UNIVERSITY OF GOTHENBURG

THE DEPARTMENT OF PHILOSOPHY, LINGUISTICS AND THEORY OF SCIENCE

# AN EXPERIMENTAL EVALUATION OF GROUNDING STRATEGIES FOR CONVERSATIONAL AGENTS

**Yiqian Zou**

| | |
|---|---|
| Thesis: | 15 hp |
| Program: | Master's Programme in Language Technology |
| Level: | Advanced level |
| Semester/year: | Spring, 2020 |
| Supervisor: | Staffan Larsson |
| Co-supervisor: | Vladislav Maraev, Christine Howes |

# Abstract

With the continuous development of technology, dialogue system's technology penetrates into human's life. Grounding also becomes more and more important for dialogue systems. It is important to choose a suitable grounding strategy in a conversational agent. Two grounding strategies are compared in this article, explicit feedback and implicit feedback. The explicit feedback in this article is different from interrogative explicit feedback. It has been modified to make a system says ”Ok, x” in response to utterance x.

The aim of this paper is to compare two grounding strategies and to find out which one is better. Additionally, how users respond to false feedback is also the research question in this article. In order to draw a conclusion, a dialogue system was implemented. This article uses a mix of quantitative method and qualitative method. Questionnaires are used to investigate the subjective judgments of participants. Participants evaluated the dialogue system through questionnaires. In the questionnaire, users rate the system from two aspects, naturalness and ease. From June 8th to 14th, the system was officially available. The data were analyzed by t-test and the result was presented in this article with diagrams.

Most participants mentioned that they prefer the system with explicit feedback in the evaluation. According to the average score, the system with explicit feedback in this paper is more natural and easier to communicate than the system with implicit feedback. However, there is no significant difference between these two grounding strategies according to the results of the T-test. This does not mean that there are no differences, but that such differences may not be obvious because of the little sample size. In addition, user's response to the wrong feedback is summarized in this article. Four kinds of reactions are described in this article, hesitation, repetition, point out the wrong feedback and correction.

**Key words:** Grounding strategy, Grounding, Dialogue system

# Acknowledgement

My deepest gratitude goes first and foremost to my supervisor, Staffan Larsson, for his constant encouragement and guidance. He led me through all stages of writing this thesis. He gave me instructive advice and useful suggestions for my thesis. Without his consistent and enlightening guidance, this paper would not have reached its current form.

High tribute shall be given to my co-supervisor, Vladislav Maraev, who continuously provide me with technical support. He put forward constructive comments on my system and helped me solve some programming problems.

I am also deeply indebted to another co-supervisor, Christine Howes. She gave me a lot of advice on research methods. She often takes time out of her busy schedule to guide my thesis.

Special thanks to my friends for always encouraging me. I also want to thank my parents. They cheered me up during the days when I couldn't hold on.

# Table of content

1. Introduction ........................................................1
1.1 Background........................................................1
1.2 Purpose and research question ........................................................1
1.3 Delimitation ........................................................3
2. Related Work ........................................................4
2.1 Previous research ........................................................4
2.2 Feedback strategies ........................................................5
2.2.1 Interrogative and positive explicit feedback ........................................................5
2.2.2 Implicit feedback........................................................6
2.2.3 The relationship between positive explicit feedback and implicit feedback ........................................................6
2.3 Response to wrong feedback ........................................................7
2.4 The state of the art of grounding strategies in conversational agents........................................................7
3. Methodology........................................................10
3.1 Experimental quantitative research ........................................................10
3.2 Human evaluation ........................................................10
3.2.1 Field experiment........................................................10
3.2.2 Telephony platform: Voxeo ........................................................11
3.2.3 Subjective Assessment of Speech System Interface (SASSI)........................................................11
4. Experimental Setup........................................................12
4.1 Dialogue design ........................................................12
4.1.1 Main dialogue........................................................12
4.1.2 Tutorial dialogue ........................................................12
4.2 Implementation ........................................................13
4.2.1 Randomize ........................................................13
4.2.2 Audio output and TTS ........................................................14
4.2.3 Audio input ........................................................14
4.2.4 Record........................................................15
4.3 Data collection ........................................................15
4.4 Data analysis........................................................16
4.5 Ethical considerations........................................................17
5. Results........................................................18
5.1 Questionnaires ........................................................18
5.1.1 Background of participants ........................................................18
5.1.2 The evaluation of dialogues with positive explicit feedback ........................................................19
5.1.3 The evaluation of dialogues with implicit feedback........................................................19
5.1.4 The result of T-test ........................................................21

5.2 The user’s response to wrong feedback....................................................................................21
6. Discussion ................................................................................................................................22
6.1 Comparison of two feedback strategies.....................................................................................22
6.2 User response when communicating with the system..............................................................23
7. Conclusion.................................................................................................................................24
7.1 Research questions.....................................................................................................................24
7.1 Limitation and future work........................................................................................................24
Reference list...................................................................................................................................26
Appendix 1. The content of audio files............................................................................................29
Appendix 2. Users’ response to wrong feedback .............................................................................30

# 1. Introduction

*In this chapter, the development prospects and market share of the dialogue system are described at the beginning. It also describes the main areas of the dialogue system and its importance to human life. Then the definition of grounding and research status of grounding are described, leading to the purpose and research questions of this article. Finally, some limitations of this research are further clarified.*

## 1.1 Background

A dialogue system is a program that communicates with human users in natural language (Bui, 2006). In recent years, with the increasing penetration of connected devices and the continuous development of technology, the market of dialogue systems has increased significantly. According to MarketsandMarkets (the second largest market research institution in the world), the global market of the dialogue system in 2019 is $ 4.2 billion. By 2024, this market will increase to $ 15.7 billion. The main conversational agents across the market include Microsoft Corporation (US), IBM Corporation (US), Google (US), Baidu (China), etc (Ken Research, 2019).

A dialogue system has a wide range of uses in various fields. International Data Corporation (IDC) previously predicted that by 2020, the penetration rate of the dialogue system in the three key areas of smart home, portable devices, and smart vehicles will be 27%, 68% and 51%. In addition, it will bring value enhancement to traditional industries such as finance, retail, and medical treatment in the next five years, and the increase will reach 64%, 62% and 48%.

With the continuous popularization of the mobile Internet, the spoken dialogue system has been closely integrated with mobile products such as Siri and Google Assistant. Users can get access to information and services through the dialogue system (Zhao et al., 2019). To achieve natural interaction with humans, the system needs to be able to be awakened, can recognize and understand human conversations, and give satisfactory and natural feedback (IDC, 2018). In many scenarios, a dialogue system can collect effective information from communication with users, such as travel reservations, information inquiry, taxi booking, route planning and etc (Arora et al., 2013).

Grounding is the foundation of communication (Clark et al., 1991). In communication, common ground will be updated through the grounding process (ibid.). Similarly, grounding is also a crucial part of the dialogue system. Although there is a lot of knowledge about grounding strategies in human-human dialogue, today's conversational agent platforms are not very powerful in supporting the grounding process.

Under such circumstances, users always have problems when communicating with conversational agents. Sometimes, they feel that the conversation with a conversational agent is unnatural. They will be annoyed when utterances are continuously recognized incorrectly by a dialogue system. Therefore, this paper hopes to compare two grounding strategies from the user's subjective evaluation.

## 1.2 Purpose and research question

The rapid development of big data and deep learning has greatly promoted the latest development of scientific research fields such as dialogue systems and computer vision (Chen et al., 2017). Therefore, most researchers use deep learning and big data to study spoken dialogue systems from multiple perspectives (Zhao et al., 2019).

Spoken dialogue systems involve most language-related sub-fields, such as ASR, dialogue management, affective modeling, etc. One of the most critical components of spoken dialogue systems is the speech recognizer (Callejas et al., 2014). Therefore, ASR errors are very fatal in dialogue systems. The main research areas to solve this problem are: detecting and eliminating background noise, recognizing in real-time even a certain prediction of the next input, integrating affect and emotion recognition as part of ASR (Batliner, Seppi, Steidl and Schuller, 2010) and some new technologies, such as deep neural networks (Dahl, Yu, Deng and Acero, 2012).

The process of grounding can also cope with this error well. Through grounding, it could be confirmed that an utterance has been heard, understood and accepted correctly. In human-human dialogue, grounding is used to confirm that they understand correctly. In most conversational agents, grounding is also used to confirm users' utterances. However, in the research field of spoken dialogue systems, little research focuses on grounding. In particular, there are few articles on grounding strategies from the perspective of user experience. Therefore, trying to use different grounding strategies in a dialogue system may change the user experience. Additionally, it is of great significance to try to transform a large amount of grounding strategy knowledge in human-to-human dialogue into human-machine dialogue.

Two grounding strategies are chosen to be compared in this paper. One is positive explicit feedback, that is the system says “OK, x” in response to “x”. Another one is implicit feedback, that is the system embeds “x” in its next utterance. For example, the following scenario of a user scheduling a meeting with the system.

U: I want to book a meeting on Wednesday.
S: **OK, Wednesday**. What time is the meeting? (positive explicit feedback)

U: I want to book a meeting on Wednesday.
S: What time do you want to reserve a meeting **on Wednesday**? (implicit feedback)

It can be seen from the above dialogue that when the user gives the feedback utterance "Wednesday" that needs to be received by the system, the system with explicit feedback clearly points out the feedback utterance "Wednesday" and continues to ask for the next required information, while the implicit feedback embeds the feedback utterance "Wednesday" in the next question.

Both types of feedback have their advantages and disadvantages. Explicit feedback does help users confirm the utterances more directly, but in some cases, it is not too natural. In contrast, it is indeed more natural for the system with implicit feedback to embed utterance into a question, but users are more likely to ignore confirming utterances. Therefore, a dialogue system was implemented to compare the two grounding strategies based on the characteristics.

The main goal of this paper is to compare two grounding strategies through an experimental evaluation. Through the evaluation by participants, these two grounding strategies can be compared subjectively. This paper is expected to provide a good basis for choosing which grounding strategies to implement in a conversational agent.

At the same time, it is also expected to collect users' responses to wrong feedback utterances. According to Skantze (2007), error handling should not be a single process in the system. Error handling detection includes explicit and implicit feedback, positive and negative evidence and ASR confidence score estimation. Most examples of negative evidence involve negative words like “not”. However, users may not only use negative words to point out an error. It is expected to summarize more negative evidence by collecting user's response to wrong feedback utterances, which can lead many dialog designers avoid using implicit feedback (ibid.)

Thus, the research question is stated as followed:

1. Which grounding strategy is better from users' perspectives?
2. How does user response to wrong feedback?

## 1.3 Delimitation

Once a hearer receives an utterance, a reaction phase is necessary to interpret the utterance and decide what to do next. Both Allwood (1995) and Clark (1996) mention the reaction phase and used different terms to divide it into consideration and feedback. In the feedback phase, the grounding of utterance will be processed. The two concepts of feedback and grounding have overlaps and differences. According to Larsson (2002), in numerous studies of practical dialogue systems, grounding is often reduced to verification or confirmation of utterances. In this paper, feedback refers to this concept.

There is a lot of knowledge about the grounding strategies in human-human dialogue, but so far, only a part of them have been implemented. In this paper, positive explicit feedback and implicit feedback are focussed on.

According to Jurafsky and Martin (2004), explicit feedback strategy is a question-style. A system with interrogative explicit feedback asks the user a direct question to confirm their understanding, such as the dialogue (a) shown below.

**(a). Interrogative explicit feedback:**

U: I want to book a meeting on Wednesday.
S: Do you want to book a meeting on Wednesday?
U: Yes.
S: What time is the meeting?

Interrogative explicit feedback plays an important role in dialogue system. However, this paper focuses on explicit feedback and implicit feedback that give correct feedback. Therefore, positive explicit feedback has been used in this paper. Implicit feedback strategy in this paper is the same as the definition in Jurafsky and Martin (2004), that is, the system embeds user’s value in its next utterance. The explicit feedback and implicit feedback in this paper are as followed:

**(b). Positive explicit feedback in this paper:**

U: I want to book a meeting on Wednesday.
S: OK, Wednesday. What time is the meeting?

**(c). Implicit feedback in this paper:**

U: I want to book a meeting on Wednesday.
S: What time do you want to book on Wednesday?

# 2. Related Work

*In this chapter, previous research is described. A set of theories and concepts that can be used in the later analysis and conclusion is described, such as grounding, explicit feedback and implicit feedback. These theories and concepts serve as a basis for our research. Finally, the art of grounding strategy in conversational agent is being discussed.*

## 2.1 Previous research

Clark and Schaefer (1989) defined grounding as the process of properly updating the common ground. In conversation, grounding ensures that all participants understand the discussion, especially when they are talking about the same topic (Burgan, 2017).

However, participants in a dialogue can never understand perfectly. Clark and Schaefer (1989) theorized the grounding criterion: The contributor and the partners mutually believe that the partners have understood what the contributor meant to a criterion sufficient for current purposes.

In the same article, they established Clark and Schaefer's Contribution model of grounding, which is widely discussed (Cormac, 2002). They divided the contribution into two phases, presentation phase and acceptance phase:

1. **Presentation phase**: **A** presents utterance **u** for **B** to consider. He does so on the assumption that, if **B** gives evidence **e'** or stronger, he can believe that **B** understands what **A** means by **u**.
2. **Acceptance phase**: **B** accepts utterance **u** by giving evidence **e'** that he believes he understands what **A** means by **u**. He does so on the assumption that, once **A** registers that evidence **e'**, he will also believe that **B** understands.

Clark redefined grounding in 1996, stating that grounding should occur at all levels of communication, such as attention, identification, recognition and consideration. In 1999, Traum pointed out deficiencies of the Contribution model and proposed to use the Strength of Evidence principle to solve this problem.

Both Allwood (1995) and Clark (1996) argued that all dialogues involve four different levels of actions. But their definitions are slightly different, so Larsson (2002) further summarized each action level:

1. **Reaction**: whether **utterance** has been integrated.
2. **Understanding**: whether **utterance** has been understood.
3. **Perception**: whether **utterance** has been perceived.
4. **Contact**: whether **hearer** and **speaker** have contact.

Clark and Schaefer (1989) also list five different types of evidence that can show a contribution has been understood. They listed from the weakest to the strongest:

1. **Continued attention**: B shows that he is continuing to attend and therefore remains satisfied with A's presentation.
2. **Relevant next contribution**: B starts in on the next contribution that would be relevant at a level as high as the current one.
3. **Acknowledgement**: B nods or say "Mm-hmm", yeah or something like that.
4. **Demonstration**: B demonstrates all or part of what he has understood A.
5. **Display**: B displays verbatim all or part of A's presentation.

The design of grounding is very important for a conversational agent. In human-human dialogue, when A does not understand B, A uses these ways to indicate the problems to B. In a conversational agent, if the system does not give the user an acknowledgment signal after processing the user's utterances, users are usually confused (Stifelman et al., 1993; Yankelovich et al., 1995).

Cooper and Larsson (1998) discussed advantage and disadvantage of optimistic grounding strategy and pessimistic grounding strategy. It is shown that middle ground between optimistic and pessimistic strategies is needed. Cahn and Brennan (1999) proposed a computational model of grounding based on Clark and Schaefer’s contribution model. In 2002, Larsson discussed feedback and grounding strategies available for the Issue-Based Information System.

In many studies of dialogue systems, grounding is not the main concern (Cormac, 2002). Although a lot of research has been done on ways to represent common ground, few have an overview and evaluation of different grounding strategies, from the perspective of the user’s experience of interacting with the system.

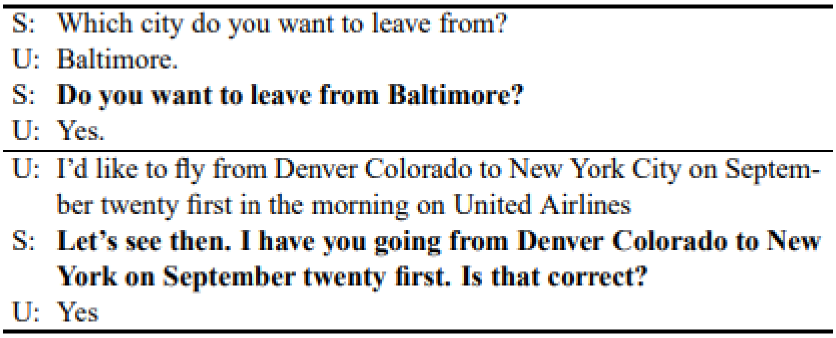
## 2.2 Feedback strategies

In a conversation, there will be many errors because of misunderstanding and mishearing. Humans use grounding to confirm that they understand correctly. Because of limitations in the system’s models, errors are inevitable in a dialogue system (Skantze, 2007). Although developers constantly improve dialogue systems and models, there are still many ambiguities and uncertainties. Compared with human-human dialogue, dialogue systems need to do more grounding (Delgado & Araki, 2005). This is usually done by feedback strategy (Jurafsky & Martin, 2004).

Feedback (known as confirmation or verification in other studies) is very important for a conversational agent. The most famous technique of grounding is explicit feedback and implicit feedback. The grounding strategies which are closest to human-human dialogue is no feedback. The grounding strategy used by most conversational agents is explicit feedback. This strategy will extend the transaction time, but when the speech recognition effect is poor, it is more reliable in terms of task success (McTear, 2004).

### 2.2.1 Interrogative and positive explicit feedback

According to Jurafsky & Martin (2004), explicit feedback uses the form of question to confirm mutual understanding. There are two types of explicit feedback. Here is the example (ibid.):



S: Which city do you want to leave from?
U: Baltimore.
S: **Do you want to leave from Baltimore?**
U: Yes.

U: I’d like to fly from Denver Colorado to New York City on September twenty first in the morning on United Airlines
S: **Let’s see then. I have you going from Denver Colorado to New York on September twenty first. Is that correct?**
U: Yes

**Figure 1. Two types of explicit feedback**

Both types of questions are yes-no questions. The first one uses a single sentence, the other uses declarative sentences followed by a short question like “Is that correct?”.

Interrogative explicit feedback can also be used after all the information have been collected (McTear, 2004). However, using this kind of interrogative explicit feedback will have difficulty when information is wrong. In the following example, it is hard to correct “Sunday” to “Friday” in one turn (ibid.):

S: Where are you travelling to?
U: London.
S: What day?
U: Friday.
S: **So, you want to travel to London on Sunday?**

Although the interrogative feedback plays an important role in a dialogue, it is not the subject of this paper. It is more interested in comparing these two different ways (explicit and implicit) of providing positive feedback. Therefore, the explicit feedback has been modified in this paper, that is, a system says "OK, x" corresponds to the utterance x given by user. Here is an example:

S: Where are you travelling to?
U: London.
S: **OK, London. What day do you want to leave?**
U: Friday.

The system with interrogative explicit feedback will increase the number of necessary turns, so the transaction time will also be increased. It can be seen from the above two examples that when the travel destination and departure time are also asked, interrogative explicit feedback spends four turns to confirm the destination, while positive explicit feedback obtains the destination and departure time in four turns.

### 2.2.2 Implicit feedback

Implicit feedback occurs when the system embeds the slot value extracted from user’s response in its next question (McTear, 2004). The system confirms to the user the understanding of the system by repeating what the user has said. Here is an example (Jurafsky & Martin, 2004):

U: I want to travel to Berlin.
S: **When do you want to travel to Berlin.**

It is worth noting that feedback utterance can also be placed in other places than at the end of the utterance, which may affect the evaluation results. In this experiment, feedback utterance is always placed at the end of the utterance. In human-human dialogue, no matter where the keywords are in the sentence, they will be emphasized by accenting the pronunciation. Therefore, to reduce the impact on results, TTS is used in this experiment to emphasize all feedback utterances.

This implicit feedback in this article is the same as the definition given before.

### 2.2.3 The relationship between positive explicit feedback and implicit feedback

It can be seen that these two grounding strategies both confirm the feedback utterance to the user in one utterance. However, positive explicit feedback confirms feedback utterance with users in a separate sentence. Implicit feedback embeds feedback utterance into the next utterance.

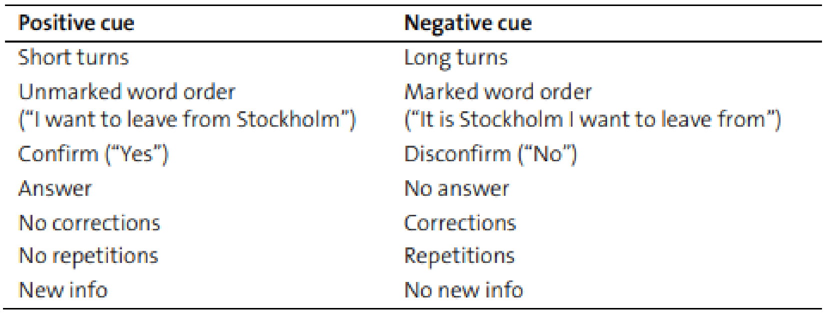
According to Jurafsky & Martin (2004), explicit feedback can help users find the system’s misrecognition more easily by simply answering “no”. However, explicit feedback also has a disadvantage. If explicit feedback appears all the time, it will make conversations boring and unnatural.

Conversely, with implicit feedback, the dialogue will be more natural. But for users, without the system's confirmation, it is easy to ignore the process of correction and difficult to find errors.

## 2.3 Response to wrong feedback

Skantze (2007) gave the definition of later error detection that it usually happens after a system displays a wrong feedback and a user initiates a repair. In the process of later error detection, it could be seen that problem signals may look very different and may depend on subtle prosodic cues, or the user may just ignore the problem.

Krahmer et al. (2001) listed possible positive and negative cues from the user after the system gives feedback.



| Positive cue | Negative cue |
|---|---|
| Short turns | Long turns |
| Unmarked word order<br>("I want to leave from Stockholm") | Marked word order<br>("It is Stockholm I want to leave from") |
| Confirm ("Yes") | Disconfirm ("No") |
| Answer | No answer |
| No corrections | Corrections |
| No repetitions | Repetitions |
| New info | No new info |

**Figure 2. possible positive and negative cues**

By analyzing a hand-labelled corpus based on spoken dialogue systems which provided train timetable information, Krahmer et al. (2001) found that users seldom gave negative cues, such as "no" when a system gives wrong feedback utterances. He also found that the prosody of correction is very different from non-correction, the pitch is higher, louder and longer, and the pause time before the correction is longer.

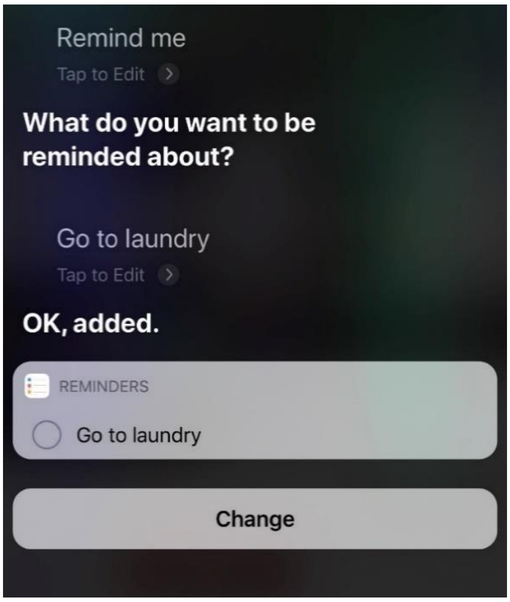
## 2.4 The state of the art of grounding strategies in conversational agents

In human-human dialogue, grounding is mostly implicit (O'Brien, 2002). In human-machine dialogue, the early dialogue systems tended to only use explicit or implicit feedback strategy. In 2001, San-Segundo stated that the feedback strategy has an important impact on the general performance of the system. A good feedback strategy can avoid asking the user redundant questions to confirm the words the user has said.

Recently, dialogue systems change the feedback strategy of each sentence according to different factors. The most important factor of the measurement is ASR performance (Jurafsky & Martin, 2004). For example, some dialogue systems refer to the acoustic confidence assigned to an utterance by the ASR system to determine which feedback strategy to use. Another important factor is the cost of errors (Jurafsky & Martin, 2004). For example, when a user uses the dialogue system to book a plane ticket, if any information is wrong, the consequence will be very serious.

Siri and Google Assistant are the most common conversational agents on the market (López et al., 2018). In the following paragraphs, the feedback strategies of these dialogue agents are described.

Siri is developed by Apple, which was released in 2010 (Hoy, 2018). It can receive many commands such as "set an alarm". When the utterance is clear and accurate, Siri usually uses implicit feedback. As shown in Figure 3, Siri accurately recognized the utterances "remind me" and embedded "remind" into the next utterance. In this case, the implicit feedback strategy is chosen. At the same time, the relevant information will be displayed on the screen. Conversely, when Siri is not sure about the user's utterance, a similarly pronounced word will be recognized to take the next action. In this case, the user can only start over and give the command again. Sometimes, there will be a red wavy line under the uncertain utterance. Users can click on the screen to modify what they have said.



Remind me

Tap to Edit

What do you want to be reminded about?

Go to laundry

Tap to Edit

OK, added.

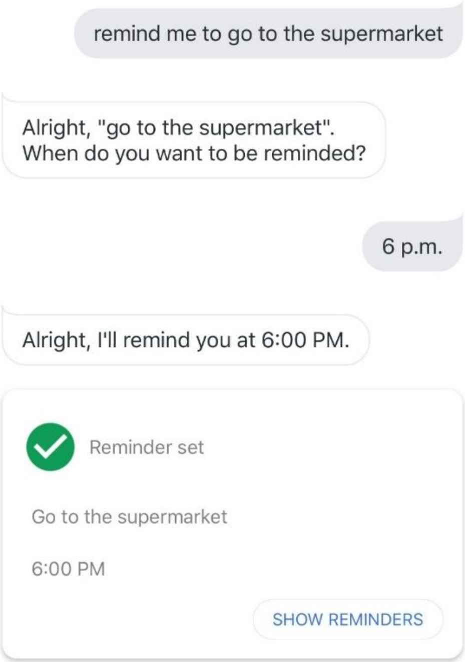
REMINDERS

○ Go to laundry

Change

**Figure 3. Communication with Siri**

Google Assistant is developed by Google, which is released in 2016. Besides some normal tasks like managing schedules, controlling smart home devices, Google has newly tested a feature that can use voice to confirm purchases (Hager, 2020). Google Assistant on mobile devices has similar feedback strategies to Siri. As shown in Figure 4, the system says “Alright, go to the supermarket” to respond to the feedback utterance. In this case, Google Assistant is confident about the user’s utterance and gives explicit feedback. If it can’t recognize what a user says, some words with similar pronunciation will be used to take the next action.



remind me to go to the supermarket

Alright, "go to the supermarket". When do you want to be reminded?

6 p.m.

Alright, I'll remind you at 6:00 PM.

Reminder set

Go to the supermarket

6:00 PM

SHOW REMINDERS

**Figure 4. Communication with Google Assistant**

Google has also developed a series of smart speakers called Google Home. Through these devices, users can simply speak commands to interact with Google Assistant. There are different types of google home, among which google home mini cannot give user verbal grounding through the text on the screen. Due to accent or other reasons, Google Home often misunderstands user's commands. For example, Google Home sometimes mishears the command “play music on **both** speakers” as “play music on **Bose** speakers”. It is really annoying to repeat the command and carefully enunciating a word.

This kind of smart speakers uses the system with no feedback so that the user is hard to correct an error in a conversation. Additionally, there is no screen to show any text, so users can only communicate with the system through voice. It is difficult to ground in the conversation. When the user listens to music on Google Home and says “next”, Google Home mishears as “necks”. But Google Home chooses no feedback strategy and just starts talking about the neck. In daily life, this kind of mishearing has little effect. However, it is necessary to use the feedback strategy to confirm the information when booking a flight ticket or depositing money through a voice-only smart speaker.

# 3. **Methodology**

*In this chapter, the method used in this paper is introduced. This paper uses experiments and human evaluation to draw conclusions.*

## 3.1 Experimental quantitative research

This paper uses experimental quantitative research. According to Aliaga and Gunderson (2002), quantitative research is explaining phenomena by collecting numerical data that are analyzed using mathematically based methods.

In this paper, a dialogue system was implemented in the experiment to find out which grounding strategy is better. The basis of the experimental method is an experiment conducted under controlled conditions to test the validity of the hypothesis (Muijs, 2011). When designing the experiment, only explicit and implicit strategies are compared. These two different strategies are given to users in the same situation which is the system gives three questions with two feedback utterances. One of the feedback utterances is correct and the other is wrong.

The questionnaire is a common method for quantitative research. Most social surveys are completed through questionnaires (Moser, 1984). In order to evaluate the two feedback strategies from the user's perspective, a questionnaire will be given to the participants after testing the system. Because of the way of dialing the phone number which is set beforehand, it is difficult to trace the participants themselves and ask for more opinions and suggestions.

This paper uses a mix of quantitative method and qualitative method. Although questionnaire is the only way of collecting data, subjective judgment is counted.

## 3.2 Human evaluation

Evaluating dialogue systems is difficult and challenging. It is also the subject of much research (Deriu, Agirre et al., 2020). Much progress has been made in the automated evaluation of dialogue system. Although human evaluation is time-consuming and laborious, it still exists and has its merits. Human evaluation can get real opinions from users, which is of great significance to the improvement of the system.

### 3.2.1 Field experiment

There are many ways to conduct human evaluation: lab experiment, field experiment and so on. In general, test subjects have the following main tasks: interacting with the system, rating dialogue or utterances, or both (Deriu, Agirre et al., 2020). The field experiment collects feedback from real users. Field experiments generate a large amount of data in a relatively short time. For the participants, the environment of the field experiment is relatively free, but at the same time, it will be easily distracted due to noise or other interference.

In 2013, Sun and May considered that field experiments are more suitable for studying various factors that affect the overall preference of a designed dialogue system. These factors include the impact of system function and actual usage context. An open and relaxed dialogue environment is important, this can be facilitated more easily by using the field experiment.

In this paper, the dialogue system was implemented on a telephony platform. Participants can dial the designated phone number to evaluate the system. Participants voluntarily evaluate the system, and they are free to choose the convenient time and place of the evaluation. In this experiment, participants not only need to interact with the system but also rate the dialogue.

### 3.2.2 Telephony platform: Voxeo

Systems implemented in a telephony platform can have real-time conversations with participants. Unlike other platforms like Apple Siri, participants cannot see the text content of the conversation. In this case, the participants will focus more on the spoken conversation.

The entire dialogue between the participants and the dialogue system needs to be recorded in order to verify data and collect user responses. Therefore, Voxeo was chosen in this paper. Voxeo provides an XML based telephony application platform with multiple flexible functions, including W3C standard and VoiceXML 2.0. This platform also supports call recording and direct-dial numbers. Recorded calls are stored on the Voxeo account management site and can be downloaded or played at any time. Unlike other services that require to enter a pin or extension code for each call, the feature called direct-dial numbers connects phone number to the VoiceXML application instantly. This feature not only helps participants enter the dialogue system more easily, but also helps save time during repeatedly calling the application for testing and debugging (Voxeo Document, 2019).

### 3.2.3 Subjective Assessment of Speech System Interface (SASSI)

In the process of evaluation, participants also need to rate the dialogue. The design of the questionnaire refers to Subjective Assessment of Speech System Interface (SASSI).

Subjective Assessment of Speech System Interface is a questionnaire measure of user satisfaction (Hone & Graham, 2001). For a speech-based interface, it is considered to be the most widely used questionnaire (Wechsung, 2014). SASSI has 50 statements and divided into 6 factors, system response accuracy, likeability, cognitive demand, annoyance, habitability, speed and feedback.

Since research users’ preferences for explicit and implicit feedback do not need to evaluate all six factors, the questionnaire has been adjusted and only two factors, likeability and habitability, have been evaluated. The questionnaire is as follows:

1. How natural the system is?
2. How easy is it when you communicate with the system?

The reason why the entire questionnaire has only two questions is that the entire evaluation process should not take too much time. During the evaluation process, participants scored two questions within a five-point scale. A question will be added at the end of the system: "Which dialogue do you like best", allowing participants to point out which one they think is best.

# 4. Experimental Setup

*In this chapter, the details of the experiment are described. The dialog design and implementation are explained in detail. Finally, the method of data collection and analysis are described.*

## 4.1 Dialogue design

### 4.1.1 Main dialogue

Two scenarios, a date scenario and a week scenario, are designed to test positive explicit feedback and implicit feedback. Each scene has three questions with two feedback utterances from the system, one of which is correct and the other is wrong. The two scenarios are as follows:

- Date scenario:

  S: What month is it today? [question 1]  
  U: **June** [utterance 1]  
  S: What date is it in **June**? [question 2]  
  U: **26th** [utterance 2]  
  S: What time is it on **1st**? [question 3]  
  U: … [utterance 3]

- Week scenario:

  S: What day of the week is it today? [question 4]  
  U: **Monday** [utterance 4]  
  S: Is it **Monday** morning or afternoon**?** [question 5]  
  U: **morning** [utterance 5]  
  S: What time in the **afternoon** is it? [question 6]  
  U: … [utterance 6]

Take the date scenario as an example, the user's utterance 1 “June” is the answer to question 1. The system embeds utterance 1 “June” into question 2. No matter what utterance 2 is, the system will always give a wrong feedback in question 3 and see the user's reaction when facing the wrong feedback of the system. At the same time, two feedback strategies can be evaluated when giving correct feedback and wrong feedback. Each feedback strategy has different performance in different scenarios. The result of evaluation will be fairer by testing both feedback strategies in both scenarios. Setting a correct feedback and a wrong feedback allows participants to judge the variable in different situations.

Scenarios and feedback strategies are paired into four combinations. Participants will randomly test one of the combinations after dialling the phone number. In other words, participants will test two dialogues, which have different scenarios and different variables. For example, the first dialogue to be tested is the date scenario with the positive explicit feedback, then the second is the week scenario with the implicit feedback. By disrupting the pairing of scenes and variables, participants can evaluate the performance of variables in the two scenarios as evenly as possible, so that the credibility of the data can be guaranteed.

### 4.1.2 Tutorial dialogue

In the process of debugging the dialogue system, the problem was found that new participants will feel confused about what they should do during the first dialogue. It takes some time for a new participant to get used to speech synthesis. At the same time, participants also need a tutorial to understand the purpose and process of the evaluation. Therefore, the tutorial dialogue is designed and given before the main dialogue.

Unlike the main dialogue, there is no feedback from the system in the tutorial dialogue. The main intention of the tutorial dialogue is to help participants familiarize themselves with speech synthesis and have a preliminary understanding of the process of the main dialogue.

The tutorial dialogue is as follows:

S: Hi, I am the robot. I will give an example so that you will learn what will happen in the formal task. Now, let's have a try.
**What's your gender? Male or female?**
U: …
S: OK. **How old are you?**
U: …
S: OK. **What's your educational background?**
U: …

By the way, by designing such a tutorial dialogue, it is easy to collect the background of the participants and analyze different aspects of data. However, participants were informed that their basic background would be collected at the beginning of the evaluation. If they do not agree to the collection of personal information, they could hang up at any time.

## 4.2 Implementation

Voice Extensible Markup Language (VoiceXML) is a technology that enables users to interact with systems through a voice browser or phone. It has a lot of functions, such as synthesized speech, digitized audio, speech recognition and DTMF key input, voice input recording, phone calls, and mixed active conversations (W3C, 2004).

Another programming language is JavaScript. It is an object-based programming language. Due to the popularity and utility, JavaScript has also been integrated into VoiceXML. JavaScript enables operations not supported by the VoiceXML language to be performed.

### 4.2.1 Randomize

After the tutorial dialogue, there will be two main dialogues that need to be evaluated. As it said before, scenarios and variables are paired into four combinations, which are date scenario with positive explicit feedback, date scenario with implicit feedback, week scenario with positive explicit feedback and week scenario with implicit feedback. During the evaluation process, different scenarios with different feedbacks will appear in the system at the same time to facilitate user evaluation. For example, if the first main dialogue is date scenario with implicit feedback, the second main dialogue should be week scenario with positive explicit feedback. Similarly, the order of the scenarios can also be different.

**Table 1. ID of these four combinations**

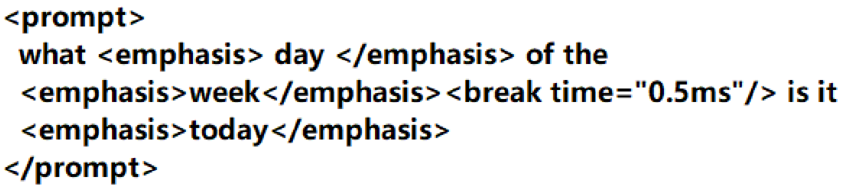
| ID | Combination |
|---|---|
| 1 | Date scenario with implicit + Week scenario with explicit |
| 2 | Date scenario with explicit + Week scenario with implicit |
| 3 | Week scenario with explicit + Date scenario with implicit |
| 4 | Week scenario with implicit + Date scenario with explicit |

Therefore, IDs were given to the four combinations. After the participants dialed the phone number, the system randomly selected a number, and the corresponding combination will be evaluated. The combination corresponding to id is shown in Table 1.

### 4.2.2 Audio output and TTS

Considering that only using text-to speech (TTS) or audio files in the system will affect the participants' ratings, the audio files which are recorded in advance are used in the introduction of the system and the process of evaluation, while the remaining tutorial and main dialogues used TTS.

TTS enables to convert text or Speech Synthesis Markup Language (SSML) into synthetic human speech. The text of the tutorial and main dialogue is shown in Chapter 4.1. Besides, in order to help understand the text well, SSML is used to enhance speech synthesis. In this paper, SSML is mainly used to insert pause and emphasize keywords. In Figure 5, the content is “what *day* of the *week* is it *today*” and the SSML code is as follows:



```
<prompt>
 what <emphasis> day </emphasis> of the
 <emphasis>week</emphasis><break time="0.5ms"/> is it
 <emphasis>today</emphasis>
</prompt>
```

**Figure 5. SSML in the code**

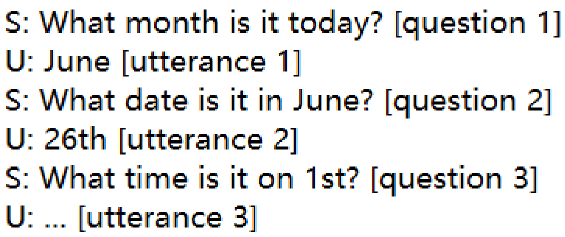
Seven audio files were recorded in advance and used to connected the testing and rating stages. The text content of these seven audio files is attached in Appendix 1. The format of the audio files was transferred to 8-bit 8Khz mu-law format. These files are stored in the same folder ‘audio’ and are operated by the following code:

```
<audio src="audio/bye.wav"/>
```

### 4.2.3 Audio input

In the experiment, the participants can provide input in two ways, spoken input and character input (DTMF). In the main experimental dialogue, participants can only use spoken input. However, in the process of rating, participants can choose to say the number to represent the score and they can also choose to enter the number 1-5 to score. In this experiment, the system's speech recognition ability was not considered. What’s more important is that the participants' ratings accurately expressed their preferences. Therefore, using DTMF in the rating process allows to receive more accurate results from participants by entering numbers.

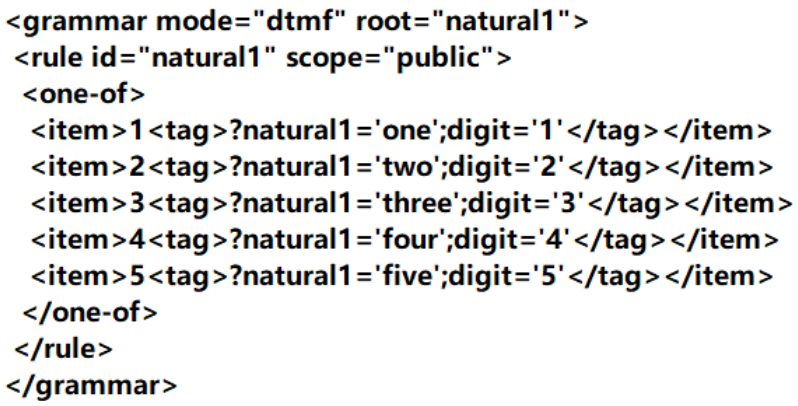
Participants perform spoken input by speaking and then the system converts what they say into text through speech recognition. Since the speech recognition capability of the platform is not really good, it is assuming that participants always answer the system questions correctly. All feedback is set in advance and changed every day in order to achieve the desired effect of the system, that is, all dialogues have a correct feedback and a wrong feedback. For example, for the participant in the scenario of Figure 6, today is 6/26. The two feedbacks of “June” and “1st” will change with the situation, “June” will always be the correct feedback, and “1st” will always be the wrong feedback.



S: What month is it today? [question 1]
U: June [utterance 1]
S: What date is it in June? [question 2]
U: 26th [utterance 2]
S: What time is it on 1st? [question 3]
U: ... [utterance 3]

**Figure 6. The date scenario**

DTMF enables users to enter numbers by dialing through phones. In VXML, the <grammar> element can provide a DTMF grammar to specify keys that users can use to perform actions or provide information (W3C, 2004). The mode attribute indicates to the VoiceXML interpreter whether the grammar is a voice or DTMF grammar. In this experiment, keys 1-5 are set to the corresponding score. The code is shown in Figure 7.



```
<grammar mode="dtmf" root="natural1">
 <rule id="natural1" scope="public">
  <one-of>
   <item>1<tag>?natural1='one';digit='1'</tag></item>
   <item>2<tag>?natural1='two';digit='2'</tag></item>
   <item>3<tag>?natural1='three';digit='3'</tag></item>
   <item>4<tag>?natural1='four';digit='4'</tag></item>
   <item>5<tag>?natural1='five';digit='5'</tag></item>
  </one-of>
 </rule>
</grammar>
```

**Figure 7. The code of DTMF**

### 4.2.4 Record

At the beginning of the call, the user will be informed that the entire conversation will be recorded. All recordings and personal backgrounds will be only used for research.

All participants’ interaction with the system will be fully recorded. Therefore, a Voxeo-specific element is used. Unlike the element, it enables to record both sides of a call. This element is more in line with the needs of the experiment. To use this element, the xmlns:voxeo attribute must be declared and mapped to the correct Voxeo namespace.

## 4.3 Data collection

In May 2020, this system was continuously tested and many details adjusted. From June 8th to 14th, the system was officially available. The phone number was posted on social networking sites such as Facebook. Many friends shared this post to help increase the number of participants. This is the snowball sampling method. This method helps reduce the time to find participants and increase the number as much as possible (Parker & Scott & Geddes, 2019). But there is also a problem with this method which is the background of the participants is relatively similar, such as age, education. There are no restrictions on participants. As long as participants can dial the phone number, they can evaluate the system.

In order to protect personal privacy, GDPR has set up "personal data", "data subject" and "data subject rights". It has also adopted strict personal privacy protection requirements. According to GDPR, user permission is needed to collect and use personal information.

In the process of evaluation, participants will be told that the entire conversation will be recorded, and some basic background such as gender will be collected. However, there is no information allowing identification of the caller was collected. Participants will also be promised that all data is only for the research of this paper and will not be misappropriated for other purposes. If a participant agrees, he can continue the process of evaluation. If he doesn’t agree, he can exit the conversation at any time.

After the participants finish the evaluation, the data is saved in the university server. The data is verified by listening to the recording to ensure the accuracy. After the research is accomplished, all data will be cleared from the server to prevent the disclosure of personal information. After data collection, the data was imported into a website to automatically generate diagrams.

## 4.4 Data analysis

The T-test is a kind of inference statistic used to determine whether there is a significant difference between the two groups of means, which may be related to certain characteristics (Kenton, 2020). It is suitable for data with a small sample size. Since there are only 22 participants to evaluate the dialogue system, T-test is used for analyzing the data.

There are three kinds of T-test, paired T-test, equal variance T-test and unequal variance T-test. Since the number of samples in explicit and implicit feedback is the same, equal variance T-test is chosen. It belongs to the independent T-test, which means that the data sets in the two groups don’t refer to the same values(ibid.).

The t-value measures the difference relative to the change in sample data. In other words, T is just the calculated difference expressed in standard errors.

The p-value is the probability of obtaining results as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. The larger the absolute value of the t-value, the smaller the p-value, and the more evidence against the null hypothesis (Minitab Blog, 2016). Generally, p<0.05 is considered statistically different, P<0.01 is considered to have a significant statistical difference, and P<0.001 is considered to have an extremely significant statistical difference.

Degrees of freedom is equal to the number of "observations" minus the number of required relationships between observations (Minitab Blog, 2016).

The calculation formula of the t-value and degrees of freedom is as follows:

$$\text{T-value} = \frac{mean1 - mean2}{\frac{(n1-1)\times var1^2+(n2-1)\times var2^2}{n1+n2-2} \times \sqrt{\frac{1}{n1}+\frac{1}{n2}}}$$

**Figure 8. The formula of the t-value**

$$\text{Degrees of Freedom} = n1 + n2 - 2$$

**Figure 9. The formula of degrees of freedom**

In Figure 8, mean1 and mean 2 are average values of each of the sample sets. Var1 and var 2 are variances of each of the sample sets. n1 and n 2 in both Figure 8 and Figure 9 are numbers of records in each sample set. The p-values are calculated using p-value tables in this paper.

## 4.5 Ethical considerations

Four ethical principles were strictly followed in the experiment: whether to cause harm to participants, such as physical injury; whether to lack informed consent; whether to infringe privacy; whether to involve deception (Diener & Grandall, 1978).

This paper only collects participants' basic background (gender, age and educational background) and voice recordings during the experiment which will not cause any harm to participants. There is no information allowing identification of the caller was collected.

Before participants evaluate the system, they were informed that some personal information will be collected and the entire conversation with the system will be recorded. Not only that, before the experiment started, participants also had a general understanding of the purpose of the experiment. At the same time, they will also be notified that the collected data is only used for research and the data will be stored on the university server, which meets the storage requirements of GDPR. The data will be cleared until the end of the research.

If the participant agrees to be collected personal data for research purposes, the evaluation will continue. If anyone has any discomfort during the experiment, they can hang up the phone at any time to interrupt the evaluation and data will not be retained.

The data is stored in the server. These data are for research only. After the study is over, all data will be erased. Even if the data come into the wrong hands, no information that can accurately identify the participant is collected. Although it may not affect the participants, if a data leak occurs, the information will be posted on social networks as soon as possible to remind participants who have participated in the experiment.

# 5. Results

*In this chapter, the results of the questionnaire are summarized. Participants' reactions to wrong feedback are also summarized.*

## 5.1 Questionnaires

From June 8th to 14th, 22 people participated in the evaluation of this dialogue system and rated the two feedback strategies. The diagram is generated from data on Sojump (an efficient questionnaire website that can automatically generate diagrams). In this system, not only the evaluation data of the feedback strategies is collected, but also the gender, age and educational background of the participants are collected.

### 5.1.1 Background of participants

As shown in Figure 10, the background of the participants is summarized. The proportion of female is much higher than the proportion of male, reaching 16/22 (72.73%). There are still male participants in this experiment, accounting for 6/22 (27.27%).

The age of the participants is concentrated in the 21-25 years old, accounting for 68.18%. At the same time, 6/22 (27.27%) of the participants are 26-30 years old and 1/22 (4.55%) are 31-35 years old. No participants in other age groups participated in the experiment.

In terms of educational background, 14/22 (63.64%) of the participants has a master's degree. Followed by participants with a bachelor degree, accounting for 7/22 (31.82%). Participants with a high school degree are the least, accounting for only 1/22 (4.55%).

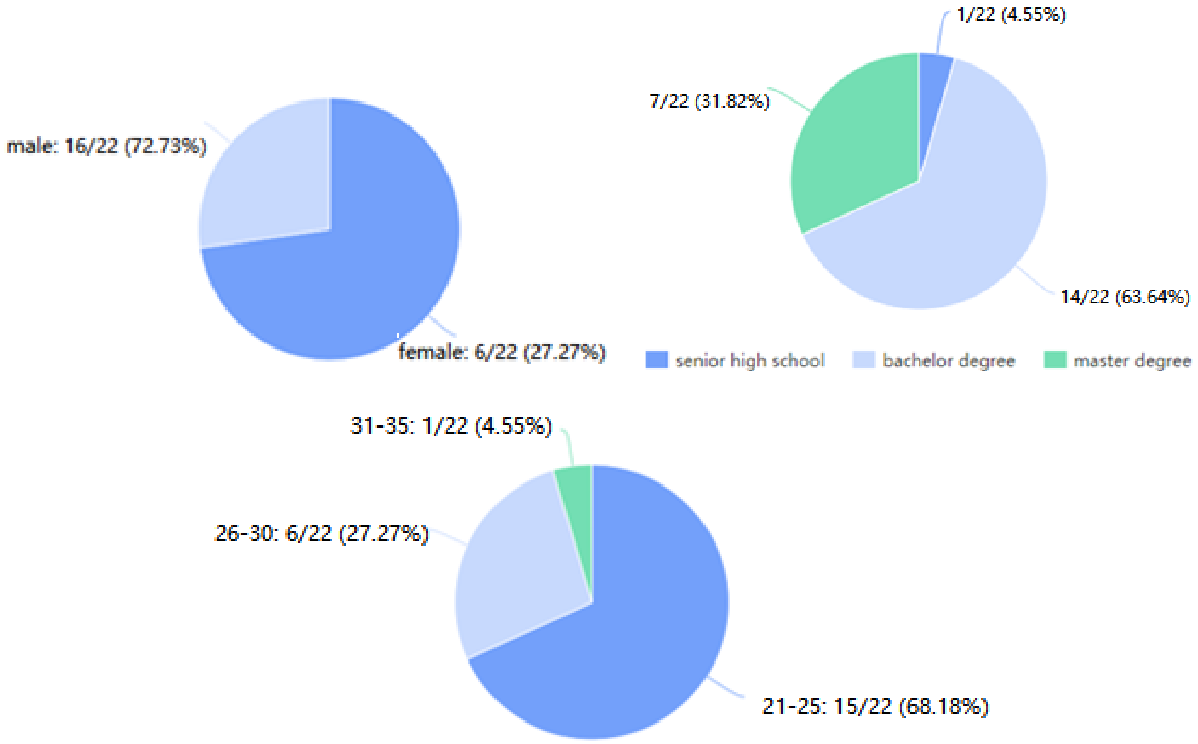


**Figure 10. background of participants**

### 5.1.2 The evaluation of dialogues with positive explicit feedback

In general, most participants are satisfied with positive explicit feedback. As shown in Figure 11, when they were asked how natural the system is, nearly half of the participants considered that the system with positive explicit feedback is natural and gave 4 points. Meanwhile, 6/22 (27.27%) participants thought it is average and gave 3 points. Additionally, no one gave 1 point and only 3/22 (13.64%) participants gave 2 points. The average score for this question is 3.59.

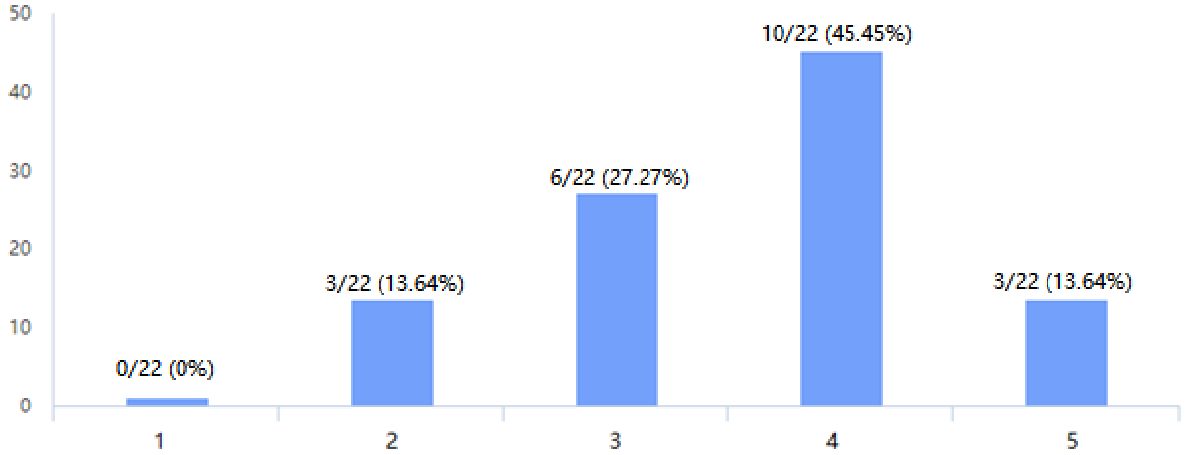


**Figure 11. The evaluation of whether the system with positive explicit feedback is natural**

When faced with the next question "How easy is it to communicate with the system?", the participants' views are almost the same. They thought this system is easy and gave a score of 4 or more. Only 3/22 (13.64%) of the participants gave 3 points. What’s more interesting is that no one gives a score below 2. The result is shown in Figure 12. The average score for this question is 4.27.

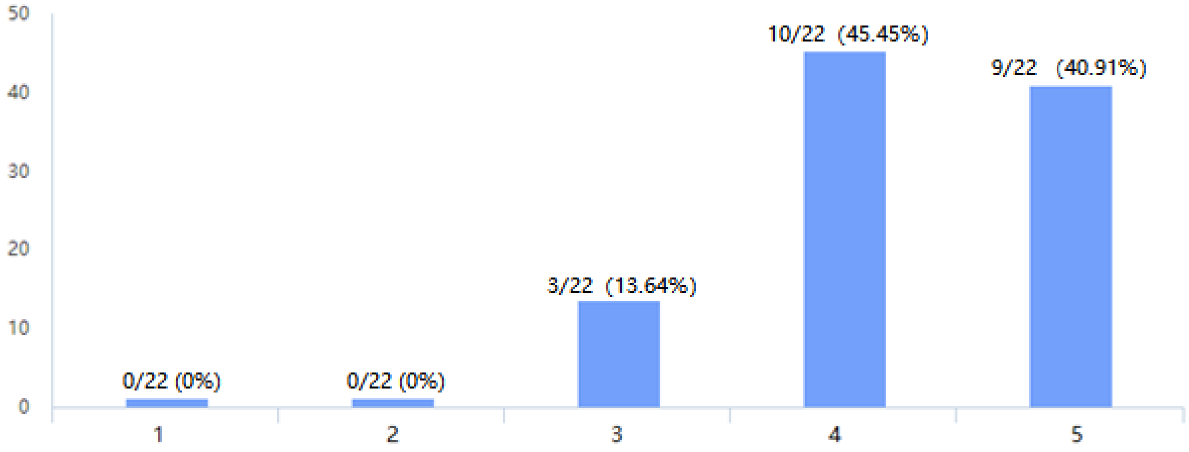


**Figure 12. The evaluation of how easy is the system with positive explicit feedback**

### 5.1.3 The evaluation of dialogues with implicit feedback

Compared with the result of positive explicit feedback, the scores given by participants to implicit feedback are quite different. Although there are still a large number of participants who give scores of more than three points, some are not very satisfied with this strategy.

According to Figure 13, when the participants were asked how natural the system is, the same result as the positive explicit feedback was that 10/22 (45.45%) of the participants gave 4 points. The difference is that the same percentage of participants felt average or unnatural. Only 2/22 (9.09%) participants thought that the system with implicit is very natural and gave 5 points. The average score for this question was 3.27.

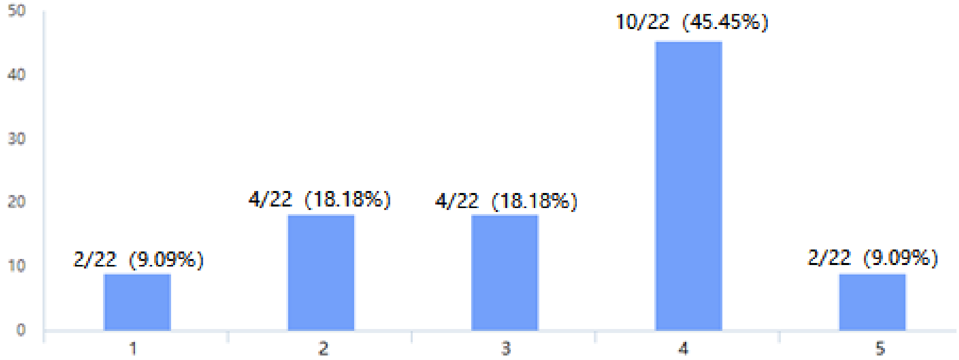


**Figure 13. The evaluation of whether the system with implicit feedback is natural**

When facing the next question, most of the participants also thought the system with implicit was easy. According to Figure 14, 15/22 (68.18%) participants gave a score of 4 or more. Even 10/22 (45.45%) of participants thought this system was very easy. But there was still 6/22 (27.27%) of the participants thought the system is average and 1/22 (4.55%) of the participants found it uneasy. The average score for this question was 4.09.

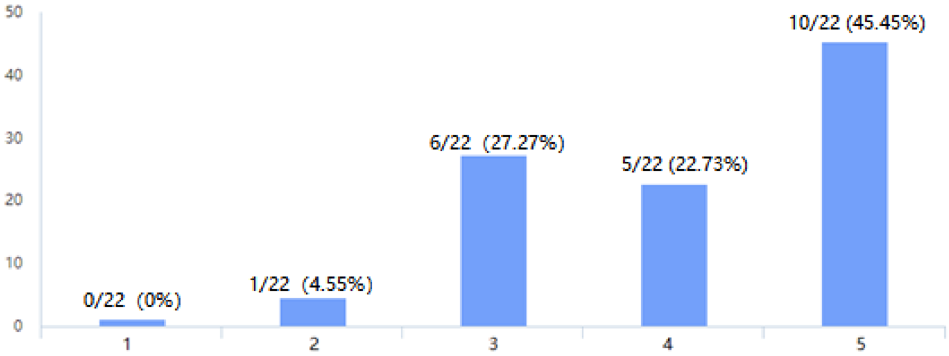


**Figure 14. The evaluation of how easy is the system with implicit feedback**

At the end of the call, the participants were asked which dialogue they preferred. Most participants chose the system with positive explicit feedback. According to Figure 15, only 7/22 (31.82%) participants prefer implicit feedback.

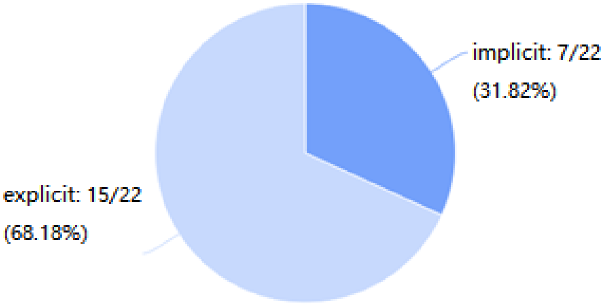


**Figure 15. participants' favourite strategy**

### 5.1.4 The result of T-test

This paper wants to find out which feedback strategy is better. Therefore, the ratings of positive explicit feedback and implicit feedback were analyzed by the T-test. In terms of naturalness, the t value is 0.74. The p value is 0.232 which is larger than 0.05. It shows that there is no significant difference in naturalness between these two feedback strategies. In terms of ease, the t value is 0.55. The p value is 0.291 which is larger than 0.05. It shows that there is also no significant difference in ease between these two feedback strategies.

Considering that the results may be affected by the order of appearance or scenes, the data is classified and analyzed. According to Table 2 and Table 3, There was also no significant difference in ease and naturalness. Similarly, the order of appearance and the scenario does not affect the results.

**Table 2. the result of the order classification**

| Order | Factor | T value | P value |
|---|---|---|---|
| Implicit feedback appeared first | Naturalness | 0.63 | 0.26 |
| | Ease | -0.38 | 0.35 |
| Explicit feedback appeared first | Naturalness | 0.39 | 0.34 |
| | Ease | 1.29 | 0.10 |

**Table 3. the result of the scenario classification**

| Scenario | Factor | T value | P value |
|---|---|---|---|
| Date scenario | Naturalness | -0.03 | 0.48 |
| | Ease | 0.10 | 0.45 |
| Week scenario | Naturalness | 1.03 | 0.15 |
| | Ease | 0.54 | 0.29 |

Although there were no significant differences found in the data, this does not mean that there are no differences. Because of the small sample size, the difference between the two feedbacks may not be so obvious.

## 5.2 The user’s response to wrong feedback

In the last question of each dialogue, participants were given wrong feedback by the system. Among the 22 participants, 6 participants did not respond to any wrong feedback of the two dialogues at all. In addition, there were five other participants who did not react to the wrong feedback of the system with implicit feedback.

In the system with positive explicit feedback, most participants first gave negative words (such as "no" and "nah") to reject the wrong feedback given by the system and then correct it. There are also some participants who use hesitation sounds (such as "em" and "eh") to doubt about wrong feedback and organize sentence to correct it. A few people made corrections directly without any negative words.

In the system with implicit feedback, nearly half of participants did not respond to wrong feedback. Most participants used question words (such as “what?”) or repeated the utterance of wrong feedback to raise doubts and then made correction or pointed out errors. Some participants used negative sentences to point out wrong feedback. There are also a small number of participants who directly make corrections. The complete set of user responses to wrong feedback is attached in Appendix 2.

# 6. Discussion

*In this chapter, the research questions raised in the first chapter are discussed. Mainly compared two feedback strategies and proposed what kind of strategy is a good grounding strategy. Users' response to wrong feedback is discussed.*

## 6.1 Comparison of two feedback strategies

According to the results of Chapter 5, it can be seen that most participants prefer positive explicit feedback. Although the scores of both feedback strategies are high (above 3 points), the average score of positive explicit feedback is higher.

In terms of naturalness, the average score of positive explicit feedback is 3.59 and the average score of implicit feedback is 3.27. Based on Jurafsky & Martin (2004), explicit feedback is less natural than implicit feedback. According to the average score, the result is different from Jurafsky & Martin. The explicit feedback in this article is slightly different from Jurafsky & Martin's. Although both of them are explicit feedback, they still have subtle differences. As mentioned in section 2.2.1, the system with interrogative explicit feedback will increase the number of necessary turns, which will make the conversation unnatural. Thus, at the same number of turns, the system with positive explicit feedback is more natural than the implicit one.

In terms of ease, the average score of positive explicit feedback is 4.27 and the average score of implicit feedback is 4.09. It can be seen that in terms of ease, participants are satisfied with the two feedback strategies. Additionally, participants found it easier to interact with the system with positive explicit feedback. This conclusion is consistent with Jurafsky & Martin's (2004) view. The system with implicit feedback makes it harder to detect and correct the wrong feedback. According to Appendix 2, it can be seen that there are more participants who did not respond to wrong feedback in the system with implicit feedback.

Although both positive explicit feedback's and implicit feedback's scores of naturalness and ease are high, the results of the T-test show that there is no significant difference between positive explicit feedback and implicit feedback. Regardless of the order of appearance or the scenario, they do not affect the result. One of the reasons may be that the sample size is too small.

In addition, the positive explicit feedback strategy in this paper has some limitations. In simple scenarios, when positive explicit feedback is repeatedly given, the unnatural shortcomings may not be obvious. In this paper, there are only two sentences in the entire dialogue that the system gives feedback. In addition, users cannot make real corrections in the system of this paper. Once the user can make corrections, the system with positive explicit feedback will continue to provide feedback of utterance. The dialogue will continue unless the feedback is correct. The conversation will be boring and unnatural. For example:

S: what date is it today?
U:10th.
S: **Ok,1st**. What time is it now?
U: no, it's 10th.
S: **Ok,8th**. What time is it now?
U: not 8th! I said 10th.

Therefore, it is suggested that a more complex system can be implemented in future research. Perhaps it can make the feedback strategy more natural, that is, interchange "got it" or similar words with "OK" in the system. At the same time, the related functions of the system should be improved, such as correction. If it could be evaluated in a real dialogue system, a more comprehensive conclusion may be drawn.

## 6.2 User response when communicating with the system

As mentioned in section 1.1, few researchers have studied grounding strategies from user experience in previous studies, especially the user's reaction to the system in different situations. In this paper, users' reactions to wrong feedback were collected.

From the collected responses, there are four kinds of reactions after the user receives the wrong feedback, that is, hesitation (or pause), repetition, point out the wrong feedback and correction.

Hesitation refers to the pause that the user generates at the moment of receiving the wrong feedback. It can be seen that participants spend less than a second on average thinking about the wrong feedback. This is generally reflected in some hesitation sounds, such as "em" and "eh". It is guessed that in this short hesitation time, the user compares what he said with the feedback received, confirms that it is wrong and organizes the sentence.

Repetition is that user repeats wrong feedback, for example, "The first?" and "Afternoon?". According to the data collected in this experiment, most of repetition is in a rhetorical sentence.

Point out the wrong feedback is generally reflected in some negative words such as "no", "not", "nah". Correction is that user corrects the mistake by saying the correct utterance.

These four reactions may appear in combination. It can be seen that users rarely correct wrong feedback directly. Most users combine hesitation or repetition and then point out the wrong feedback. Understanding these responses can be used to improve future dialogue systems. More reactions can be collected in future research.

According to Krahmer et al. (2001), there are seven kinds of negative cues: long turns, marked word order, disconfirm, no answer, corrections, repetitions and no new info. Their views are basically the same as this article, but the method of classification is slightly different. Among these factors, the concept "hesitation" in this paper is consistent with their concept of "long turns". Meanwhile, "point out the wrong feedback" is the same concept as "disconfirm". The concept of "repetition" is the same as Krahmer et al.

The difference is that this paper integrates the two concepts of "marked word order" and "correction" from Krahmer et al. into one concept "correction". As an example of this concept in Figure 2, the marked word order is "It is Stockholm I want to leave from". Similar sentences appeared in the data in this paper, such as "it's afternoon.". Perhaps the scenario in the experiment is too simple so that participants do not need to change the order of words when making a correction. The concept "no new info" is also involved in "correction" in this paper. No new info means that users speak correct feedback utterance one more time. This is also the definition of "correction" in this paper. Additionally, "no answer" from Krahmer et al. is considered that participants didn't recognize the wrong feedback utterance.

Because Krahmer et al. (2001) uses a hand-labeled corpus and this paper collects data from participants. As expected, the conclusion is slightly different. In the future, the number of samples can be increased to improve the summary of user responses to wrong feedbacks.

# 7. Conclusion

*In this chapter, the general content of this article is summarized. The research questions and conclusions of this paper are described. The research limitations of this article and future research directions are proposed*

The main aim is to compare two grounding strategies, explicit feedback and implicit feedback. This paper provides a method of choosing which grounding strategy to implement in a conversational agent. It can find a suitable grounding strategy by implementing a system with different grounding strategies and evaluating it by users.

Explicit feedback in this paper is that system says "OK, x" corresponds to the utterance x given by user which is different from previous studies. Implicit feedback is that system embeds the utterance x in to next utterance which is the same as previous studies. Users evaluated these two feedback strategies by two factors, naturalness and ease. The following two questions are given to users to rate. Data is collected by recording the entire conversation between the user and the system.

- How natural is the system?
- How easy is it when communicating with the system?

## 7.1 Research questions

For the first research question “Which grounding strategy is better?”, it was found that most users preferred explicit feedback. According to the average score, explicit feedback in this paper is more natural than implicit feedback. At the same time, users can easily communicate with systems with explicit feedback. Compared with implicit feedback, explicit feedback makes it easier for users to find wrong feedback. However, the results of T-test show that there is no significant difference between two grounding strategies.

For the second research question “How does user response to wrong feedback?”, four types of reactions are summarized by observing users’ response to wrong feedback. These four reactions appear through related words.

- Hesitation: the pause that the user generates at the moment of receiving the wrong feedback. E.g. “em”, “eh”.
- Repetition: user repeats wrong feedback. E.g. “The first?” and “Afternoon?”.
- Point out wrong feedback: point out mistakes through negative words. E.g. “not”, “no”.
- Correction: speak the correct utterance

This paper compared two grounding strategies from the perspective of user experience. It also provides a method for how to choose the grounding strategy in the future. At the same time, it also analyzes the user's response to the wrong feedback and as a reference for improving the dialogue system in the future.

## 7.1 Limitation and future work

First of all, it is a pity that p value is bigger than 0.05. Although the score of explicit feedback are higher than implicit, they still have no significant difference. One of reasons maybe the small sample size. Because of the time, only 22 participants evaluated the dialogue system. The sample size can be increased in future research and more participants can be invited to evaluate the dialogue system.

Secondly, it is mentioned in section 6.1 that the system is simple and the shortcomings of explicit feedback may not be fully shown to users. The system of this paper only gives explicit feedback in two utterance. Although every sentence will have "ok, x" in response to the utterance x, the conversation is not so unnatural and boring. Considering that the results may be different in a more complex dialogue system, it is recommended to use a complex dialogue system or a real dialogue system to compare grounding strategies in future research.

Thirdly, participants are found through social networks, not randomly selected. From the background of the collected participants, it can be seen that most of them are master’s degree, and the age is also concentrated in 21 to 25 years old. In the future, it is expected to obtain data from users of different backgrounds

Fourthly, this article uses questionnaires to collect data. This is indeed a time-saving method. In order not to make the evaluation time too long, there are only two questions in the questionnaire. In the future, the number of questions in the questionnaire can be increased, which allows users to score the system from different factors. At the same time, interviews are also recommended. It can get more suggestions and opinions from the participants.

Finally, it could also be considered to collect ratings of correct and wrong feedback separately, which might help to choose strategies based on how often the systems get things wrong.

# Reference list

Aliaga, M. and Gunderson, B. (2002) Interactive Statistics. [Thousand Oaks]: Sage

Allwood, J. (1995). An activity-based approach to pragmatics. Technical Report (GPTL) 75, Gothenburg Papers in Theoretical Linguistics, University of Goteborg.

Arora, S., Batra, K., & Singh, S. (2013). Dialogue System: A Brief Review. *ArXiv, abs/1306.4134*.

Batliner, A., Seppi, D. Steidl, S., & Schuller, B. (2010). Segmenting into adequate units for automatic recognition of emotion-related episodes: A speech-based approach. *Advances in Human Computer Interaction*, 2010. http://dx.doi.org/10.1155/2010/782802

Bui, T. H. (2006). *Multimodal dialogue management—state of the art*. Technical Report TR-CTIT-06-01, University of Twente.

Burgan, D. (2017). Dialogue Systems & Dialogue Management. [Internet] Available from: https://www.dst.defence.gov.au/sites/default/files/publications/documents/DST-Group-TR-3331.pdf [Accessed from 3rd June, 2020].

Cahn, J. E., & Brennan, S. E. (1999). *A Psychological Model of Grounding and Repair in Dialog*. Proceedings of the Fall 1999 AAAI Symposium on Psychological Models of Communication in Collaborative Systems. Sea Cliff, Massachusetts. November 5-7, 1999. Pages 25-33.

Chen, H., Liu, X., Yin, D., & Tang, J. (2017). A Survey on Dialogue Systems: Recent Advances and New Frontiers. *ArXiv, abs/1711.01731*.

Clark, H. H. (1996). *Using Language*. Cambridge University Press, Cambridge.

Clark, H. H., & Brennan, S. E. (1991). *Grounding in communication.* In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (p. 127–149). American Psychological Association.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22(1), 1–39.

Cooper, R., & Larsson, S. (1998). *Dialogue Moves and Information States*. Proceedings of the Third International Workshop on Computational Semantics (IWCS-3).

Cormac, O. B. (2002) Grounding Strategies in Dialogue Systems. [Internet] Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.19.5454 [Accessed from 3rd June, 2020].

Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 20(1), 30–42. http://dx.doi.org/10.1109/ TASL.2011.2134090

Delgado, R.L., & Araki, M. (2005). Spoken, Multilingual and Multimodal Dialogue Systems: Development and Assessment. United Kingdom.

Deriu, J., Rodrigo, A., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E., & Cieliebak, M. (2020). Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 1 - 56.

Google Design Guidelines. (2018) Available from: https://designguidelines.withgoogle.com/conversation/conversational-components/confirmations.html# [Accessed from 3rd June, 2020].

Hager R. (2020). Google confirms new voice-confirmation feature for purchases in Assistant. Available from: https://www.androidpolice.com/2020/05/25/google-assistant-gets-new-confirm-with-voice-match-setting-for-payments/ [Accessed from 4th June, 2020]

Hone, K.S., & Graham, R. (2001). Subjective assessment of speech-system interface usability. *INTERSPEECH.*

Hoy M. B. (2018). Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. *Medical reference services quarterly*, 37(1), 81–88. Available from: https://doi.org/10.1080/02763869.2018.1404391

International Data Corporation (2018) *Conversational artificial intelligence white paper-awakening everything, a new era of human-computer interaction* [Internet] Available from: http://www.rpa-cn.com/zixunyuanliku/xingyeyanjiubaogao/2019-10-14/1476.html [Accessed from 29th May, 2020].

Jurafsky, D., Martin, J. (2008). Dialogue System and Chatbots. *Speech and Language Processing: An introduction to natural language processing, Computational linguistics, and speech recognition.*

Kenton, W. (2020) T-Test [Internet] Available from: https://www.investopedia.com/terms/t/t-test.asp [Accessed from 26th June, 2020]

Larsson, S. (2002). *Issue-based Dialogue Management.* [Internet] Available from: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=8A179D8B472CD7AA09561949F8440A71?doi=10.1.1.420.8452&rep=rep1&type=pdf [Accessed from 4th June, 2020]

Litman, D., Pan, S. (2002). Designing and Evaluating an Adaptive Spoken Dialogue System. *User Modeling and User-Adapted Interaction*. Vol. 12, No. 2/3, pp. 111-137, 2002.

López G., Quesada L., Guerrero L.A. (2018) Alexa vs. Siri vs. Cortana vs. Google Assistant: A Comparison of Speech-Based Natural User Interfaces. *Advances in Intelligent Systems and Computing*, vol 592.

López-Cózar, R., Callejas, Z., Griol, D., & Quesada, J. F. (2014). Review of spoken dialogue systems. *Loquens, 1(2).*

MarketsandMarkets (2018) *Conversational Systems Market by Component (Compute Platforms, Solutions, Services), Type (Voice and Text), Application (Customer Support and Personal Assistant, Branding and Advertisement, and Compliance), Vertical, and Region - Global Forecast to 2024* [Internet]. Available from: https://www.marketsandmarkets.com/Market-Reports/conversational-systems-market-232318863.html [Accessed from 29th May, 2020].

Minitab Blog (2016) What are degrees of freedom in statisticis? [Internet]. Available from: https://blog.minitab.com/blog/statistics-and-quality-data-analysis/what-are-degrees-of-freedom-in-statistics#:~:text=For%20a%201%2Dsample%20t,values%20for%20the%20t%2Dtest. [Accessed from 29th May, 2020].

Muijs, D. (2011). *Doing quantitative research in education with SPSS*. London, GB. SAGE Publications (In Press)

Ken Research (2019) *Conversational System Market - Global Drivers, Restraints, Opportunities, Trends, and Forecasts up to 2023* [Internet]. Available from: https://www.kenresearch.com/technology-and-telecom/it-and-ites/conversational-system-market/173165-105.html [Accessed from 29th May,2020].

McTear, M.F. (2004). Spoken dialogue technology: toward the conversational user interface. Ulster University.

Moser, C. A., Kalton, G. (1985) *Survey Method in Social Investigation*. London: Heinemann Educational.

Parker, C., Scott, S., Geddes, A. (2019) *Snowball Sampling*. SAGE Research Methods Foundations. doi:10.4135/

San-Segundo-Hernández, R., Montero-Martínez, J.M., & Pardo, J.M. (2001). Designing Confirmation Mechanisms and Error Recover Techniques in a Railway Information System for Spanish. *SIGDIAL Workshop*.

Skantze, G. (2007). Error Handling in Spoken Dialogue Systems: Managing Uncertainty, Grounding and Miscommunication (PhD dissertation). KTH, Stockholm. Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-4521

Stifelman, L. J., Arons, B., Schmandt, C., and Hulteen, E. A. (1993). VoiceNotes: A speech interface for a hand-held voice notetaker. *Human Factors in Computing Systems: INTERCHI '93 Conference Proceedings*, pp. 179–186. ACM.

Sun, X., & May, A. (2013). *A comparison of field-based and lab-based experiments to evaluate user experience of personalised mobile devices (Version 1)*. Loughborough University. Available from: https://hdl.handle.net/2134/21372

Voxeo Document. (2019). Available from: https://help.voxeo.com/go/help/docs [Accessed from 26th June, 2020]

Wechsung, I. (2014). An Evaluation Framework for Multimodal Interaction - Determining Quality Aspects and Modality Choice. *T-Labs Series in Telecommunication Services*.

W3C Recommendation. (2004). Voice Extensible Markup Language (VoiceXML) Version 2.0. Available from: https://www.w3.org/TR/voicexml20/ [Accessed from 26th June, 2020]

Yankelovich, N., Levow, G.-A., and Marx, M. (1995). Designing SpeechActs: Issues in speech user interfaces. In Human Factors in Computing Systems: CHI '95 Conference Proceedings, Denver, CO, pp. 369–376. ACM.

Zhao, Y., Li, Y.L., & Lin, M. (2019). A Review of the Research on Dialogue Management of Task-Oriented Systems. Journal of Physics Conference Series, 1267, 012025

# Appendix 1. The content of audio files

### Instruction.wav

Hi, thank you for calling! All calls into this system are recorded just for my thesis. In the process, it would be better not to use the speak louder. Next, you will evaluate two dialogues and see how the bot responds to you. After that, you will be asked to rate them separately. In order to help you to learn this system well, there will be a tutorial. Now, let's have a try.

### rateinstruction.wav

Thanks for testing the first dialogue. Next, I will give you two statements. Please rate them from 1 to 5. The higher the score, the more you agree with this statement. You can also enter numbers through the keyboard.

### natural.wav

1. The interaction with system is natural.

### easy.wav

I got it. 2. It is easy to communicate with the system and find out the wrong feedback.

### thanks.wav

Thanks for rating. Let's start next dialogue

### rate2.wav

Thanks. Please rate this dialogue.

### best.wav

OK. Which one do you like best?

### bye.wav

Thanks for your help. Bye.

# Appendix 2. Users’ response to wrong feedback

| No. | Explicit feedback | Implicit feedback |
|---|---|---|
| 1 | It’s afternoon. | It is not the 1st. |
| 2 | Eh… It’s eighth. | / |
| 3 | / | / |
| 4 | I said “afternoon” | It’s ninth. |
| 5 | / | / |
| 6 | Emm… | What? |
| 7 | / | / |
| 8 | / | / |
| 9 | No, it’s 9th. | Afternoon? |
| 10 | Nah. It's not morning. | Not 9th. |
| 11 | Today is 10th, not 1st. | / |
| 12 | No..no, not 1st. I said 10th. | Em… it’s morning. |
| 13 | Er..it’s afternoon. | / |
| 14 | Wait a moment. I said “afternoon”. | 1st? I said 10th. |
| 15 | Nah, It’s in the morning. | / |
| 16 | / | / |
| 17 | Eh..morning, not afternoon. | Pardon? It’s 11th. |
| 18 | No, it isn’t 1st. | It’s half past (…). Eh..sorry but it isn’t morning.. |
| 19 | Nah-nah, it’s “afternoon”. | Eh… not 12th. |
| 20 | / | / |
| 21 | No, I said 13th. | I said “morning”. |
| 22 | Eh… | / |