participants based on the entire interaction, not just based on the way errors are handled by the prototype. According to the participants, their grades were often based on the following themes:

- Speed, the prototypes were often very slow
- Dialogue
  - The subjects of the questions
  - The replies of the prototype

4.2 Qualitative Results

When looking at the perceived pros and cons of the prototypes, there are several themes that stand out:

1. Tempo of the conversation
2. Level of intelligence of the conversation
3. Level of empathy
4. Certainty of understanding

4.2.1 Tempo of the conversation

Several participants have mentioned the tempo of the conversation in their review. Since the question is repeated after issuing the command “go back”, conversation in the unconditional prototype feels slower. As one participant stated “Sometimes I even lost my train of thought in the conversation”

The conditional prototype, on the other hand, felt faster to several participants. Since their answers are not repeated the sentences feel shorter. Besides that, asking “Sorry I did not hear you correctly, can you repeat your answer?" is considered a faster way of repairing mistakes in conversation.

4.2.2 Level of intelligence of the conversation

Another theme that occurred more often in the user studies is the perceived level of intelligence of the prototype. According to multiple participants, the conditional prototype oftentimes feels more intelligent. This opinion mostly formed based on asking participants to repeat themselves, instead of wrongly repeating the answer like the unconditional prototype. "When the prototype asks me to repeat myself it feels more independent. You don't have to hold its hand like with the unconditional prototype" As this participant explained, asking people to repeat themselves gives the idea that the prototype is smarter instead of having to decide if it is correctly understood yourself.

Some people, however, perceived the unconditional prototype as more intelligent. This opinion is based on correct repetition of the answers given by the user. "The next question sounds more intelligent when it correctly repeats my answer." Repetition showed the participant that the prototype was intelligent enough to correctly repeat their answer.

4.2.3 Level of empathy

The unconditional prototype felt more empathic in conversation. Since the answers are repeated, the participants often felt better understood by the prototype. "The unconditional prototype feels more natural to talk with". Besides that, the use of the answer in the next question makes the prototype feel more personal. "The conversation was more personal; repetition makes the conversation sound cosier". The conditional prototype therefore feels much less empathetic.

Something everyone agrees on, however, is that the prototype uses too much positive confirmation. Both the conditional and the unconditional prototype confirm the answer of the user by means of positive reinforcement. A lot of participants feel that the prototype should get more of a personality and should not agree with everything that has been said. "The continues positive affirmation of my statements makes the agent feel like a people pleaser, very unpersonal and robotic."

4.2.4 Certainty of understanding

Certainty of understanding can be considered one of the most important aspects of the conversation with the agent. Because the answer is repeated in the unconditional prototype, you have the power to confirm yourself whether the prototype has understood you correctly. As several participants stated, "I like to judge for myself if my answer was correctly understood." Besides that, multiple participants have mentioned the advantage of hearing which part of the answer the prototype has misunderstood, to be able to repair the answer more efficiently. "By repetition of the answer I can hear which part has been misunderstood and how I can pronounce it better." There are, however, also some downsides mentioned to this final check. "If you are distracted, you can miss that the agent made a mistake and you will leave the wrong answer." Because of that, mistakes are not always caught. The conditional prototype, on this front, is more secure as it indicates whether it has understood the participant correctly for itself.

But not everyone agrees with this statement. Several participants have mentioned a disadvantage of the prototype deciding for itself if it has correctly understood the answer, namely that you cannot be certain if it has correctly understood you. "You cannot be certain that the agent has correctly understood it the second time, I feel like I just say the same thing twice." The computer might think it has saved the right answer, but you cannot be certain about that. Therefore, the preference of some participants goes to the unconditional prototype to be able to check if they have been correctly understood for themselves.

5. DISCUSSION

5.1 Results and Literature

In this section, existing literature will be compared to the results of the user study.

According to literature reparation in conversation is done through the theory of least collaborative effort [5, 6]. When looking at the definitions of conditional and unconditional grounding, conditional grounding fits the theory of least collaborative effort best. By means of conditional grounding the user only has to repeat themselves, while in unconditional grounding the user first has to judge for themselves if the repeated answer is correct, if not give the "go back" signal, and then has to repeat themselves. By the theory of least collaborative effort therefore, the user prefers conditional grounding. In practice, this has also proven true. 13 out of 20 participants prefer conditional grounding.

The results are however, not in accordance with the similar user study described in section 2.2 [2]. In that study, users prefer "keyword confirmation explanation (KCE)" (unconditional grounding) over "confirmation" (conditional grounding). These results, however, can not completely be compared side by side. While there are similarities to be found between the repair strategies, KCE and unconditional grounding, and "confirmation" and conditional grounding are not completely similar. Besides that, the study is based on text-based chatbots, while this research is based on spoken dialogue systems.

5.2 Limitations

During the performance of this research, several items were identified which could have influenced the results of the user studies. These items will be discussed in this section.

5.2.1 Errors Unconditional Prototypes

As explained before, the errors in understanding the user in the unconditional prototypes were implemented by a wizard. The mistakes were always implemented at the same points in conversation, to ensure similarity in the different user studies. The mistakes themselves, however, were not always similar. Because each participant gave a different answer, the misunderstanding of the prototype was also different in each case. Two different answers and mistakes to the question, "where would you like to go on vacation sometime?" are:

Answer: Austria
Repetition: Australia

Example 7. Closely related false repetition

Answer: Norway
Repetition: North Away

Example 8. Less closely related false repetition

Example 7 shows how the false repetition of the word Austria, is still similar by taking another country with a similar sound. In example 8 however, no country which sounds like Norway can be found, and therefore two different words are created which sound similar. It has been tried to maintain similarity in the different errors, but differences still occurred. For some participants the mistakes were therefore more obvious than for other participants.

Because of that, it cannot be said that the user experiences were the same in every study, meaning that there are confounding variables which could have influenced the results of the study.

Secondly, not every participant always repaired the mistakes made by the unconditional prototype. Several times, participants skipped over mistakes even while they noticed them. Because of this, not all participants repaired the same number of mistakes in the user study.

To try and prevent this difference, the researcher has also asked people to repair misunderstandings of the unconditional prototype during the study. Since this is a form of interference with the research, people have only been requested a maximum of one time to repair misunderstanding in the form of false repetitions of the agent. After that the researcher no longer interfered to ensure minimum influence. However, even asking a participant one time introduces outside stimulation to the study. Besides that, it did not always work, and thus the participants have not all repaired the same number of mistakes in the unconditional prototypes.

5.2.2 Dialogue and waiting time

As discussed before, it has been tried to reduce the influence of the dialogue by creating multiple versions of the dialogue. However, it is not certain that the dialogue did not have any influence in the results.

Besides that, the unconditional prototypes had a longer waiting time than the conditional prototypes between the questions of the agent. This is because of the wizard input necessary for the unconditional prototypes. It has been tried to increase the waiting time for the conditional prototype to combat this, but unfortunately that did not work. Therefore, there was a difference in waiting time between the different prototypes. Participants have been told during the introduction to try and not pay attention to this, but it did possibly influence the results of the user study.

5.2.3 Prototype breaking down

During the user studies, it has happened several times that the prototype broke down. Every time the prototype was repaired again shortly, but it could still have influenced the opinion of the participants. The breakdowns of the prototypes are evenly distributed between the prototypes. In total, the prototypes ran into problems about 5 times.

5.2.4 Online environment

Because of covid 19, the experiments were conducted online. As a result of that, the environment could not be completely controlled for the research.

One case where this was a problem was with two elderly people that were interviewed. Since these participants do not have a very large house, they were present during each other's trials and interviews. To combat this, several measures were taken. Firstly, the participants were given prototypes with different dialogues, to make sure that they did not have repetitions at the exact same moments in their conversations. Secondly, the structure of the interview was changed by having the participants both interact with their first prototype, before moving on to the second prototype. By doing so, they were unable to notice the differences between the prototypes before the start of their own interview. Because of that, these participants could have been influenced by each other.

Another downside of the online environment is the way in which the prototype was controlled. The prototype only works on one computer since the wizard and the user interface were not able to communicate with each other on different computers. Because of that, the participants were not able to click on the recording button themselves, but the researcher had to do so. This means that the participants do not have the full control over the prototype that they would have in a normal situation. Since, however, each participant had the same experience, it hopefully reduced the influence on the result.

5.2.5 Participant selection

As described in section 3.2.1 the participants were gathered by asking friends and family of the researcher. Because all participants have a personal relationship with the researcher, their opinions can be influenced by this relationship. The participants for example could not have wanted to hurt the feelings of the researcher by giving negative feedback on the prototypes.

5.2.6 Order Bias

The results show a very noticeable order bias in the opinion of participants on their preferred mode of grounding. Since the results of the experiment seem to have been severely influenced by the order bias, it should be taken into consideration when drawing conclusions on the research.

6. CONCLUSIONS AND FUTURE WORK

In conclusion, the conditional grounding prototype is considered a better way to repair mistakes by 65% percent of the participants. According to the participants, the conditional prototype feels more independent, smarter, and has a better conversational flow. The biggest downside to the conditional prototype, however, is that you cannot be certain that the prototype has correctly understood your answer. In this aspect, the unconditional prototype is better than the conditional prototype. On top of that, the unconditional prototype is perceived as having a more personal conversation due to the repetition of the answers of the user. On the other hand, multiple participants have stated that the repetition of wrong answers made the prototype less intelligent and slower. Besides that, participants considered having to say, "go back", which resulted in a repetition of the previous question, to be too long of a process, resulting in a loss of concentration in the conversation.

However, since there are only 20 participants who partook in the study and a strong order bias has been found, it is not possible to draw definitive conclusions from this research. Therefore, I propose a new experiment, with the following differences to the current research. Firstly, I propose to have participants interact with only one of the current prototypes. After the interaction, more specific questions regarding grounding and error handling should be asked. In these questions, some examples of changes to the prototype interacted with will be given. These examples are related to the type of grounding they did not interact with in the prototype. This means examples of unconditional grounding are given after interaction with conditional grounding and examples of conditional grounding after interaction with unconditional grounding in the prototype. Secondly, I propose to create even more similar experiences in the different prototypes, by making sure the wait time is more similar than it currently is, by increasing the waiting time of the conditional prototype. Lastly, the proposed experiment is advised to be conducted with a larger group of participants to gain more statistical relevancy. Besides that, it is advised to also include participants without a personal relation to the researcher.

As a final note, the BLISS project is advised to take this research project as a guideline for the implementation of grounding within the Whappbot.

7. REFERENCES

[1] Aneja, D., McDuff, D. and Czerwinski, M. Conversational Error Analysis in Human-Agent Interaction. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents* (New York, 2020).

[2] Ashktorab, Z., Jain, M., Vera Liao, Q. and Weisz, J. D. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Conference on Human Factors in Computing Systems - Proceedings* (MAryland, USA, 2019).

[3] Bohus, D. *Error Awareness and Recovery in Conversational Spoken Language Interfaces*. Ph.D. Dissertation. Carnegie Mellon University Pittsburgh, PA, 2007.

[4] Bohus, D. and Rudnicky, A. Sorry I didn't Catch That: An Investigation of Non-understanding Errors and Recovery Strategies. *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue* (01/01 2005).

[5] Bresnahan, M. Discourse Structure and Anaphora: Written and Conversational English. *Studies in Second Language Acquisition*, 12 (03/01 2008), 105.

[6] Clark, H. H. and Brennan, S. E. Grounding in Communication In: L.B. Resnick, J.M. Levine and S.D. Teasley, eds., *Perspectives on Socially Shared Cognition*. American Psychological Association, (2004) 127–149.

[7] Davies, B. Least Collaborative Effort or Least Individual Effort: Examining the Evidence. *Not Found* (01/01 2007).

[8] Heeman, P. A. and Allen, J. F. Speech Repairs, Intonational Phrases and Discourse Markers: Modeling Speakers' Utterances in Spoken Dialog. *Comput. Linguistics*, 25 (1999), 527-571.

[9] Komatani, K. and Kawahara, T. Flexible Mixed-Initiative Dialogue Management using Concept-Level Confidence Measures of Speech Recognizer Output. *Proc. COLING*, 1 (01/11 2003).

[10] Komatsu, S. and Sasayama, M. Speech error detection depending on linguistic units. *In PervasiveHealth: Pervasive Computing Technologies for Healthcare* (Trento, Italy, 2019).

[11] Krahmer, E. J., Swerts, M., Theune, M. and Weegels, M. F. Error Detection in Spoken Human-Machine Interaction. *International Journal of Speech Technology*, 4 (2001), 19-30.

[12] Marge, M. and Rudnicky, A. I. Miscommunication detection and recovery in situated human-robot dialogue. *ACM Transactions on Interactive Intelligent Systems,* 9, 1 (2019).

[13] Park, C., Lim, Y., Choi, J. and Sung, J. E. Changes in linguistic behaviors based on smart speaker task performance and pragmatic skills in multiple turn-taking interactions. *Intelligent Service Robotics*, 14, 3 (2021), 357-372.

[14] Schegloff, E., Jefferson, G. and Sacks, H. The Preference for Self-Correction in the Organization of Repair in Conversation. *Language,* 53 (06/01 1977), 361-382.

[15] Skantze, G. *Error Handling in Spoken Dialogue Systems - Managing Uncertainty, Grounding and Miscommunication*. Ph.D. Dissertation. KTH Computer Science and Communication, Department of Speech, Music and Hearing, 2007.

[16] Stoyanchev, S. and Johnston, M. Localized error detection for targeted clarification in a virtual assistant. *In ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (South Brisbane, Queensland, Australia, 2015).

[17] Stoyanchev, S., Salletmayr, P., Yang, J. and Hirschberg, J. Localized detection of speech recognition errors. In *Spoken Language Technology Workshop (SLT)* (Miami, USA, 2012).

[18] Tominaga, Y., Tanaka, H., Ishiguro, H. and Ogawa, K. Standard Dialogue Structure and Frequent Patterns in the Agent Dialogue System. *In Distributed, Ambient and Pervasive Interactions,* Norbert StreitzPanos Markopoulos, 2021, 348-360.

[19] Visser, T., Traum, D., DeVault, D. and op den Akker, R. A model for incremental grounding in spoken dialogue systems. *Journal on Multimodal User Interfaces*, 8, 1 (2014), 61-73.