Bauhaus-Universität Weimar
Faculty of Media
Degree Programme Computer Science and Media

Query Clarification in Voice Search

The case of ambiguous terms and false memories

Master's Thesis

Arefeh Bahrami Mirabadi
Born May 4, 1987 in Esfahan, Iran

Matriculation Number 115602

1. Referee: Prof. Dr. Benno Stein
2. Referee: Prof. Dr. Matthias Hagen

Submission date: September 28, 2018

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Weimar, September 28, 2018

.....
Arefeh Bahrami Mirabadi

Abstract

Despite the widespread use of voice-based search, few studies have addressed its complications. Specifically, clarifying user intent in the voice-based setting is not as straightforward as in the text-based one. In this thesis, we tested systems to resolve a user's intent for two types of unclear queries in a voice-based search setting: queries that are *ambiguous* and queries that contain *false memories* (i.e., wrong details that the user misremembered about the item they are searching for. To this end, we identified three key research questions for voice query clarification regarding adaptations to the user's background, optimal query clarification options in terms of their number, length and phrasing as well as how they impact user satisfaction.

In the first part of this work, we designed and conducted the first user study that measures user satisfaction for clarification options, which we presented with seven different methods. Our findings include that (1) user satisfaction depends significantly on language proficiency levels, (2) user experience of the system does not decrease when the system asks for clarifications, and (3) the most effective way of query clarification depends on the number and lengths of the possible answers.

In the second part of this thesis, we present the first user study on how voice-based search systems may communicate false memory corrections to their users. We designed a study in order to estimate user satisfaction for three different false memory clarifications methods and one baseline case where the system only answers with "sorry, I don't know that one!". Our findings indicate that (1) users are more satisfied when they receive a clarification that and how the system corrected a false memory, (2) users even prefer failed correction attempts over no such attempt, and (3) the tone of the clarification has to be considered for the best possible user satisfaction as well.

Finally, in the last part of this thesis, we compare the two studies and provide design implications for voice assistants in clarifying a user's intent. Our observations indicate that voice assistants should allow users to interrupt the system in order to identify their information need whenever they want, that the user's English proficiency should be taken into account based on the level of interaction the user and the system are supposed to have, and that it is important to consider the tone of responses when clarifying false memory.

Contents

Acknowledgements

First and foremost, I would like to give a very special gratitude to my supervisor Johannes Kiesel. Without his dedication, help and support, this work would have never been achieved.

My sincere appreciation to Prof. Dr. Matthias Hagen for his supervision and support.

I would like to thank Prof. Dr. Benno Stein for accepting my work under his supervision.

In addition, a thank to all my friends who were always around to help, especially Michael Schindler, Masoud Allahyari, Jaleh Babajani, Negin Yaghoubsharif, Fahime Same, and Milad Alshomary.

I also would like to thank the volunteers who participated in our user studies.

Last but by no means least, I am highly grateful to my father, mother, siblings and niece who have supported me along the way.

Chapter 1

Introduction

Interpreting the information need of a query is an essential task in information retrieval (Wang and Agichtein [2010]). Studies show that a great number of web queries are often short and/or ambiguous (Cui et al. [2003], Dou et al. [2007], Jansen et al. [2000]). Traditional information retrieval systems help users to clarify their information need by query reformulation suggestions. Many commercial search engines provide the *did you mean* functionality that seeks to clarify user's intent. Furthermore, there has been considerable amount of research on how to suggest and complete queries to assist users in clarifying their intent (Bhatia et al. [2011]). Generally, traditional information retrieval systems interact with users by offering alternative suggestions to their search query and displaying a list of possible results. As soon as the user begins to enter their query in the search box of a search engine, relevant suggestions are displayed. Moreover, after entering the query, alternative query suggestions are offered on the screen (see figure 1.1). They can be either corrected grammatical problems and misspelled words or expanded contextual suggestions.

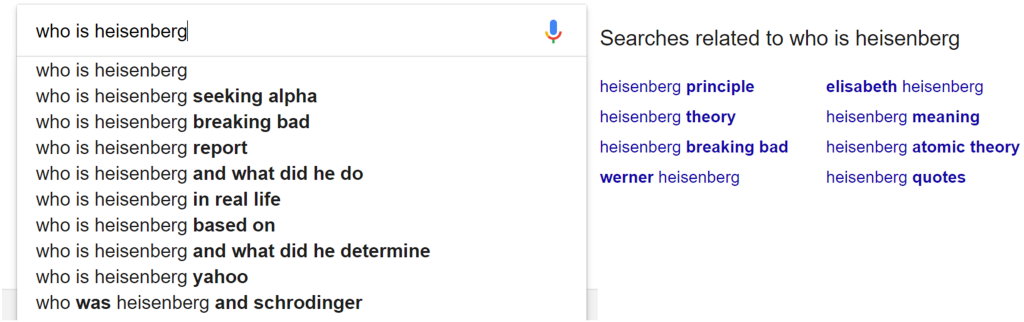

Figure 1.1: Left: when user starts entering a query, different suggestions are offered for completing the query. Right: suggestions related to the submitted query.

However, in the nowadays popular voiced-based search systems, where there is no visual channel of communication, resolving user's intent has been an ongoing challenge (Allan et al. [2012], Lai and Yankelovich [2006], Luger and Sellen [2016]). How should the system offer possible suggestions without overloading user's short-term memory? How many options should it present to user and how long should they be? How are voice-based systems supposed to provide query reformulation suggestions, when all the interaction is performed through voice?
In this thesis, we focus on resolving the user's intent for two types of unclear queries in a voice-based search setting: queries that are *ambiguous* and queries that contain *false memories*. As an example for ambiguous queries, suppose the user asks their voice assistant "How to do a B-52?". The user could refer to the B-52 cocktail, the B-52 hairstyle or maybe the B-52 Stratofortress bomber. With which of the meanings of B-52 should the voice assistant respond? How should the formulated response be, so the user has the most satisfaction? Do users prefer to be asked back for clarification or to be given meanings or entire categories of meanings to choose from? These are the problems that this thesis seeks to remedy in the first part. Clarifying false memories is the topic of the second part of this thesis. False memory means that the user mis-remembers one or more details of the item they are searching for. For instance, imagine you are searching for a movie you have watched before; the only thing you remember is some parts of the plot and the name of one actor, which in fact, is not correct. Should the system give the answer in respective to what the user said? Should it tell the user that there were no results but found something else? How should the system formulate the answer in order not to confuse the user? What if there are actual results with the given misremembered details? We conducted two user studies to seek answering the following research questions, which are important to incorporate query clarification functionality in voice interfaces:

- **RQ I** Does the user background affect user satisfaction when experiencing query clarification?
- **RQ II** How much do the length and the number of the clarifying options affect the user satisfaction?
- **RQ III** What is the best way of clarifying the user's intent?

This thesis consists of six chapters: Chapter 2 describes the literature review and the related work. To the best of our knowledge, this work is a first foray in the topic of voice query clarification. In chapter 3, we discuss voice-based search, its recent advances and challenges as well as how the Amazon Alexa

voice interface works. Chapter 4 is dedicated to our first study in which ambiguous queries are investigated. To this end, 14 participants each solved 13 different ambiguous information needs. We collected 708 judgments on the user experience of the tasks and analyzed them. The results show that the participants' English proficiency level affect their satisfaction of the experience significantly. Moreover, participants are more stratified when the different possible answers, in this case short ones, are listed to them. Chapter 5 explains our user study on clarifying false memories. We recruited 12 participants each solved 14 different task which were information needs containing false memories. We collected 672 judgments and analyzed them. We found that participants are more satisfied when the correction of a false memory is explained. Besides, the way of correcting false memories play an important role in increasing the satisfaction of the user. In chapter 6, we discuss the comparison of the two user studies. In the final chapter, we present the conclusion and future steps in integrating query clarification into voice interfaces.

Chapter 2

Related Work

The related work of this thesis is categorized separately into four subsections: query Clarification in text-based information retrieval systems, query clarification in conversational search, voice-based search, and re-finding and known items.

Query clarification in text-based information retrieval systems Query clarification including query disambiguation in the traditional text-based systems is an essential task of the commercial search engines and has been studied broadly . There are a lot of works focusing on how to generate query suggestions.
Firstly, query logs are the main source of information for building rich models of user searching activities aiming to improve users' web search experience such as query recommendations [Boldi et al., 2009]. Silvestri [2009] showed the query mining techniques that can be used to extract useful information and how they are applied in the search applications to enhance user experience. Fonseca et al. [2003] presented a method to generate suggestions for a web search engine using association rules extracted from query log of the search engine. Query graphs extracted by query logs are another source of generating query suggestions. Query graphs have nodes as queries and edge linking the two nodes represent some similarities between the two queries. Boldi et al. [2009] improved query recommendations based on short random walks on the query-flow graph.
Moreover, clustering of query logs is another way to obtain query suggestions. Wen et al. [2001] proposed to cluster similar queries according to their contents to recommend URLs to frequently asked queries of a search engine. Baeza-Yates et al. [2004] presented a method for suggesting related queries based on clustering of query logs.
Jones et al. [2006] introduced the notion of "query substitution" that is to

generate related queries to replace the user's original query. They proposed a model for query modification based on query similarity combined with a ranking of the proposed queries. Their Experimental results showed that their model is capable of generating highly relevant query substitutions.
Query auto-completions offer users useful queries while they are typing queries in real time. This works as Cai and de Rijke [2016] provided a survey of query auto completion works. Bar-Yossef and Kraus [2011] suggested to use context such as the user's recent queries to improve the prediction quality of the auto complete suggestions and proposed the first context-sensitive query auto completion algorithm. Bhatia et al. [2011] proposed a probabilistic method for generating query suggestions by extracting the correlated phrases of the documents in the corpus and suggest it to the user as they are typing the query.
Query expansion is another way of dealing with query inaccuracy. The main motivation of query expansion is to help the user to clarify the original query by adding meaningful terms to the original query and express more details in explaining the information need [Ooi et al., 2015]. One way to generate query expansion is by using relevance feedback. Relevance feedback is the approach to expand the queries in which the system chooses the terms and phrases that users identified as relevant to the original query [Ooi et al., 2015]. Another popular approach for query expansion is to build a language model for the query based on the probability distribution over terms [Carpineto and Romano, 2012] and choose the terms with the highest probability for the query expansion.
How to present the diverse search intents in the result is another challenge that has been studied extensively. Dou et al. [2011] suggested that the search results should be presented in a multi-dimensional way as queries are usually ambiguous at different levels and dimensions. Kato and Tanaka [2016] focused on ambiguous intents and proposed a method to optimize search result. They developed a probabilistic searcher model of users interacting with query suggestions and conducted a user study to examine the effects of query suggestions on search behaviors, and based on this user study, adjusted the parameters used in the searcher model. Their results demonstrated that their search result optimization was effective, especially for patient users and queries with a limited number of intents.
Another novel way to interact with the users in the text-based retrieval system and using their feedback to generate new search result was introduced by Kotov and Zhai [2010]. They proposed a new framework in which the retrieval system automatically generates questions that offer specific context to the query in order to guide users to the answers they are searching for. This can be also helpful for the ambiguous queries when the system offers multiple

diverse questions for different meanings of the query submitted. They used the resulting document collections as well as the query itself to generate the questions and present them to the user. However, these questions are not straight forward in identifying user's information need and are detailed version of the query with more context which is supposed to potentially be what user exactly meant. For example, as the article explained, if a user searches for "John Kennedy" the generated questions would be "Who is John Kennedy?", "When was John Kennedy born?", "What number president was John. F.Kennedy?", "Who killed President Kennedy?". Although these questions can be helpful for the users to specify their information need, still they do not provide different meanings for the the particular ambiguous part of the query. In another recent paper from the CAIR workshop 2018, Wambua et al. [2018] described an approach of clarifying ambiguous search queries by automatically asking users a series of clarifying questions. These questions narrow down the search result aiming to clarify the user's intent. Clearly, these two last works deal differently with the ambiguous queries and provide contextual suggestions while we focus on clarifying the ambiguous parts of the query and present different meanings.

Query clarification in conversational search Query clarification in conversational search needs more research (Allan et al. [2012], Lai and Yankelovich [2006], Luger and Sellen [2016]). Allan et al. [2012] discussed the convergence of question answering systems and information retrieval. They proposed research challenges in conversational answer retrieval (CAR) and called for researchers and designers to provide techniques to reach effective CAR systems. They declared that the purpose of the dialogue is to "refine the understanding of questions and improve the quality of answers" and clarifying any forms of ambiguity of the question is a way to reach this goal. This was one of the primary stimulants to motivate the researchers to improve person-machine conversations.
One study regarding query clarification in dialogues is conducted by Braslavski et al. [2017] who explored the dialogues between the users on a community question answering website aiming to understand the clarification question posted by users when the query intents are unclear to ultimately helping to generate automatic clarification questions that make the interaction between user and system more natural. Their experiment showed promising research direction toward the query clarification automation in the future. Radlinski and Craswell [2017] provided a framework of conversational search in which they defined a conversational search system as follows:

> A **Conversational search system** is a system for retrieving information that permits a mixed-initiative back and forth between

> a user and agent, where the agent's actions are chosen in response to a model of current user needs within the current conversation, using both short- and long-term knowledge of the user.

They proposed a theoretical framework for conversational information retrieval system in which, similar to traditional information retrieval system, users are allowed to query and the system should present the result as well as asking the user for clarification if needed. They presented the scenarios in which conversion is necessary to elicit the information need of the user. One of these scenarios that we use in our work is to give users bounded choices as it is easier for the user to choose between the options instead of coming up with the terms themselves.
One of the main characteristics of a conversational information retrieval system is to be able to elicit the intent if necessary. Luger and Sellen [2016] investigated the factors that affect the use of conversational agents in everyday life by conducting a series of interviews of 14 people who were regular users of conversational agents. Their findings show that the users did not have a proper image of how conversational agents work and the agents' poor feedback reinforced it. Consequently, they found that there is still a lot of interaction work needed to make conversational agents as natural as they are supposed to be. This is not the only research observing the interaction between users and conversational agents; Vtyurina et al. [2017] investigated human behavior when using conversational systems for search tasks. A user study was conducted in which participants used three conversational agents: automatic, human and wizard. Based on their observations some recommendations for a future conversational agent design was given: sticking to the context of the conversation helps establishing short length answers and improves user experience, providing sources of answer adds credit to the systems, and using feedback from users' answer can be used to recover from system failure.

Voice-based search Nowadays, with the evolving technology in accurate speech recognition as well as popularity of smart phones, voice-based search becomes one of the inevitable part of everyday lives. However, the voice-based search should be treated differently from the text-based search as it is more complex Trippas et al. [2018]. There are a number of studies analyzing the conversational character of voice-based search. For example Trippas et al. [2018] investigated conversational search challenges and opportunities based on the study they conducted in which pairs of people completed search tasks verbally. Their study demonstrated that the complexity of interaction increases instantly when the result is no longer displayed to the users and is presented through voice and also when the users are allowed to convey their information

need freely and naturally. On the other hand, one positive aspect of searching through audio is that the interaction between the users and the system increases, which leads to information need clarification of the user.
In Trippas et al. [2017], a controlled laboratory study was conducted to investigate the conversational strategies in search and observations were recorded, transcribed, and annotated. The recordings were analyzed and coded into conversational patterns and classified into themes. Query refinement offer, which is related to our work, is among meta-communication theme where retrievers engaged in communication with users about the query.

Re-finding and Known Items Blanc-Brude and Scapin [2007] investigated the attributes that are needed to be improved in the retrieval systems, so people can retrieve their files more easily. They conducted a study on 14 participants who had to recall the features of their documents and then to retrieve them. The results show that people have difficulty in remembering the keywords to search for the documents. Another similar memory failure is observed in the studies Elsweiler et al. [2008] Elsweiler et al. [2011a] with the emails re-finding. Tyler and Teevan [2010] shows that although Web search engines aim to help people find new information, people tend to use them to re-find Web pages they visited before and explored the differences between the first time queries to the re-finding ones, which tended to be better.
Generating simulated queries for evaluation objectives has been an increased interest. Azzopardi et al. [2007], Elsweiler et al. [2011b] and Kim and Croft [2009] worked on how to generate automatic know-item queries for personal search evaluation. Kim and Croft [2010] used human computation game to predict the type of the known-items searched by users. However, these studies failed to take into considerations the realistic scenarios with false memory Hagen et al. [2015], Hauff et al. [2012]. Hauff and Houben [2011] explored know-items which consist of books, movies, songs, etc. from Yahoo! Answers and found that about 10% contain false memories. These crawled questions and answers are the source of inspiration of our use cases in the second study we conducted.

Chapter 3

Voice-based Search

Voice-based search trends have been increasing significantly in recent years. Users tend more to use vocal commands to handle their search queries. Moreover, there are settings that make voice-based search preferable, from while driving a car, riding a bike or using wearable devices, to when you would rather ask your voice interface while you are laying on the couch than write it down in the search box. Statistical postulations say 50% of all searches will be voice searches by 2020[1]. Google has announced that 72% of people who own voice-activated speakers say that their devices are used as part of their daily routines[2]. Moreover, the popularity of virtual digital assistants is growing rapidly, with dominant technology companies such as Google (Home), Apple (Siri), Microsoft (Cortana) and Amazon (Alexa). It is estimated that the number of people using digital assistants will reach 1.8 billion by 2021[3].
Despite this popularity, digital assistants lack certain accuracy in their responses. According to a 2018 study[4], there are many cases that digital assistants proved to be wrong, for example when a query has multiple meanings, or when they answer the query with the closest topic of the information need asked by the user. The former problem is what we focused on in the first part of this thesis.
In the following sections, we explain the design principles of a voice interface, as well as introducing Amazon Alexa and how to develop a skill on its developer console.

[1]https://www.campaignlive.co.uk/article/just-say-it-future-search-voice-personal-digital-assistants/1392459

[2]https://www.thinkwithgoogle.com/consumer-insights/voice-assistance-consumer-experience/

[3]https://www.go-gulf.com/blog/virtual-digital-assistants/

[4]https://www.stonetemple.com/digital-personal-assistants-study/

Amazon Alexa Alexa, Amazon cloud based voice service, is a virtual assistant developed by Amazon which is capable of establishing voice interaction with the users. For example, it is able to search for questions, play music, set alarms or timers, show weather forecast and etc. The advantage of Amazon Alexa is that it provides a platform for developers to build custom skills. Recently, Google also provides a platform for developers to build actions for the Google Assistant. At the time of this thesis, the only open platform was provided by Amazon. In the following section we explain how we can create a custom skill on Amazon.

Figure 3.1: How Alexa skills process user's spoken query. Skill interface translates the audio from the user to events that Skill Service can handle.

Building a Skill Skills are voice driven applications for Alexa voice interface. An Alexa skill contains of two main components, *Skill Service* including HTTPS server or AWS Lambda and *Skill Interface* which is on Amazon's Alexa developer console. The interaction between these two components provides a working skill (see Figure 3.1).
The Skill Service, which we call it *Server Component* in our setup (4.1.5 and 5.1.3), lives in the Cloud and hosts the codes we write, which determines how to respond to user's speech query. It can be implemented in any language that can be hosted on HTTPS server and return JSON responses. Amazon provides a platform, AWS Lambda, in which we can code in Node.js. It is also possible to use our own HTTPS server, as we did in our work. The Skill Service implements event handler methods which define how the skill behaves when the user triggers the event by speaking to Alexa. Main event handlers implemented by the Skill Service are as the following:

- **OnLaunch Event** This event is sent to our skill once the user uses the skill's invocation name. In other words, the skill is launched by the user.

- **Intent Handler** This Event is an indication of what the user wants to do, which maps to the interaction defined in the Skill Interface. There should be as many intent handlers as the defined intents on the Skill Interface each triggered by different spoken utterances defined in the Skill Interface.

The Skill Interface configuration, which we call *Cloud Component* in our setup (see 4.1.5 and 5.1.3), is the second part of creating a skill, where we specify the utterances needed to invoke each of the intent handlers on the Skill Service. The Skill Interface translates what the user says for the Skill Service, where the respective event handles that. In the Skill interface we specify:

- **Invocation Name** which is what users say when they want to launch the skill leading to OnLaunch Event's invocation.
- **Intents** which are basically the actions users want to be done by Alexa. Intents optionally can have slots. For example, "Heisenburg" in the query "who is Heisenburg?" can be defined as a slot, so all similar queries (who is *slot*?) can be resolved by the same intent.
- **Sample Utterances** that represent all the possible ways users are talking to the skills. The more the sample utterances are, the more natural the interaction is. These sample utterances are defined for each intent separately and are ultimately resolved by an intent handler in the Skill Service.
- **Slot Types** that are lists of possible slots. For instance, for the query we presented above, "Heisenburg" would be one of the entries on the slot type lists.

Figure 3.2 shows how a simple skill named Greeter[5] works. This skill is called by saying Greeter as invocation name. When we define the sample utterances, the Skill Interface can resolve it to the specific event that the Skill Service can handle. Here the sample utterance is "say Hello". Once the event is recognized by the Skill Interface, it triggers the responding event handler in the Skill Service which then returns the output speech "Hello" to the Skill Interface and then is returned to the Alexa enabled device where it is spoken to the user.

[5]https://www.youtube.com/watch?v=QxgdPI1B7rg

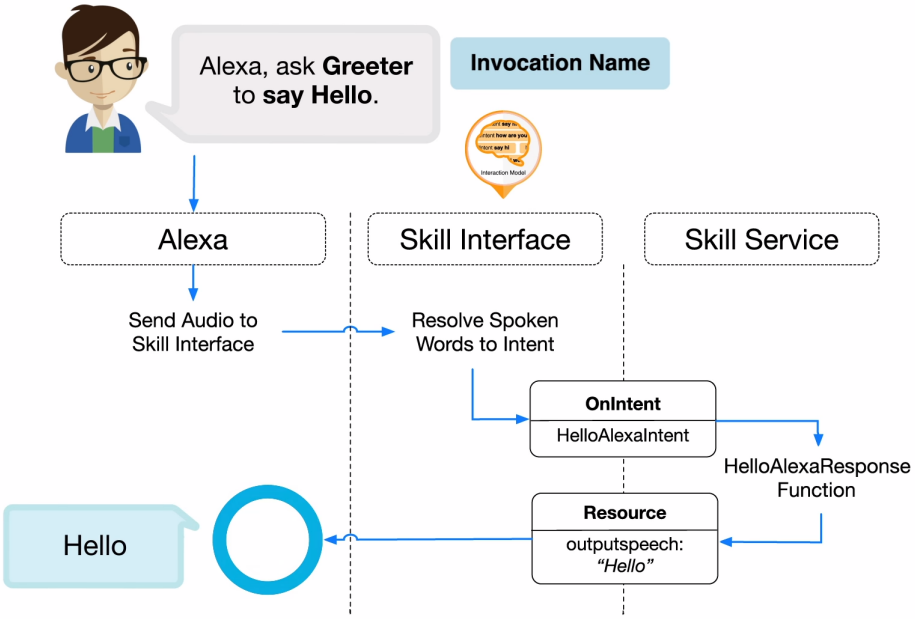

Figure 3.2: Procedure of the Greeter skill. The user says "Alexa, ask Greeter to say Hello"; Alexa is the wake word which should be used to initialize the conversation with Alexa, Greeter is the invocation name that calls the skill, and "say Hello" is an utterance resolved to an intent on the Skill Interface and is sent to the Skill Service which triggers the corresponding event handler there. The event Handler method (here: HelloAlexaResponse Function) returns the output speech of "Hello" to the Skill Interface and ultimately to the Alexa enabled device where it is spoken to the user (Source: Alexa Developers YouTube Channel)[6]

.

Chapter 4

Clarifying Ambiguous Queries

In order to analyze user behavior and preferences for a voice-based query clarification, we conducted a user study with a mock-up Alexa system. To this end, we adapted the generic research questions from the introduction as follows:

- **RQ I** Does the user background affect user satisfaction when experiencing query clarification?
- **RQ II** Do the length and the number of the clarifying options affect the user satisfaction?
- **RQ III** Does the user satisfaction decrease when the system ask for clarification?

To this end, we conducted a user study in which 14 participants had to resolve 13 ambiguous information needs using our mock-up skill for Amazon Alexa. In this chapter, we detail the procedure of the user study, the methods we used for clarifying ambiguity in the tasks, and the implementation of our mock-up skill.

4.1 Setup

In the first study we had 14 participants fulfilling 14 different ambiguous information need tasks using Amazon's Alexa voice assistant (called *system*). The study consisted of the following parts:

- One page containing consent form, the pre-study questions, and the instruction.

Figure 4.1: The setting of the user studies. The participants sat on the black chair next to the Alexa enabled device while the study instructor sat on the wooden chair handing participants different tasks as well as taking note and recording their voice.

- 14 small sheets of paper were presented to the participants randomly, containing a short scenario description, an interaction phase between the participant and the voice assistant that has the start of the interaction and an ambiguous query, and an after-interaction questionnaire containing 4 questions regarding how the interaction with the system perceived by the participants (see Figure 4.2).
- 2 optional post study questions that handed over by the instructor at the end of the study. We asked the participants how they would improve the system and let us know if they have any comments.

The study was conducted in one of the university laboratories, where Computer Science students and employees work. We used a private room and hung "study is going on" sign at the door, so people would know, and the participants would not get distracted. Figure 4.1 shows the setting of our two user studies. Participants sat on the black chair with Alexa next to them while the instructor sat on the wooden chair handing participants each tasks as well as taking notes and recording participants' voice.

In the following sections we explain the entire process and structure of the study in details.

4.1.1 Consent Form

At the beginning of the study the participants were asked to sign a consent form in which they were briefed about the duration of the study, that they stay anonymous and could quit at any time they wanted, and that their voice is recorded during the study for the research purpose. We assured them that we will not share their voice with anybody although we could not guarantee that Amazon would not make use of the recording. The consent form is available in the Appendix (see A).

4.1.2 Pre-study Questions

After signing the consent form, the participants were asked to fill out 6 pre-study questions, consist of some background questions. We explain the questions in the following parts.
The first question asked for the gender of the participants. Since we recruited participants mostly from the Computer Science and Civil Engineering department, it was likely to have more male participants than the female ones [Nelson and Rogers, 2003].
The second question asked for the age of the participants. The age ranges that we chose are the ones used in the typical demographic surveys. *17 or younger*, *18 to 30*, *31 to 49*, *50 to 64* and *65 years or older*. Our participants were among the second and third age group as they were either master's students or PhD students of the university.
The Third question asked for the use frequency of the voice assistants. The answers include *frequently*, *rarely*, and *never*. Participants that use voice assistants more frequently are expected to be more comfortable in working with the system and thus have a better user experience.
The fourth question was answered if the participants were users of the voice assistant. It asked which of the voice assistants (Amazon Alexa, Microsoft Cortana, Google Home, Siri,...) the participants use. This question also is to measure if using different voice assistants affect the overall user experience of the participants.
The next question asked for the tasks the participants use their voice assistants for. Last question was about the English proficiency of the participants, as it could be an important factor in the satisfaction of the participants experience with the system. They had to self rate their English proficiency as *Proficient*, *Intermediate*, or *Beginner*. The participants that consider themselves as the

beginner level should be omitted of the final results since it is very important to be able to read the tasks fluently and understand the options that Alexa provides and choose among them.

4.1.3 Task Instructions

After the pre-study questions and on the same page, the instruction of how to do the study was handed over to the participants. It explained that the study consists of a series of voice query tasks and each task contains a small scenario and the respective ambiguous query. The instruction also asked the participants to imagine the scenario and the information need even if they knew the answer. It described that to resolve this information need, they should communicate with Alexa and to invoke Alexa they should First say "Alexa" and wait for the blue ring to appear, and then say "Find"! The participants were instructed that this blue ring demonstrates that Alexa is listening and that Alexa then answer with "Yes?". Afterward, they should continue with the query which was written at the second bullet point on the paper. Alexa then responds with different options. The participants were supposed to continue the interaction until the information need was resolved. In case Alexa responds with "Sorry, please try again", they had to start over the interaction. After they have reached the answer or have tried 5 times, they had to answer 4 post-interaction ratings (see Figure 4.2) and then continued with the next task the instructor provided and at the end answered the post study questions.

4.1.4 Tasks

The first study was contained of 1 training task, which we did not consider in the final results and 13 others tasks. The tasks are the central part of the study, where each task consists of a small scenario description. The topics are mostly inspired by the Webis-Ambient-15[1] corpus which is an extension of the Ambient data set created by Carpineto and Romano[2]. The list of the tasks and their corresponding query is in Appendix A. We specified the query and asked the participants to read the exact query which is provided to make sure the query is ambiguous.

[1] Webis-Ambient-15

[2] http://search.fub.it/ambient/

Scenario: You want to surprise your Irish partner with an Irish cocktail called B-52, but you don't know how to make it.
Interaction start:

- Alexa. Find!
- How to do a B-52?

After interaction:	Agree		Neutral		Disagree	Don't know
The system answered my question	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Figure 4.2: One of the 14 tasks the study participants were told to do. They were instructed to start the interaction by saying "Alexa, Find!", wait for the system to react, and then to follow up with the provided question. They should then continue the interaction until the system responds with an answer. After that, they should rate their system experience for the task using the provided checkboxes.

Clarification Methods

In pursuance of investigating how to best present different clarification options, we programmed our Amazon Alexa voice assistant to respond to the participant's query for a task in different ways. Specifically, we used 11 tasks in which the ambiguity results from one word. According to Kiesel et al. [2018], the paper we submitted to the SIGIR conference, the following 7 response methods (here grouped into baseline, standard, and many-option methods) are used for clarification:

Baselines (no clarification)

Direct (2 tasks) This method answer the query directly without giving any options. The direct answer could be either the desired one(1 task, called hit) or incorrect(1 task, called miss).

Concatenate (1 task) Answer the query with three possible meanings (including the desired one) with one short sentence each.

Standard (clarification for few options)

3-meanings (2 tasks) Ask to choose from 3 meanings of the ambiguous word (described by 1-5 words). The desired meaning is in the list

(1 task, *hit*) or not, in which case participants can describe the meaning themselves or ask for more (1 task, *miss*).

3-long-meanings (2 tasks) Like 3-meanings, but meanings are described by 8-16 words (speaking takes about twice as long).

Verify (2 tasks) Ask to verify if a specific meaning is the desired one. Either "yes" (1 task, *hit*) or "no," in which case continue with 3 meanings to choose (including the desired, 1 task, *miss*).

Many-options (clarification for queries with many meanings)

5-meanings (1 task) Like 3-meanings, but with 5 meanings presented at a time.

3-categories (1 task) Like 3-meanings, but first ask for a category, then continue with 3-meanings within that category. This is inspired by Wikipedia disambiguation pages, where meanings are often grouped by category.

In order to avoid biases in the study, not only were the tasks presented to each participant in a random order, but also was the desired answer position in the options-lists randomized. Furthermore, we also used Fisher's exact test to measure significant bias which there was not any.
For comparison, we included 2 tasks where the ambiguity stems from an acronym and use 3-meanings response method in this case.
In the next part we explain how we implemented the above mentioned tasks using Amazon Alexa. Note that we used Amazon Alexa in order to provide a natural interaction for participants, as well as the fact that Amazon has an open source platform and researchers can develop their own customized skill.

4.1.5 Custom Alexa Skill

In this section, we explain how we developed our Amazon Alexa custom skill. In order to do so, the first step is to build a server that handles all Alexa requests. Webis group of our university has already implemented and provided the classes needed for handling requests. After creating the server, Alexa should be able to access it via HTTPS, which is supported by Webis 16 machine at the university. After getting the server up and running, we can start building the cloud components.

Cloud Component

After building the server and have it running, we can start adding our skill on Amazon Alexa Skills Kit Developer Console[3]. Once the skill is named, developing the interaction model can be started. Interaction model contains the invocation name, intents and slot types.

Invocation Skill invocation name is the keyword users should say to invoke the skill. We used "find" as the invocation name which is short and easy to pronounce.

Intents & Slot Types Our skill consist of 2 custom intents and 5 built-in intents containing Amazon *CancellIntent*, *HelpIntent*, *NoIntent*, *StopIntent* and *YesIntent*. As their names imply, *CancellIntent* and *StopIntent* are used for the cancellation of the interaction and *HelpIntent* handles the cases users need help about the skill. *NoIntent* provides option of a negative confirmation to a yes/no question by Alexa, which we needed when the desired answer was not on the first offered options. The purpose of the *YesIntent* is to let user provide a positive response to a yes/no question.
Our two custom intents are named *query* and *select*. *Query* intent includes *sample utterances* of the queries with a *task* slot for different queries. For example, if "How to do a B-52" is our query, in the intent *query*, it appears as "how to do a {task}", and in the list of the tasks we define a slot value B-52. As soon as the user reads the query it invokes the *query* intent (see Figure 4.3). Our second intent is *select*, abstracted name for the confirmation option users answer with the *selection* slot for the actual option. For instance if Alexa answer the query "How to do a B-52?" by "Do you mean B-52 the Irish cocktail, B-52 hairstyle, or B-52 the chess opening?", *select* is the desired option which is here "the cocktail". The *select* intent is invoked as soon as the user says "the cocktail", "b-52 cocktail", "the Irish cocktail" and many more variations that could be defined in the list of selections. These variations are listed in the Slot Types part of the interaction model(see Figure 4.4). As it is shown in the figure, it is possible to specify different synonyms for the slots in the Skill Interface, to make the system more flexible with the users' spoken words. despite the possibility, this feature was not active yet and we had to specify these synonyms on the server side. In order to avoid voice recognition errors, in addition of adding different spoken variations of the ambiguous word in the interaction model on the Amazon Console, we also added those variations to our server component. In other words, this makes Alexa recognition model trained to listen specifically for the keywords and phrases in the clarification

[3]https://developer.amazon.com

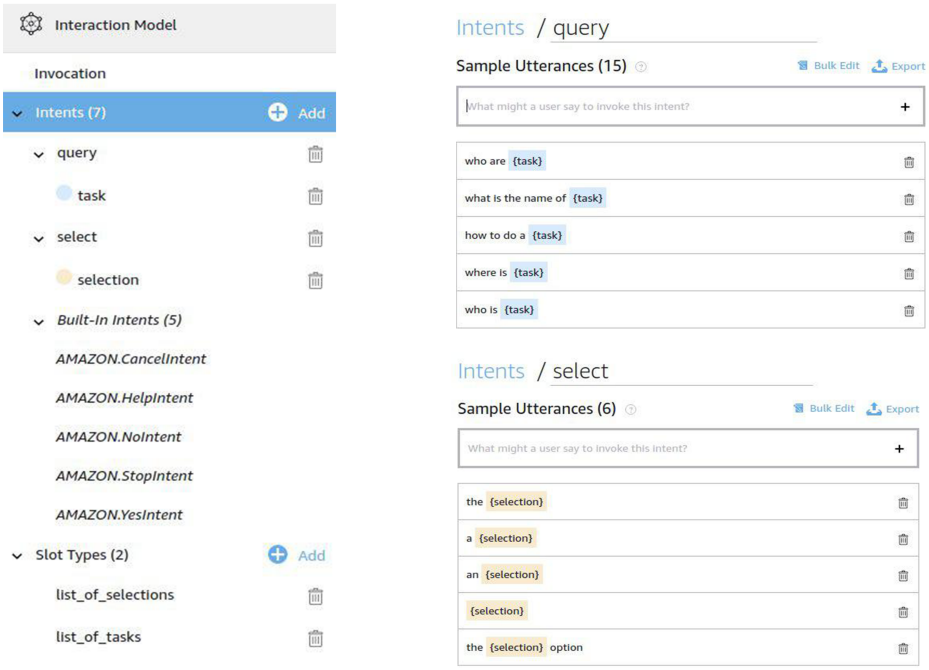

Figure 4.3: Left: Our interaction model of Alexa developer skill kit which consists of invocation, intents (custom and built-in), and slot types. Right Top: Some *sample utterances* of the *query* intent. These are basically the queries that are written for each task. Right Bottom: Some of the *sample utterances* of the *select* intent. These are the options participants were supposed to answer with during their interaction with Alexa.

options. For example for B-52 topic, we specified *b52, fifty, five two, b. 52, b 52* in our program to make sure the system understands the multiple ways of saying B-52. We also improve the system so it allows participants to use index words such as first, second, last,etc. to specify their desired option position. Since our mock-up skill is restricted to these few phrases (around 100 in total), voice recognition worked properly, with only few expectations where participants spoke too quietly.

Server Component

In this part we discuss the implementation procedure of our mock-up skill's server component. We categorize this section based on our clarification methods variations introduced in 4.1.4 and how Alexa answers in each case.
Since we had variations in providing the results to the participants, we organized presenting the tasks based on the number sequence of the participants taking part in our study. So, the participants with even number heard different

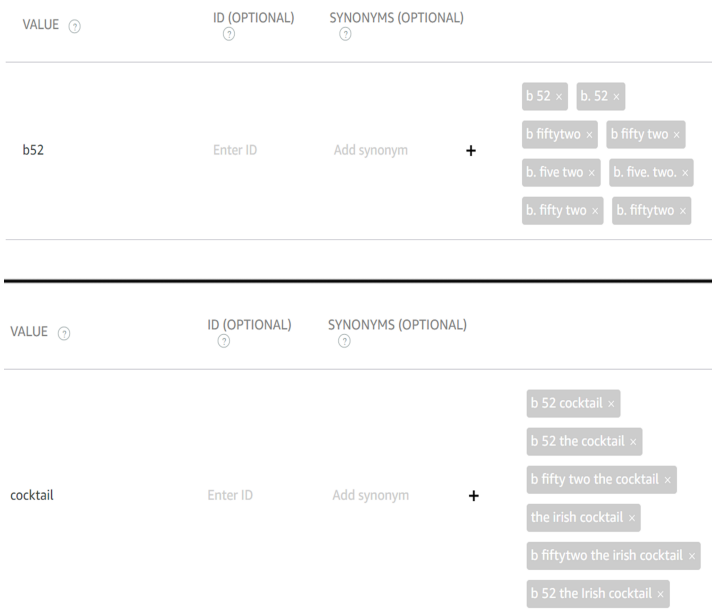

Figure 4.4: Different variations of the options that participants are supposed to say are added to avoid voice recognition mistakes and make sure Alexa understands what participants say.

answers from Alexa from the participants with odd number.

Baselines We had 2 tasks for our *Direct* clarification method that answer the query directly with one meaning of the ambiguous query, either the desired answer or not (*hit* or *miss*). Based on the participants order of participation in the study(odd and even), the system returned either of the 2 responses. As an example, the answers to the query "who are the scorpions?" are shown below. Based on the scenario, the desired answer to this query is the German rock band named Scorpions.

- Desired response: "Scorpions the German rock band consists of Rudolf Schenker, Klaus Meine, Matthias Jabs, Pawel Maciwoda and Mikkey Dee."
- Incorrect response: "The Scorpions, the British beat group consists of Peter Lewis, Thony Brierley, Anthony Postill, Rodney Postill and Mike Delaney."

The system's answer to *Concatenate* method's task, in which the system answers the query with three possible meanings including the desired one, contains short sentences of the meanings.

Standard We had 4 topics for the *3-meanings* and *3-meanings-long* clarification methods. For each participant, we used 2 of them with short and the other 2 with long options based on their participation order(odd or even). For instance, the short 3-meanings options presented to the user for the query "How to do a B-52?" are:

- the Irish cocktail
- the hairstyle
- the chess opening

3-meanings options consists of 1-5 words and it takes around 6 seconds for Alexa to tell each of them. to give the clarification options the system always started with *Do you mean* and a little pause between each options. The long 3-meanings options contains 8-16 words and it takes around 12 seconds for Alexa to tell them. The long 3-meanings options of the same query of the topic B-52 look as follows:

- the Irish cocktail which is a layered shot composed of a coffee liqueur
- the woman's hairstyle also called beehive hairstyle
- the initial moves of a chess game

As it is stated before, in order to avoid bias, the target index of the options were presented differently to users. The desired option in one 3-meanings and one 3-meanings long methods turned among 1st,2nd, and 3rd place while for the other two turned among the 4th, 5th and 6th place. For these index options, participants tended to give the right meaning before the system could give them the second 3-meanings options.
We had 2 tasks for the *Verify* clarification method n which the system asks to verify if the given option is the desired one ("Do you mean the Irish cocktail?") and the participant will say yes if it the desired one, and say no if it's not. if not, then the system continues with presenting 3 meanings which contains the desired one (in a random index). Here is a conversation example of this task in which the desired option is not on the first verified option and is on the second option of the next 3-meanings options:

User : What year is Matador produced?

System : Do you mean Matador, the Danish TV series?

User : No.

System : Do you mean 2005 film starring Pierce Brosnan, the US TV series of 2014 or the Danish TV series?

USER : US TV series.

SYSTEM : Matador is an American television series premiered on July 15, 2014, on the newly launched channel El Rey Network.

Many-options We used two topics for *5-meanings* task and *3-categories*. 5-meanings works like 3-meanings, but with 5 meanings options. For the 3-categories task, the system first asks for a category and with the confirmation of the user continue with the meanings like 3-meanings method. An example of the conversation for 3-categories is shown below.

USER : Where is Palm company?

SYSTEM : Do you mean Palm in business, in technology, or in music?

USER : business

SYSTEM : Do mean Palm Breweries, Palm the personal digital assistant, or Palm Pictures?

USER : Breweries

SYSTEM : Palm Breweries is a brewery company. It owns several different Belgian breweries. the headquarters of palm brewery is in Steenhuffeldorp, Belgium.

Acronyms 2 tasks for clarifying acronyms were added as the controller of the study which work the same as 3-meanings clarification method.

4.2 Pilot Study

Before the main study, we tested our system with three people from the Computer Science faculty to evaluate the performance of the system and improve any potential problems. The results of this primary study were not considered in our final evaluation. Overall, the system worked fine. We made some minor changes in the content of the tasks and enhanced some of the final responses to make them more comprehensible for the participants.

4.3 Participants

For the main study, 14 participants including 9 males and 5 females were recruited from our university's Computer Science and Civil Engineering Departments. As they were all Master's or PhD students, their age distribution was between 18-30 (9 participants), and 31-49 years old (5 participants).

Concerning voice assistant's use frequency, 8 participants never used one, 5 used their Siri or Google home rarely, for weather forecast and simple search, and only 1 participant used his Amazon Alexa frequently. Therefore, our participants were all adults and could be seen as novice users of voice assistants. English proficiency, which is an important factor in our study, was requested to be rated by the participants themselves. We found that in some cases, the ratings were incompatible with the participants' performance. Some participants could speak fluently, but they had rated themselves as "intermediate, or some rated themselves as "proficient" although they were not fluent enough and had difficulty understanding the meaning of some words. Therefore, we changed the ratings for 5 participants based on our observation of their interaction with the system and realized our modified rating is more compatible with the post task question of how easily they could understand the system. In this thesis we present, the modified version.

4.4 Data

The time each participant spent to complete the study was between 15 and 25 minutes. We had 14 participants for the main study fulfilling 13 tasks for a total of 182 interaction phases. We had to ignore 5 of these phases, because 3 people were not familiar with the scenario concept, 1 person said he did not pay attention to the scenario, and 1 person failed to fulfill the task due to the system bug. Moreover, we gathered 4 ratings for each interaction phase (see Figure 4.2) for a total number of 728 ratings.

4.5 Results

In this section, we investigate the research questions introduced in the introduction, based on the results from the study. Afterwards, a qualitative analysis of the study was performed.
The 4 post-task ratings collected at the end of each task focus on the user satisfaction and experience which is the main target of the research questions.

1. The system answered my question.
2. The system behaved as I expected.
3. The system was easy to hear/understand.
4. The system was pleasant to use.

In the following parts, we are investigating the research questions based on the ratings collected from the above questions.

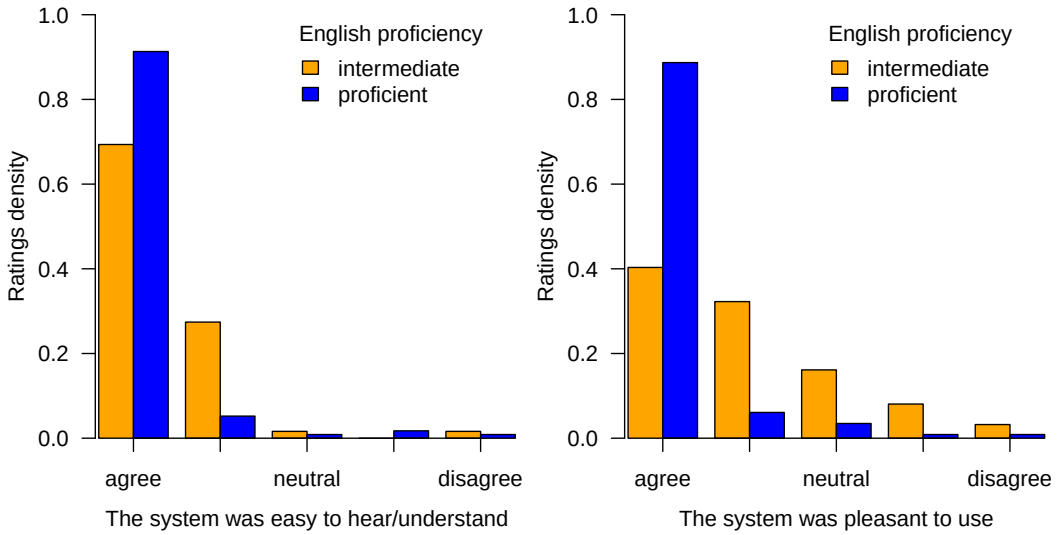

Figure 4.5: Overall ratings for understandability and pleasantness by English proficiency of the participants.

4.5.1 Does the user background affect user satisfaction when experiencing query clarification?

The background questions which have key importance for the analysis of this question are: the English proficiency and the use frequency of voice assistants. We analyzed whether the ratings of the participants with different English proficiency levels and with different use frequency of voice assistants vary remarkably.

As it is illustrated in Figure 4.5, participants with proficient English level had a more pleasant experience and the system was also easier for them to hear/understand($p < 0.001$). The Pearson correlation coefficient for the pleasantness of the system between the users with proficient English and intermediate English is -0.44 which is relatively high. This result indicates that voice assistants' designers should take users' English proficiency into considerations when designing the system, which is able to ask for clarification.

Moreover, we analyzed the ratings of the participants with different usage experience. Since we only had 1 frequent user of voice assistant, we excluded them from our statistical analysis. The effect of Voice assistants usage is shown in Figure 4.6. As the figure indicates, the participants that had experience of working with voice assistants, found the system easier to hear/understand and also more pleasant to use ($p < 0.001$ for both). The expected behavior of the system for those with more usage experience is more extreme where $p < 0.05$. This could be stemmed from the fact that more experienced users

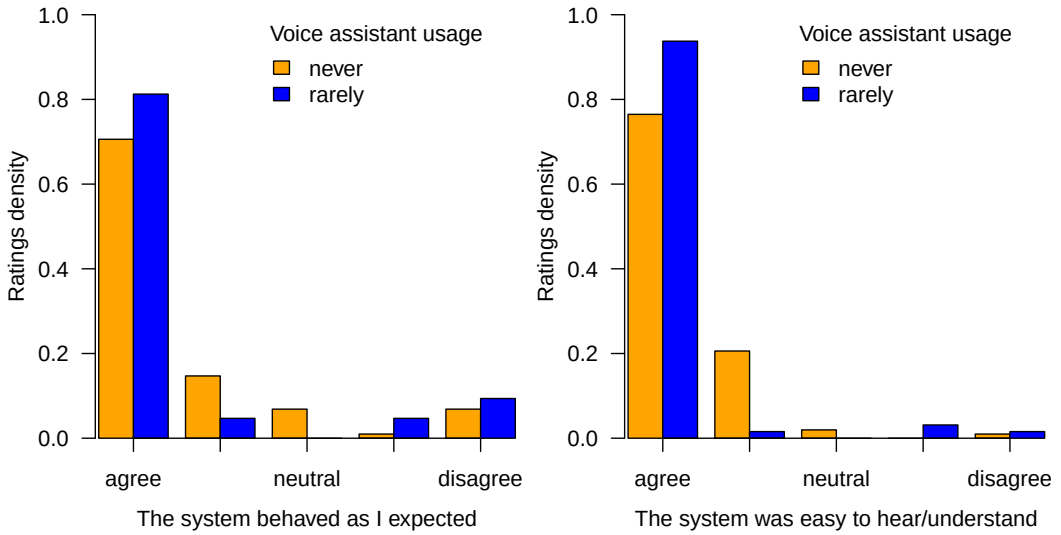

Figure 4.6: Overall ratings for predictability and understandability by frequency of a participant's voice interface usage.

had a predefined expectation from the voiced-based interfaces and this could lead them to either supporting the system or opposing to it. These results indicate that the effect of voice assistant usage is not as prominent as the effect of English proficiency ($|r| < 0.09$), and therefore, considering user familiarity seems to be not as essential in designing voice assistants.

4.5.2 Do the length and the number of the clarifying options affect the user satisfaction?

To answer this research question and as the English proficiency proved to be an important fact in 4.5.1, we focus on different English proficiencies, intermediate and proficient, separately. Since we had 5 ratings consisting of agree, somehow agree, neutral, somehow disagree and disagree, we map them onto a range from 1 to 5, where lower numbers show more satisfaction (closer to 1). We used μ_p as the satisfaction indicator in this comparison. The means of the ratings of the participants with proficient English, for different response methods were close to each other ($\mu_p < 1.4$); however, the means for the participants with intermediate English proficiency are more dispersed (see Figure 4.7). As it is illustrated in the Figure 4.7, participants were the most satisfied with the 3-meaning response method ($\mu_p = 1.4$). 3-meanings-long and verify were the next preferred methods of all. There is also a significant difference between 3-meanings and 3-categories ($p < 0.01$) and between 3-meanings-long and 3-categories ($p < 0.05$). This result indicates the importance of consid-

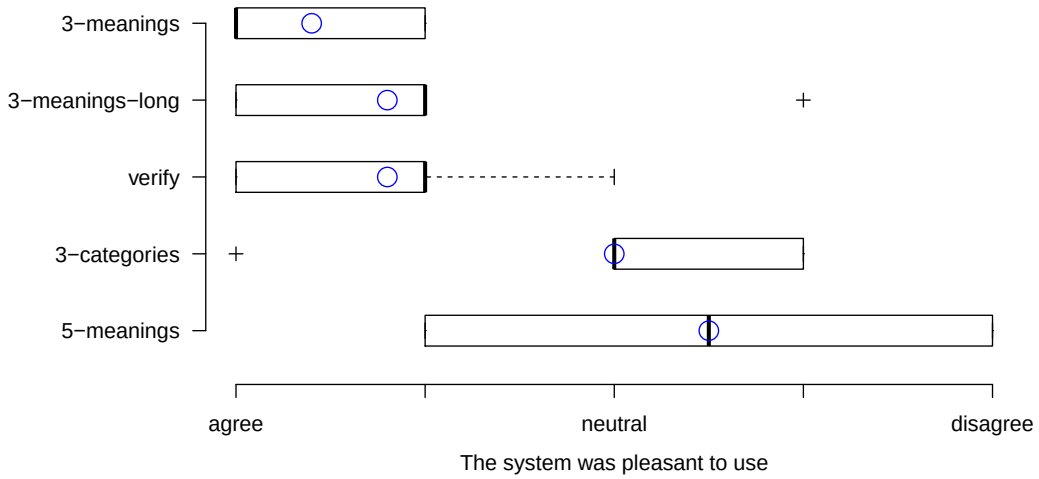

Figure 4.7: Distribution and mean (circles) rated pleasantness for participants with intermediate English level by response method.

ering 3-meanings response method in designing voice assistants for users with intermediate English proficiency, as it is the most preferred method among all we tested in our study.

4.5.3 Does user satisfaction decrease when asked for clarification?

To answer this question, we compare the verify response method, where the user has to verify the meaning before the final answer, and the direct baseline method, where the answers are given without clarification. As Figure 4.8 shows, participants rate both verify response method and direct response method similarly for the type that the system assumes correctly what the desired answer is and gives it to the user (i.e. a "hit"). This indicates that giving clarification does not have negative effect on the user satisfaction. However, when the system considers a meaning which is not the desired one regarding the scenario (i.e. a "miss"), participants' ratings for both cases drop. To our delight, five participants spontaneously stated, after the study, that they had fun interacting with the system. (add 2 3 quotes here). All in all, these results indicates that voice assistants should always seek clarification requests when there is ambiguity in the query.

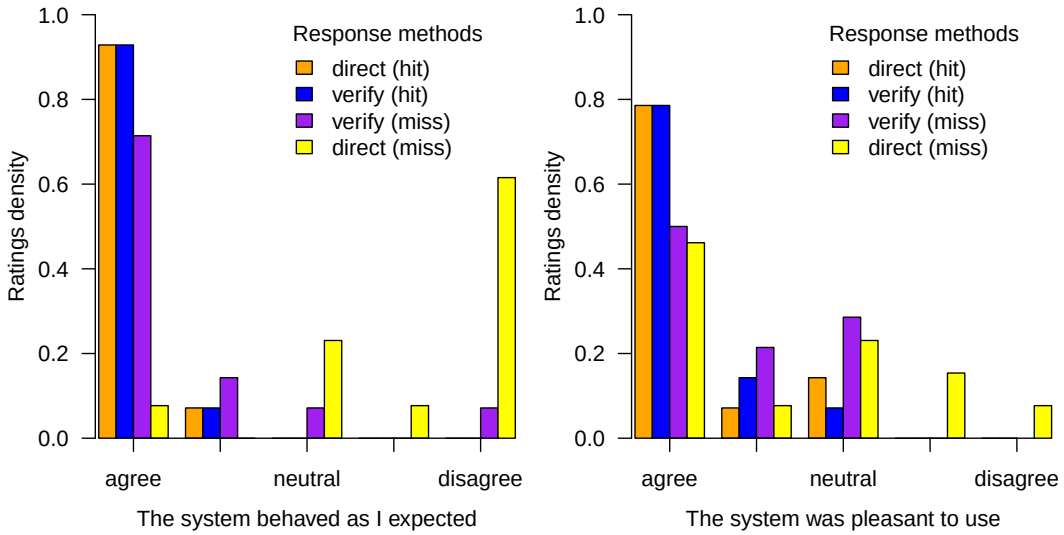

Figure 4.8: Response-specific ratings for predictability and pleasantness.

4.5.4 Qualitative Analysis

In this part, we give an overview of the qualitative analysis of our study and the design suggestions our study implies.

First of all, our observations suggest that for queries with many clarification options, users should be allowed to specify the meaning they intended. During the study, most of the participants preferred to specify the meaning themselves than to ask for more options. In detail, 10 out of 14 participants tried to interrupt the system immediately to specify their desired option before waiting for the system to finish offering the options. Surprisingly, this occurred for all the response methods except for 3-meanings and 3-meanings-long when the desired meaning is in the first list. This signifies how a list of 3 options is favored over lengthier options. The current setup of Amazon Alexa only allows interruption by saying "Alexa", which none of our participants were able to discover on their own.

Moreover, we investigated the impact of giving all the possible answers for different meanings without asking for clarification. In fact, one of the tasks was to query about a person ("Who is Heisendberg?") and the answer given by the system was one sentence containing 3 short clauses introducing the 3 famous Heisenbergs. Interestingly, participants were very pleased with this type of response method ($\mu_p = 1.2$) which indicates that on voice assistants, different presentations are suitable for different number of meanings with short answers.

In order to inquire whether participants can distinguish between queries containing ambiguous words and queries with ambiguous acronyms, we added two

extra tasks of ambiguous acronyms. Analyzing the result, we found no statistically significant different in participants' ratings. Thus, our results can be applied to both cases.

Chapter 5

Clarifying Queries with False Memories

As mentioned in the introduction (1) and reviewed throughly in the related work (2), re-finding and known-item search are common retrieval tasks. The problem emerges when users seek to search for an item which a part of it, is no longer in their memory or has been replaced. This leads to the *false memories*, i.e. the misremembered "properties" the desired item actually does not have. Queries containing such false memories might correspond either no or unsatisfactory results which make users frustrated and confused.
The ideal retrieval system should help users to resolve such confusion. This is a particular challenge for the modern voice assistants, where limited capacity of voice channel requires a vigilant design in order not to overload user's short-term memory.
In this very first research on identifying the characteristics of a voiced-based retrieval system that can detect and correct false memories, we conducted a user study on false memory clarification and seek to answer the following research questions:

- **RQ I** Does language fluency affect user satisfaction?
- **RQ II** Do wrong clarifications degrade user satisfaction?
- **RQ III** How to best clarify false memories?

Note that the above mentioned research questions are different from the ones introduced in the introduction, since we adapted them specifically for the second study. In the following sections, the details of our study on clarifying queries containing false memories will be described in voice-based search.

5.1 Setup

In the second study, we had 12 participants fulfilling 14 different information need tasks with false memories using Amazon's Alexa voice assistant (called *system*). Similar to our first study, this study consists of the following parts:

- One page containing consent form, the pre-study questions and the instruction.
- 14 small sheets of paper were presented to the participants randomly, containing a short scenario description, an interaction phase between the participant and the voice assistant that has the start of the interaction and a query containing false memory and an after-interaction questionnaire containing 4 questions, regarding how the interaction with the system was perceived by the participants (see figure 5.1).
- 2 optional post study questions that handed over by the instructor at the end of the study. We asked the participants how they would improve the system and let us know if they have any comments.

In the following sections we explain the above mentioned parts of the study in detail. For details about the consent form and the pre-study questions, see 4.1.1 and 4.1.2. We proceed with the instruction part, which is different from the first study.

5.1.1 Task Instruction

After the pre-study questions and on the same page, the instruction of how to do the study was handed over to the participants. It explained that the study consists of a series of voice query tasks and each task contains a small scenario and respective query containing false memory. The instruction also asked the participants to imagine the scenario and the information need, even if they knew the answer. It described that to resolve this information need, they should communicate with Alexa, and to invoke Alexa they should First say "Alexa" and wait for the blue ring to appear, and then say "Explore"! The participants were instructed that this blue ring demonstrates that Alexa is listening and that Alexa then answers "Yes?". Afterward, they should continue saying the query which was written at the second bullet point on the paper. Each query consisted of three facts highlighted with grey background. Alexa then either answered the query or said she does not know. The study supervisor confirmed it to the participant whether The answer is indeed the desired one with showing a thumbs-up or not, showing a thumbs-down. In

case Alexa didn't know the answer or the study supervisor showed a thumbs-down, the participants were supposed to continue querying by dropping out one of the details that are indicated by the grey background and query again with the 2 remaining details. This process should be repeated until the study supervisor showed a thumbs-up. After each task, the participants answered 4 post-interaction ratings (see Figure 5.1) and then continued with the next task the instructor provided, and at the end of the study, answered the post study questions.

5.1.2 Tasks

In this study we had 2 training tasks which are not considered in our results and 14 main tasks. Each task consists of a brief scenario description along with a corresponding query that contains a false memory. The scenario descriptions are based on real known-item queries with false memories collected by Hagen et al. [2015]. They crawled a set of 2,755 known-item intents (movies, songs, books, poems, etc.) from Yahoo! Answers and found that about 10% contain false memories. Figure 5.1 shows a task with an example scenario, query, and questionnaire.

Scenario: You try to remember the title of a controversial book that came out back in the 1990s and claimed scientific evidence that whites are genetically superior to blacks. You think it was called something like "The *something* Factor."

Interaction start: Alexa. Explore!
What is the title of the book from the 1990s that claimed superiority of Whites and is called "The *something* Factor"?

Post-interaction questions:	Agree		Neutral		Disagree	Don't know
The system was helpful	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Figure 5.1: Example task from our study. Participants should start by saying "Alexa, Explore!", wait for the system to react, and then read the provided question. The question details—one of which is a false memory—were highlighted with a grey background. After task completion, the participants rated their experience using the checkboxes.

As it is shown in the Figure, there are three details highlighted in the query from which, one is wrong. The participants were not informed, which of these highlighted details is wrong. In 12 tasks of the 14 main tasks, after the user queries the original query, the system changes a detail in it. In 6 of these 12 tasks, the system corrects the detail which is a false memory, and in the other 6 tasks changes a different one. If the false memory was not corrected, the participants had to drop one of the details and use the other two to query again. This procedure continued until the participant addressed the detail containing false memory.

Clarification Methods

To analyze how to clarify false memory corrections, our mock-up responded in one of the following ways to a participant's prompt [Kiesel et al., under review].

None (2 tasks) "Sorry, I don't know that one."

Direct (4 tasks) Answer the query for a changed detail, but do not explicitly inform the user that a change has happened. However, since it is best practice for voice-interface design to repeat the query in the response[1] (so-called *landmarking*), the system does so with all details, including the changed one. Participants, therefore, may notice that the system has changed a detail.

Negativley Clarified (4 tasks) Respond with the standard Alexa line for no result, clarify that a result exists when a detail is changed, and then answer the changed query.

Positively Clarified (4 tasks) Instead of starting with "I don't know," the more positive suggestion "You probably mean ..." is used.

If the system changed a detail of the query, for half of the tasks the false memory was corrected, while for the other half the correction itself was wrong. Figure 5.2 illustrates the response types.

5.1.3 Custom Alexa Skill

Similar to the skill we built for the first study, after configuring the server that can handle Alexa requests, we start creating the cloud components on the Skill Interface and then implementing the server components on the server side which are explained in the following sections.

[1] `developer.amazon.com/docs/custom-skills/voice-design-best-practices-legacy.html`

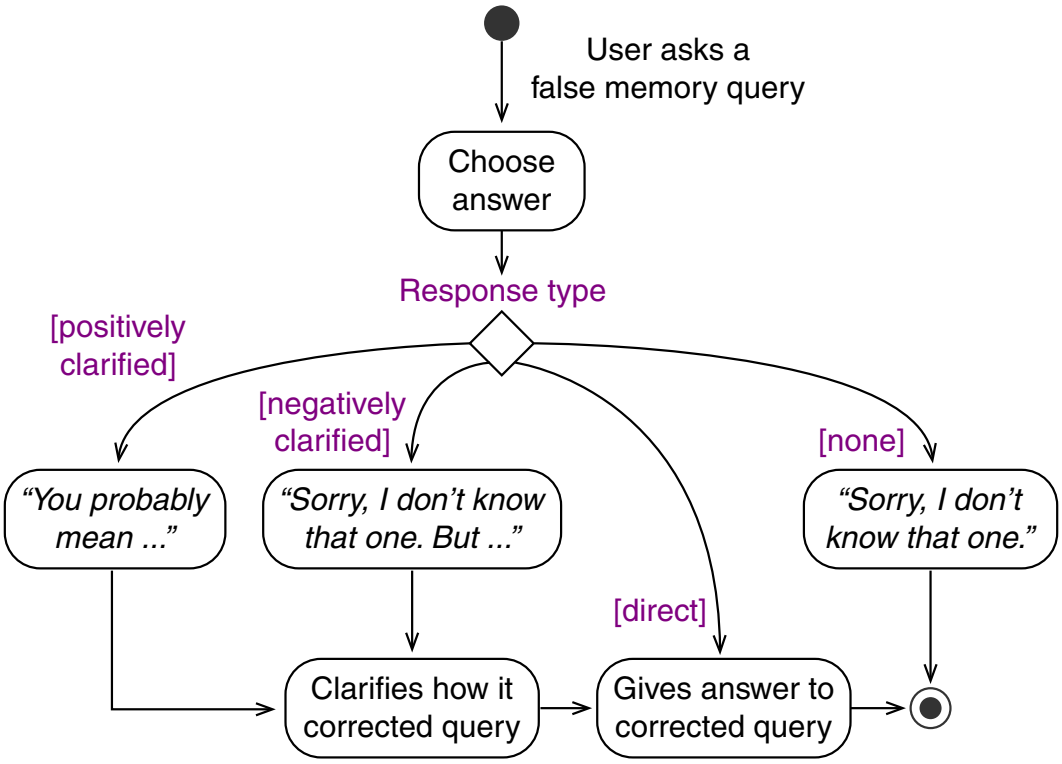

Figure 5.2: The four response types of our study with the corresponding Alexa reaction (the Figure is from Kiesel et al. [under review])

Cloud Component

As it is explained in 4.1.5 and 4.1.5, after building the server and have it running, we can start adding our skill on Amazon Alexa Skills Kit Developer Console[2]. Once the skill is named, developing the interaction model can be started. Interaction model contains the invocation name, intents and slot types.

Invocation Skill invocation name is the keyword users should say to invoke the skill. We used "explore" as the invocation name which is short and easy to pronounce.

[2]https://developer.amazon.com

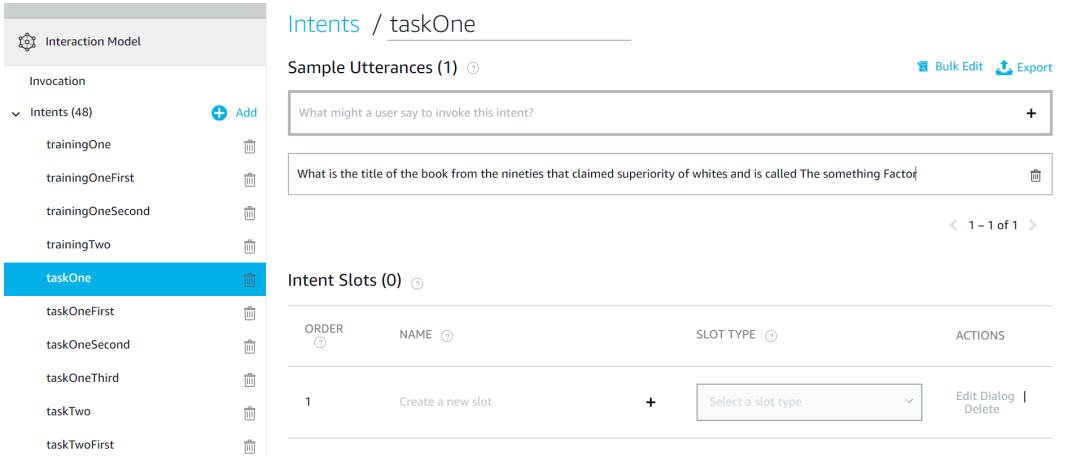

Figure 5.3: A part of our interaction model of Alexa developer skill kit which consists of invocation, intents (custom and built-in), and no slot types. On the right, there is the *sample utterances* of the first task intent. This is the original query written for the first task. The 3 successive intents (*taskOneFirst*, *taskOneSecond*, and *taskOnethird* are the 3 other variations of this query, each with two of the items (the three items here are *from the nineties*, *claimed superiority of Whites*, and *is called "The something Factor"*))

Intents & Slots Types Our skill consists of 48 custom intents and 5 built-in intents containing Amazon *CancellIntent*, *HelpIntent*, *NoIntent*, *StopIntent* and *YesIntent*. The built-in intents usage is explained in 4.1.5. Our 48 custom intents cover our 16 (2 training and 14 main) tasks of the query that are supposed to be asked by the participants. Each of these intents were invoked by different variations of combining the highlighted details in each query. Unlike our first study, this skill does not have any slots. Figure 5.3 shows the interaction model of the second study. For each task we have an intent with the *sample utterance* exactly set to the original query, written for each scenario in the task. For the 6 of our tasks, there are 3 more intents, in which two of the three details (that are highlighted in the original query) are forming a query. For instance, the query "What is the title of the book from the 1990s that claimed superiority of Whites and is called "The something Factor" ?" has one intent with exactly the sample utterance as the query, one intent with the sample utterance of "What is the title of the book from the 1990s that claimed superiority of Whites?", which is a query combining of first two details, and two other intents for the other 2 variations.

Server Component

After configuring the skill on the Amazon Alexa Console, we should implement what Alexa responds to the user. Unlike the first study that the whole procedure was done by 2 intents and we dealt with slots more than intents, here we had to assign different responses for our 48 custom intents. Since the response types are throughly demonstrated in 5.2, below we show a demo conversation for the task shown in 5.1. The clarification method in this task is *None*, so the system answers the written query with "Sorry, I don't know that one". But as we wanted the users reach the answer, we asked them to query again to reach the answer. Here the user's attempt in reaching the answer was successful in the second try of formulating the query.

USER : What is the title of the book from the 1990s that claimed superiority of Whites and is called "The something Factor"?

SYSTEM : Sorry, I don't know that one.

INSTRUCTOR : (thumbs down)

USER : What is the title of the book that claimed superiority of whites and is called The something Factor

SYSTEM : The book that claimed superiority of Whites is called "The Belief Factor".

INSTRUCTOR : (thumbs down)

USER : What is the title of the book from the nineties that claimed superiority of whites

SYSTEM : The book from the 90s that claimed superiority of Whites is called "The Bell Curve".

INSTRUCTOR : (thumbs up)

5.2 Pilot Study

We conducted a pilot study before the actual study to test our system. We had 2 participants from the Computer Science department of our university taking part in the pilot study. The overall system performance was fine. We improved some of the answers by Alexa in order to make them more comprehensible. We did not consider the pilot study results in our final findings.

5.3 Participants

We recruited 12 participants from our Computer Science, Civil Engineering, and Urbanism Department including 7 males and 5 females. As they were all Master's or PhD students, their age distribution was between 18-30 (6 participants), and 31-49 (6 participants).
Concerning voice assistant's use frequency, 2 participants never had used one, 9 had used their Siri or Google home rarely, mostly for weather forecast and search, and only 1 participant used his Amazon Alexa frequently. Therefore, our participants are all adults and can be seen as novice users of voice assistants.
Finally, 7 participants had proficient English, whereas the other 5 had intermediate English.

5.4 Data

The time that the 12 participants completed the study was between 20 and 26 minutes. Each participant fulfilled 14 tasks which gives 168 total interaction phases. After each interaction phase, a 4 post-interaction ratings were collected, which gives a total rating of 672.

5.5 Results

In this section, we investigate the research questions introduced in the introduction. We adapted the research questions to match this second study. The 4 post-interaction questionnaire were the base, on which we sought to answer the research questions. These 4 questions are as follows:

- The system was helpful.
- The system behaved as I expected.
- The system was easy to hear/understand.
- The system was pleasant to use.

The first question is different from our first study (the system answered my question), as each task continues until the participant reaches the answer and this is implied by the thumbs-up given by the study supervisor.
In the following parts, we are investigating the research questions based on the ratings collected from the above questions.

5.5.1 Does language fluency affect satisfaction?

Figure 5.4 illustrates the ratings of the participants for the 4 post-interaction questions based on their English proficiency. The expectation is that participants with less fluent English rate the system less pleasant. However, there is no correlation between the fluency and pleasantness in the ratings (Pearson's $r = 0.05$), although there is a significant difference ($p < 0.01$). This suggests that for the response types we tested in our study, English fluency does not play an important role. There was also no significant difference between participants with different English fluency in hearing/understanding the system as the $p < 0.01, r = 0.28$. Finally, as the Figure indicates, there is no significant difference in the ratings of the participants with different English levels in terms of their expectation of the system and predictability, as $p > 0.05$ for both cases. As the result suggests, designers of voice assistants do not necessarily need to consider English fluency of the users to develop a system capable of clarifying false memory with our system's response types.

5.5.2 Do wrong corrections degrade satisfaction?

Since satisfaction is the ultimate aim of an information retrieval system and correcting query may affect user satisfaction drastically, we tried to investigate whether wrong corrections degrade satisfaction. Specifically, when correcting a false memory, the system might return some results that are not user's desired information need. For such cases and for cases when the system changes the information need without giving any feedback, we expected that perceived pleasantness by the participants reduces. To our surprise, and as Figure 5.5 shows, for some response types (*Direct* and *Positively clarified* in the Figure), the participants were more pleased when the system corrected their query and gave them results aside from their desired information need (*other detail* in the Figure). Moreover, the system was pleasant to use even when there was no correction (*None* in the Figure). However, the result for *p*ositively clarified is the most significant, despite the small sample size for both the general case and when restricted to tasks where the false memory detail is changed ($p < 0.05$ for both). As the result indicates, voice assistants should rather try to directly correct their users' queries, even if there is a chance of presenting undesired results according to users' information need.

5.5.3 How to best clarify correction?

While several ways of explanation is plausible to best clarify questions, here we investigate two of the most basic parameters of explanations: should corrections be clarified at all and does the tone of the correction matter?

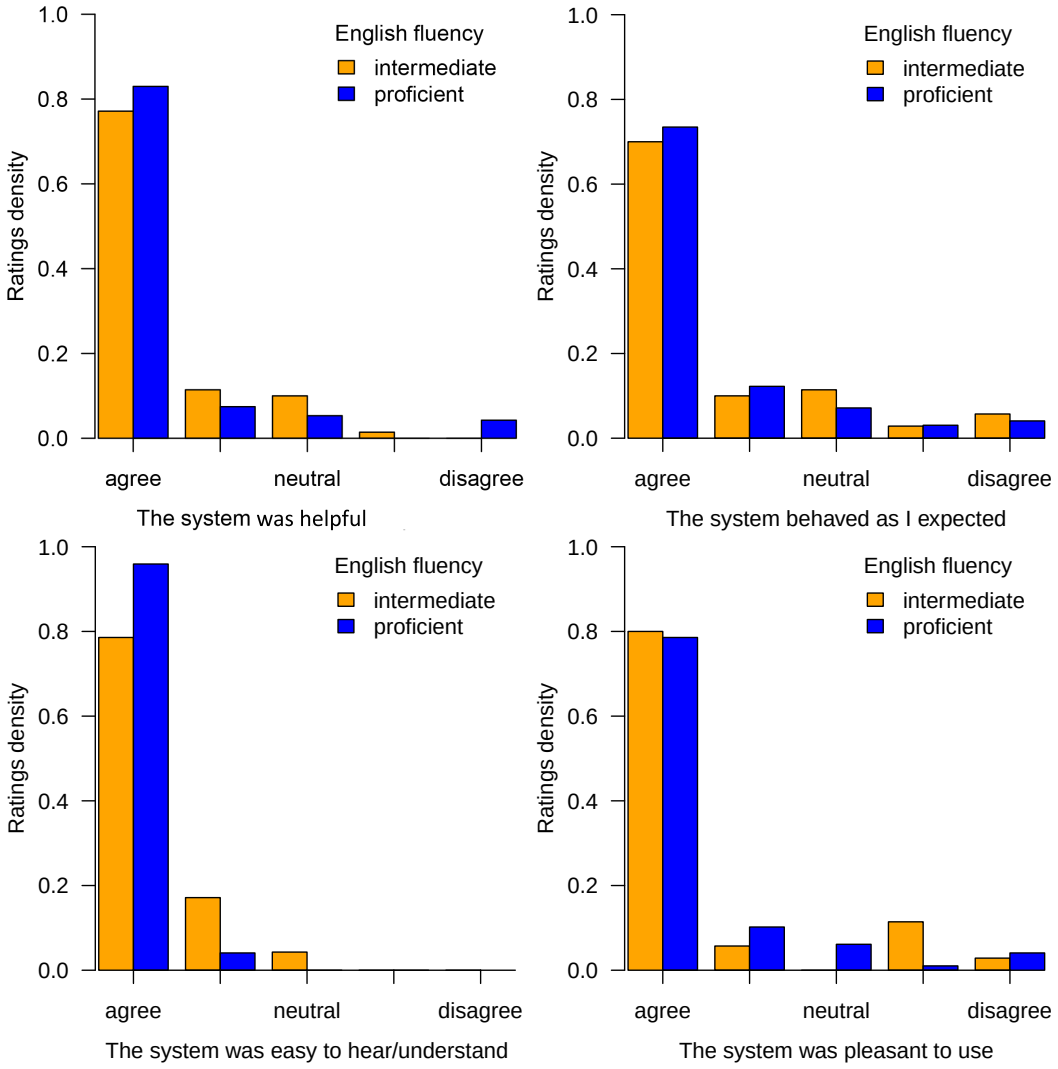

Figure 5.4: Overall ratings by English fluency.

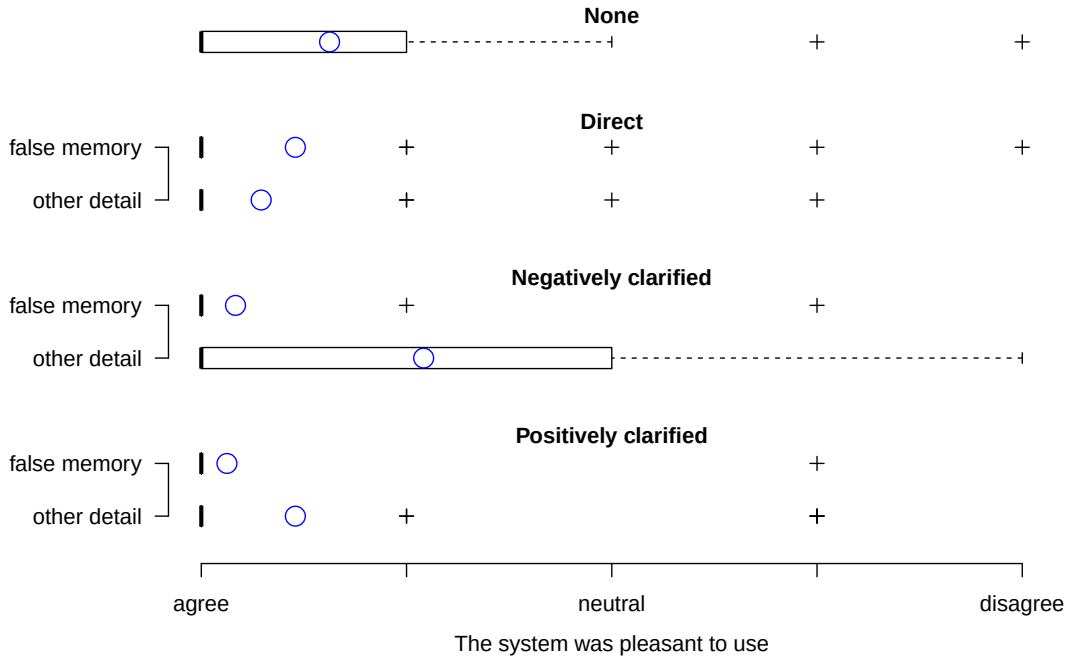

Figure 5.5: Distribution and mean (○) pleasantness by response method and changed detail.

For the case that the response does not refer to the false memory (i.e *other detail*), Figure 5.5 shows a noticeable difference between the two response types that clarify the correction, positively and negatively, where the positive one was rated as more pleasant. This finding indicates that although in both cases the responses did not satisfy participants' desired information need, they still prefer to be corrected in a positive way. However, due to our small sample size, this visually noticeable difference in Figure 5.5 is not statistically significant as $p > 0.05$ and should be validated in future studies. Despite that, the results suggest that in order to increase user satisfaction, the tone of the correction should be considered as important.
For the case that the response does refer to the false memory, Figure 5.6 illustrates the participants ratings for the predictability and pleasantness of the system in details. As it is clear in the Figure, the response types that clarify the correction, both positively and negatively, collect better ratings for both predictability and the pleasantness of the system. However, due to our small sample size there no statistically significant result as $p > 0.05$. nevertheless, two of our participants spontaneously stated that the explanations given by the system were useful and they would with for an Alexa with such functionalities in real world. All in all, the visual distribution of the ratings and the feedback we received from the participants, indicate that explanations of the query

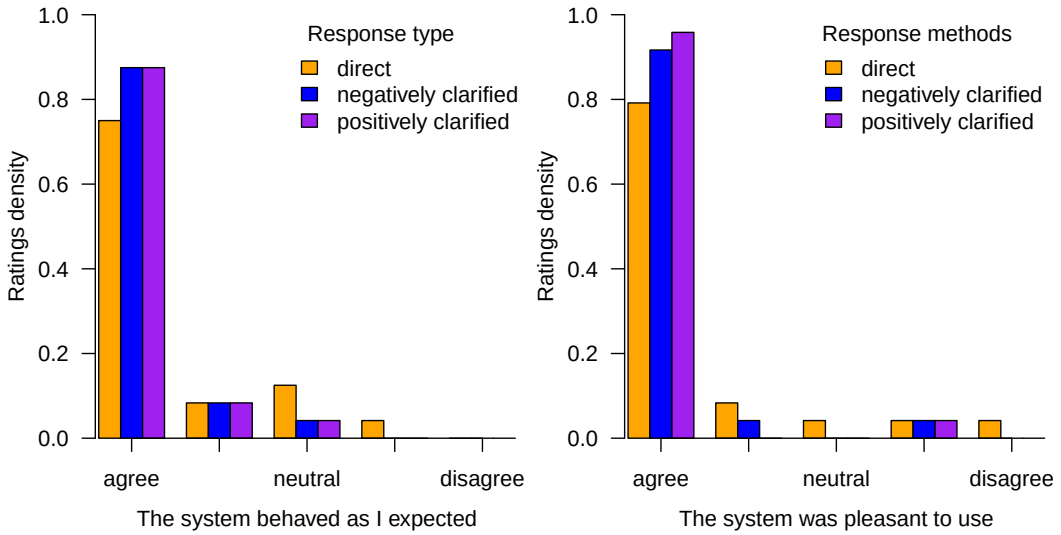

Figure 5.6: Response-specific ratings for predictability and pleasantness when the false memory detail is changed.

corrections are not only to some degree expected by the users, but also increase the users satisfaction.

5.5.4 Qualitative Analysis

In this part we explain the qualitative analysis of our observations during the second study.
Firstly, our observations suggest that a successful voice assistant should allow users to query at any length they need to with acceptable pause time. Specifically, participants found it not pleasant when the system interrupted them while formulating their query. "Alexa, be more patient!", one of our participants stated along with 4 more that suggested to extend the listening time as improvement recommendation at the end of the study. This is the current issue of Amazon Alexa, where the listening time does not suffice for long queries and with a little pause from the user side, the session is ended forcing the user to start over.
Moreover, as our statistical analysis also approved, participants preferred when the system positively corrected the false memory. One of our participants said "I liked the mode in which I got profound feedback with additional information about question asked". Therefore, this finding suggests that the designers of voice assistants should take into account the level of explanation when correcting queries that contain false memory.
Finally, we used landmarking in the responses Alexa provided which is ac-

ceptable in today's voice assistants. However, 2 of our participants found this not pleasant and stated that in some answers there was "too much repetition of the question". This indicates that for long queries voice assistants should avoid landmarking or maybe just give a brief overview of the query.

Chapter 6

Comparison of the Two Studies

In this thesis, we target resolving user's intent for two types of unclear queries in a voice-based search setting: queries that are *ambiguous* and queries that contain *false memories*. To this end, we conducted two user studies to measure users satisfaction with the methods our system provided in each user study. In this chapter, we specify the comparison between these two studies and provide design implications for voice assistants in clarifying user's intent.
One of the core differences between the two studies is the level of interaction between the system and the user. In clarifying ambiguous queries in the first study, we used 7 response methods from which, 5 involved users in a back and forth, natural conversation and the interaction continued until the user confirmed the clarification option that the system provided. However, in clarifying queries containing false memories in the second study, the end of the task was marked by the instructor providing thumbs up or down and the interaction between the system and user was one level, with the user asking the query and the system responding to that. Having more interaction in the first study, the following situations arose:

- When completing the tasks in the first study, participants tended to interrupt the system by prompting the answer themselves. Specifically, 10 out of 14 participants gave the desired option to the system before it finishes offering the options. There were also 2 cases in which the participants did not how to proceed after hearing that the desired option was not among the first list the system provided. However, these were not the case in the second study where the system answered the query right away and the instructor marked the end of the interaction. This also happens for the tasks in which the participants had to assign a category to the ambiguous word of the query. They either said "none of them" in response to the system or tried to give the answer directly from the scenario. These findings indicate that it is important for a voice

assistant attempting to clarify user's intent, to allow users to interrupt the system when they need.

- When interacting with the system in the first study, participants needed to be more fluent in English in order to have a better user experience. As it is shown in the result of the first study in chapter 4, English proficiency played a significant role. Participants with higher proficiency level found the system both more pleasant to use and easier to understand. Some participants had a problem with the meaning or pronouncing certain words; therefore, they had to do the task more than once to reach the final answer and that affected their satisfaction of the whole experience. However, in the second study, there is no correlation between the fluency and pleasantness in the ratings of the participants. These results suggest that designers of voice interfaces should take into account user's English proficiency based on the level of interaction user and the system are supposed to have.

Another considerable difference between the two studies is that in the second study we took into account the tone of the responses given by the system which we did not in the first study. The reason is that in the first study, there were several possibilities to answer the submitted query and user had no unintentional effect, while in the second study, there were queries that contain false memory which stems from the fact that users misremembers a fact. As a result, the system in the second study had to choose whether to correct the user or not. For this reason, we consider the tone of the given response when designing the second study, which proved to be an important factor in clarifying queries with false memory as it is shown in the results (see 5.5). Consequently, it is important to consider the tone of a voice assistant when clarifying false memory.
One similarity between the studies is that in both studies we used landmarking method to provide the final answer to the users. Landmarking is when the system repeats back what the user says in the query to give an implicit confirmation to assure users they are correctly heard. Our observations confirm the landmarking rule which is suggested by Amazon[1], if the repeated phrases are not too long, as our qualitative result indicates (see 5.5.4).

[1] `developer.amazon.com/docs/custom-skills/voice-design-best-practices-legacy.html`

Chapter 7

Conclusion

In this thesis we took a first step towards the task of voice search query clarification.
In the first part of this work, we conducted a user-centric study to answer three key research questions for clarifying voice search queries which are ambiguous. The question of how to present the different clarification options to the user, how much does user satisfaction decrease when the system asks for clarification and if users' background impact their satisfaction of the system.In order to achieve this aim, 14 participants completed 13 different ambiguous information need. We collected 708 judgments on the user experience of the tasks and analyzed them. Our findings show that users won't mind when the system asks for clarification and this doesn't affect their overall satisfaction of the system. Moreover, English proficiency plays an important role in increasing the level of satisfaction of the users and it is essential to be considered in the designing of clarification options on the voice interfaces. Our study also shows that presenting three clarification options is recommended over other number of options. Furthermore, users should be allowed to interrupt the system whenever they want to confirm and clarify the query themselves. Last but not least, users preferred that the system provides all the possible answers in case the answers are short.
In the second part of this thesis, we investigated how voice-based search systems should correct queries containing false memories. We identified three key research questions as how to best clarify queries with false memory, as well as inquiring if the English language level of the users affect their satisfaction of the system, and if there is a penalty for the user satisfaction when the system corrects the query containing false memory. To this end, we conducted a user study attempting to answer our research questions with 12 participants solving 14 different tasks, which were information needs containing false memories. We collected 672 judgments and analyzed them. Our results indicate

that clarifications of false memory increase satisfaction and users prefer failed clarified corrections over no correction. In addition, we found that the way system corrects the user has an impact on the user satisfaction; specifically, we found that the more positive the tone of the system's response, the higher the user satisfaction on average.

As future work, many different designs of the studies that we conducted can be performed. Specifically, alternative response methods in different scenarios can be tested in order to find the most suitable strategy in clarifying the ambiguous queries as well as those that contain false memory. In achieving better response methods for correcting queries with false memory, our findings suggest that the focus should be more on the *positively clarified* response type together with experimenting alternative ways of conveying the positive tone in the answer, as such tone of the system affects user satisfaction significantly. This can include pointing out to the user positively in which part of the query they had false memory. Besides, crowdsourcing platforms can be helpful in verifying and improving the confidence of our results. Finally, these varying response methods can be implemented and deployed as different skills on the open platforms such as Amazon Alexa or Google Home. As a result, users can test them in real scenarios and researchers can use the data to evaluate the methods.

Bibliography

James Allan, W. Bruce Croft, Alistair Moffat, and Mark Sanderson. Frontiers, challenges, and opportunities for information retrieval: Report from SWIRL 2012. *SIGIR Forum*, 46(1):2–32, 2012. doi: 10.1145/2215676.2215678. URL `http://doi.acm.org/10.1145/2215676.2215678`. 1, 2

Leif Azzopardi, Maarten de Rijke, and Krisztian Balog. Building simulated queries for known-item topics: an analysis using six european languages. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 455–462, 2007. 2

Ricardo Baeza-Yates, Carlos Hurtado, and Marcelo Mendoza. Query recommendation using query logs in search engines. In *Proceedings of the 2004 International Conference on Current Trends in Database Technology*, EDBT'04, pages 588–596, Berlin, Heidelberg, 2004. Springer-Verlag. ISBN 3-540-23305-9, 978-3-540-23305-3. doi: 10.1007/978-3-540-30192-9_58. URL `http://dx.doi.org/10.1007/978-3-540-30192-9_58`. 2

Ziv Bar-Yossef and Naama Kraus. Context-sensitive query auto-completion. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 107–116, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0632-4. doi: 10.1145/1963405.1963424. URL `http://doi.acm.org/10.1145/1963405.1963424`. 2

Sumit Bhatia, Debapriyo Majumdar, and Prasenjit Mitra. Query suggestions in the absence of query logs. In *Proceedings of SIGIR 2011*, pages 795–804. ACM, 2011. 1, 2

Tristan Blanc-Brude and Dominique L. Scapin. What do people recall about their documents?: implications for desktop search tools. In *Proceedings of the 12th international conference on Intelligent user interfaces*, IUI '07, pages 102–111, New York, NY, USA, 2007. ACM. ISBN 1-59593-481-2. doi: 10.1145/1216295.1216319. URL `http://doi.acm.org/10.1145/1216295.1216319`. 2

Paolo Boldi, Francesco Bonchi, Carlos Castillo, Debora Donato, and Sebastiano Vigna. Query suggestions using query-flow graphs. In *Proceedings of the 2009 Workshop on Web Search Click Data*, WSCD '09, pages 56–63, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-434-8. doi: 10.1145/1507509.1507518. URL `http://doi.acm.org/10.1145/1507509.1507518`. 2

Pavel Braslavski, Denis Savenkov, Eugene Agichtein, and Alina Dubatovka. What do you mean exactly?: Analyzing clarification questions in cqa. In *Proceedings of CHIIR 2017*, pages 345–348. ACM, 2017. 2

Fei Cai and Maarten de Rijke. *A Survey of Query Auto Completion in Information Retrieval*. Now Publishers Inc., Hanover, MA, USA, 2016. ISBN 168083200X, 9781680832006. 2

Claudio Carpineto and Giovanni Romano. A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, 44(1):1:1–1:50, January 2012. ISSN 0360-0300. doi: 10.1145/2071389.2071390. URL `http://doi.acm.org/10.1145/2071389.2071390`. 2

Hang Cui, Ji-Rong Wen, Jian-yun Nie, and Wei-Ying Ma. Query expansion by mining user logs. 15:829– 839, 08 2003. 1

Zhicheng Dou, Ruihua Song, and Ji-Rong Wen. A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 581–590, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-654-7. doi: 10.1145/1242572.1242651. URL `http://doi.acm.org/10.1145/1242572.1242651`. 1

Zhicheng Dou, Sha Hu, Kun Chen, Ruihua Song, and Ji-Rong Wen. Multi-dimensional search result diversification. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 475–484, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0493-1. doi: 10.1145/1935826.1935897. URL `http://doi.acm.org/10.1145/1935826.1935897`. 2

David Elsweiler, Mark Baillie, and Ian Ruthven. Exploring memory in email refinding. *ACM Trans. Inf. Syst.*, 26(4):1–36, 2008. 2

David Elsweiler, Mark Baillie, and Ian Ruthven. What makes re-finding information difficult? a study of email re-finding. In *Proceedings of the 33rd European conference on Advances in information retrieval*, ECIR'11, pages

568–579, Berlin, Heidelberg, 2011a. Springer-Verlag. ISBN 978-3-642-20160-8. URL `http://dl.acm.org/citation.cfm?id=1996889.1996963`. 2

David Elsweiler, David E. Losada, José Carlos Toucedo, and Ronald T. Fernández. Seeding simulated queries with user-study data forpersonal search evaluation. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 25–34, 2011b. 2

Bruno M. Fonseca, Paulo B. Golgher, Edleno S. de Moura, and Nivio Ziviani. Using association rules to discover search engines related queries. In *Proceedings of the First Conference on Latin American Web Congress*, LA-WEB '03, pages 66–, Washington, DC, USA, 2003. IEEE Computer Society. ISBN 0-7695-2058-8. URL `http://dl.acm.org/citation.cfm?id=951953.952387`. 2

Matthias Hagen, Daniel Wägner, and Benno Stein. A corpus of realistic known-item topics with associated web pages in the clueweb09. In *Advances in Information Retrieval - 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 - April 2, 2015. Proceedings*, pages 513–525, 2015. 2, 5.1.2

Claudia Hauff and Geert-Jan Houben. Cognitive processes in query generation. In *Advances in Information Retrieval Theory - Third International Conference, ICTIR 2011, Bertinoro, Italy, September 12-14, 2011. Proceedings*, pages 176–187, 2011. 2

Claudia Hauff, Matthias Hagen, Anna Beyer, and Benno Stein. Towards realistic known-item topics for the ClueWeb. In *Proceedings of the 4th Information Interaction in Context Symposium, IIiX 2012*, 2012. 2

Bernard J. Jansen, Amanda Spink, and Tefko Saracevic. Real life, real users, and real needs: A study and analysis of user queries on the web. *Inf. Process. Manage.*, 36(2):207–227, January 2000. ISSN 0306-4573. doi: 10.1016/S0306-4573(99)00056-4. URL `http://dx.doi.org/10.1016/S0306-4573(99)00056-4`. 1

Rosie Jones, Benjamin Rey, Omid Madani, and Wiley Greiner. Generating query substitutions. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 387–396, New York, NY, USA, 2006. ACM. ISBN 1-59593-323-9. doi: 10.1145/1135777.1135835. URL `http://doi.acm.org/10.1145/1135777.1135835`. 2

Makoto P. Kato and Katsumi Tanaka. To suggest, or not to suggest for queries with diverse intents: Optimizing search result presentation. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, pages 133–142, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3716-8. doi: 10.1145/2835776.2835805. URL `http://doi.acm.org/10.1145/2835776.2835805`. 2

Johannes Kiesel, Arefeh Bahrami, Benno Stein, Avishek Anand, and Matthias Hagen. Toward voice query clarification. In *SIGIR*, pages 1257–1260, 2018. 4.1.4

Johannes Kiesel, Arefeh Bahrami, Benno Stein, Avishek Anand, and Matthias Hagen. Clarifying false memories in voice-based search. under review. 5.1.2, 5.2

Jinyoung Kim and W. Bruce Croft. Retrieval experiments using pseudo-desktop collections. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 1297–1306, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-512-3. doi: 10.1145/1645953.1646117. URL `http://doi.acm.org/10.1145/1645953.1646117`. 2

Jinyoung Kim and W. Bruce Croft. Ranking using multiple document types in desktop search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 50–57, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0153-4. doi: 10.1145/1835449.1835461. URL `http://doi.acm.org/10.1145/1835449.1835461`. 2

Alexander Kotov and ChengXiang Zhai. Towards natural question guided search. In *Proceedings of WWW 2010*, WWW '10, pages 541–550, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772746. URL `http://doi.acm.org/10.1145/1772690.1772746`. 2

J Lai and N Yankelovich. Speech interface design. 12 2006. 1, 2

Ewa Luger and Abigail Sellen. "like having a really bad pa": The gulf between user expectation and experience of conversational agents. In *Proceedings of CHI 2016*, CHI '16, pages 5286–5297, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3362-7. doi: 10.1145/2858036.2858288. URL `http://doi.acm.org/10.1145/2858036.2858288`. 1, 2

Donna J Nelson and Diana C Rogers. *A national analysis of diversity in science and engineering faculties at research universities*. National Organization for Women Washington, DC, 2003. 4.1.2

Jessie Ooi, Xiuqin Ma, Hongwu Qin, and Siau-Chuin Liew. A survey of query expansion, query suggestion and query refinement techniques. *2015 4th International Conference on Software Engineering and Computer Systems (ICSECS)*, pages 112–117, 2015. 2

Filip Radlinski and Nick Craswell. A theoretical framework for conversational search. In *Proceedings of CHIIR 2017*, pages 117–126. ACM, 2017. 2

Fabrizio Silvestri. Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in IR*, 4(1–2):1–174, 2009. 2

Johanne R Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. How do people interact in conversational speech-only search tasks: A preliminary analysis. In *Proceedings of CHIIR 2017*, pages 325–328, 2017. 2

Johanne R Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. Informing the design of spoken conversational search. 2018. 2

Sarah K. Tyler and Jaime Teevan. Large scale query log analysis of re-finding. In *Proceedings of the Third International Conference on Web Search and Web Data Mining, WSDM 2010, New York, NY, USA, February 4-6, 2010*, pages 191–200, 2010. 2

Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles LA Clarke. Exploring conversational search with humans, assistants, and wizards. In *Proceedings of CHI 2017*, pages 2187–2193. ACM, 2017. 2

Muuo Wambua, Stefania Raimondo, Jennifer Boger, Jan Polgar, Hamidreza Chinaei, and Frank Rudzicz. Interactive search through iterative refinement. 2018. 2

Yu Wang and Eugene Agichtein. Query ambiguity revisited: Clickthrough measures for distinguishing informational and ambiguous queries. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 361–364, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 1-932432-65-5. URL `http://dl.acm.org/citation.cfm?id=1857999.1858054`. 1

Ji-Rong Wen, Jian-Yun Nie, and Hong-Jiang Zhang. Clustering user queries of a search engine. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 162–168, New York, NY, USA, 2001. ACM. ISBN 1-58113-348-0. doi: 10.1145/371920.371974. URL `http://doi.acm.org/10.1145/371920.371974`. 2

Appendix A

User Studies Questionnaires

In the following pages we present our user studies questionnaires. We discussed the user studies in Chapter 3 and 4.

Voice Query User Study

This study will take 15 minutes. You will have to fulfill some tasks using our voice-controlled service and answer some questions regarding your experience with such services and your impression of our service.

During the task, your voice will be recorded and sent to voice recognition services of Amazon outside the European Union. We will use the recording for research purposes only. We will not share it with people that are not involved in this research. Note that we can not guarantee that Amazon does not make own use of the recording.

You may quit at any time.

I understood and agree with these terms

_____ _____

Place, date Signature

Pre-study questions

What is your gender? Female ☐ Male ☐ Other ☐

How old are you? 17 or younger ☐ 18-30 ☐ 31–49 ☐ 50–64 ☐ 64 or older ☐

How often do you use voice assistants? (Alexa, Cortana, Home, Siri, ...) Frequently ☐ Rarely ☐ Never ☐

If you use a voice assistant, which one do you use? _____

If you use a voice assistant, for what tasks do you use it? _____

How would you rate your English level? Proficient ☐ Intermediate ☐ Beginner ☐

Instructions

This study consists of a series of very short voice query tasks. For each task, you will receive a piece of paper. Please now take a careful look at the example we provide you.

The **scenario** briefly describes a setting with a certain information need. Please imagine yourself having this need in this setting, even if you already know the answer.

To resolve this information need, you ask Alexa. You can try this now. First say "Alexa" and wait for the blue ring to appear. Then say "Find!". You should hear the answer "Yes?". Then continue with the question at the second bullet point of the **Interaction start** section. You should now be hearing a response from Alexa, where she offers you three choices. Now try to continue this interaction with Alexa to resolve your information need. In case Alexa responds "Sorry, please try again!", then do so, starting again with saying "Alexa".

After you reached an answer or tried 5 times, please answer the **after-interaction questions** and then continue with the next task. Finally, fill in the **post-study questions**.

Scenario: You want to start reading a new book. You remember your mom was happy with the book she was reading called The Good Life which is an autobiography of a singer. But you do not know the author.

Interaction start:

- Alexa. Find!
- Who is the author of The Good Life book?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system answered my question	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: Telegram x is on the market and you want to download it. On the play store there are photos of the app. One shows the message from Heisenberg. Your friend says this is the character from Breaking Bad series but you are sure that he is a famous physicist.

Interaction start:

- Alexa. Find!
- Who is Heisenberg?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system answered my question	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: You are walking on the street and see a poster saying Rudolf Schenker has a concert next weekend. You remember that he is the member of the Scorpions, The German rock band, but not sure about it.

Interaction start:

- Alexa. Find!
- Who are the members of the Scorpions?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system answered my question	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: As a kid you loved Jimbo, the talking airplane but you dont remember the name of the series.

Interaction start:

- Alexa. Find!
- What is the name of the series with Jimbo?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system answered my question	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: You want to surprise your Irish partner with an Irish cocktail called B-52, but you dont know how to make it.

Interaction start:

- Alexa. Find!
- How to do a B-52?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system answered my question	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: You are living in the US and planning to go on a trip to Australia with your friend. Someone recommended you to see the Magic Mountain theme park there. You want to know where exactly it is.

Interaction start:

- Alexa. Find!
- Where is Magic Mountain?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system answered my question	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: You and your friend are texting on the phone and talking about your favorite comic characters. Your friend says his favorite character is Mercury from Marvel Comics. You dont know her, but dont want to lose face either.

Interaction start:

- Alexa. Find!
- Who is Mercury?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system answered my question	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: You and your colleague friend are talking together. You want to start watching a new series. Your friend suggests Life on Mars series. You tell him that it's an old series and you prefer to watch a newer one, but your friend thinks it's not that old.

Interaction start:

- Alexa. Find!
- What year is the Life on Mars produced?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system answered my question	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: You are drinking PALM beer with your friend. Its the first time you see this kind of beer and you get curious where it is from.

Interaction start:

- Alexa. Find!
- Where is Palm company?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system answered my question	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: You are playing cards with your friend. Your friend sees the Joker card and tells you that Joker is also a comic character from DC Comics. You know there is a Joker character in Batman. You want to make sure if they are the same Joker.

Interaction start:

- Alexa. Find!
- Tell me about the character Joker.

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system answered my question	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: You and your colleague friend are talking together. You want to start watching a new series. Your friend suggests Matador, the US television series. You tell him that it's an old series and you prefer to watch a newer one, but your friend thinks it's not that old.

Interaction start:

- Alexa. Find!
- What year is Matador produced?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system answered my question	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: You are home texting your friend from Monte Carlo, Italy. He is inviting you for his wedding. You think maybe it is a good opportunity to visit the city as well. You wonder how much time you'd need to go sightseeing.

Interaction start:

- Alexa. Find!
- What can I do in Monte Carlo?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system answered my question	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: Your father's friend is at the hospital in the Intensive Care Unit (ICU) and your father is visiting him. You are supposed to pick your father up. You are close to the place but dont know exactly where it is.

Interaction start:

- Alexa. Find!
- Where is the closest ICU?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system answered my question	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: You are talking to your friend about different Application Programming Interfaces (APIs) and run into question when exactly the first web API was created.

Interaction start:

- Alexa. Find!
- How old is the web API?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system answered my question	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Post-study questions

How would you improve the system?

If you have general comments on this study, please add them here

Voice Query User Study

This study will take 25 minutes. You will have to fulfill some tasks using our voice-controlled service and answer some questions regarding your experience with such services and your impression of our service.

During the task, your voice will be recorded and sent to voice recognition services of Amazon outside the European Union. We will use the recording for research purposes only. We will not share it with people that are not involved in this research. Note that we can not guarantee that Amazon does not make own use of the recording.

You may quit at any time.

I understood and agree with these terms

Place, date _____ Signature _____

Pre-study questions

What is your gender? Female ☐ Male ☐ Other ☐

How old are you? 17 or younger ☐ 18-30 ☐ 31–49 ☐ 50–64 ☐ 64 or older ☐

How often do you use voice assistants? (Alexa, Cortana, Home, Siri, ...) Frequently ☐ Rarely ☐ Never ☐

If you use a voice assistant, which one do you use? _____

If you use a voice assistant, for what tasks do you use it? _____

How would you rate your English level? Proficient ☐ Intermediate ☐ Beginner ☐

Instructions

This study consists of a series of voice query tasks. For each task, you will receive a piece of paper. Please now take a careful look at the example we provide you.

The **scenario** briefly describes a setting with a certain information need. Please imagine yourself having this need in this setting, even if you already know the answer.

To resolve this information need, you ask Alexa. You can try this now. First say "Alexa" and wait for the blue ring to appear. Then say "Explore!". You should hear the answer "Yes?". Then continue with the question at the second bullet point of the **Interaction start** section. Alexa will give you an answer or say she does not know. Your study supervisor will show you whether the answer is indeed the desired on (thumb up) or not (thumb down). In case Alexa does not know the answer or the study supervisor shows a thumb down, repeat this process, but drop one of the three facts (indicated by a grey background). If you still get a thumb down, try again with another two of the three facts.

After you reached an answer or tried 5 times, please answer the **after-interaction questions** and then continue with the next task. Finally, fill in the **post-study questions**.

Scenario: You try to remember the title of what you think is a German horror movie. In the moview, there are two college student girls vacationing and then attacked by a murderer.

Interaction start:

- Alexa. Explore!
- What is the title of the German horror movie in which two girls vacation and are attacked by a murderer?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system was helpful	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: You try to find the title of a movie that you think Morgan Freeman stars in. In the movie, Morgan Freeman offers a sniper a job to kill a person, but then tries to shoot the sniper himself.

Interaction start:

- Alexa. Explore!
- What is the title of the movie in which a sniper man takes an offer by Morgan Freeman but is then shot by his employer?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system was helpful	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: You try to remember the title of a controversial book that come out back in the 90s and claimed scientific evidence that whites are genetically superior to blacks. You think it was called something like "The *something* Factor."

Interaction start:

- Alexa. Explore!
- What is the title of the book from the 90s that claimed superiority of Whites and is called "The *something* Factor"?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system was helpful	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: You try to remember the name of a character who you think is from a Marvel comics, has wrappings around its face like a zombie, and wears a gray coat.

Interaction start:

- Alexa. Explore!
- What is the name of the Marvel comics character who has wrappings around its face and wears a gray coat ?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system was helpful	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: You are looking for lyrics of a country song from 2005, but you dont know the title or the name of the artist. The lyrics you remember are "when Im 90 sitting in my rocking chair and she is more generous than I could ever be".

Interaction start:

- Alexa. Explore!
- What is the title of the country song from 2005 with the lyrics : "when Im 90 sitting in my rocking chair and she is more generous than I could ever be" ?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system was helpful	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: You are trying to remember a movie from 2006 which you think Maggie Smith stars in. There's a teacher and she has an affair with a student and gets into trouble and Maggie Smith tries to take care of her.

Interaction start:

- Alexa. Explore!
- What is the name of the movie from 2006 in which a teacher has an affair with a student and gets into trouble and Maggie Smith tries to take care of her ?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system was helpful	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: You are trying to remember the name of a website; you think it starts with a "d" and the site's layout is blue, and it is for sharing files.

Interaction start:

- Alexa. Explore!
- What is the website for sharing files with a blue layout which starts with a "d" ?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system was helpful	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: You are trying to remember the name of the video game. The character looks similar to Rayman with invisible limbs and you think it has a yellow head.

Interaction start:

- Alexa. Explore!
- What is the name of the video game in which the character looks similar to Rayman with invisible limbs and a yellow head ?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system was helpful	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: You are trying to remember the title of one 2004 song that you think it is by Ruben Studdard. The lyrics you remember are : "Sunshine after rain, you have taken away my pain, somehow and some way, its over now, The storm is over now".

Interaction start:

- Alexa. Explore!
- What is the title of the 2004 song by Ruben Studdard that says : "Sunshine after rain, you have taken away my pain, somehow and some way, its over now, The storm is over now" ?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system was helpful	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: You are trying to remember the title of a movie in which a little boy takes home a baby crocodile and then his parents flush it down. Years later the baby crocodile turns big and launch an attack on the city.

Interaction start:

- Alexa. Explore!
- What is the title of the movie in which a little boy takes a baby crocodile home where his parents flush it down and years later the crocodile launch an attack on the city?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system was helpful	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: You are trying to remember the poem you think is from the Victorian era. It is about an urn and the painting on it, was of two lovers whose hands are just outside of each others reach.

Interaction start:

- Alexa. Explore!
- What is the name of the poem from the Victorian era which is about an urn and a painting of two lovers?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system was helpful	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: You are trying to remember the title of the cartoon that you think it starts with a "C" and its two main characters are the father and son bear and they escaped from the circus.

Interaction start:

- Alexa. Explore!
- What is the title of the cartoon that starts with a "c" and its two main characters are a father and son bear and they escape from the circus?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system was helpful	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: You are trying to remember a game you think it is a web-based game developed by Google. In the game you are a single celled organism that can move around and eat things.

Interaction start:

- Alexa. Explore!
- What is the name of the web-based game developed by Google in which you are a single celled organism that can move around and eat things?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system was helpful	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: You are trying to remember the name of the the 1999 rock album with a cover you think it had a single goldfish on.

Interaction start:

- Alexa. Explore!
- What is the name of the 1999 rock album that had a cover picture of a single goldfish?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system was helpful	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: You are trying to remember the title of the 2008 movie directed by Eric Valette in which you think people get killed by the cell phones.

Interaction start:

- Alexa. Explore!
- What is the title of the 2008 movie directed by Eric Valette in which people get killed by the cellphones?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system was helpful	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Scenario: You are trying to remember the title of the Japanese anime in which a girl named Ellen has a tattoo of a shark on her hand.

Interaction start:

- Alexa. Explore!
- What is the title of the anime from Japan in which the character's name is Ellen and has a tattoo of a shark on her hand?

After-interaction questions:	Agree		Neutral		Disagree	Don't know
The system was helpful	☐	☐	☐	☐	☐	☐
The system behaved as I expected	☐	☐	☐	☐	☐	☐
The system was easy to hear/understand	☐	☐	☐	☐	☐	☐
The system was pleasant to use	☐	☐	☐	☐	☐	☐

Post-study questions

How would you improve the system?

If you have general comments on this study, please add them here