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

Argument Search with Voice Assistants

Master's Thesis

Kevin Lang
Born Sept. 5, 1989 in Greiz

Matriculation Number 110010

1. Referee: Prof. Dr. Benno Stein
2. Referee: Prof. Dr. Ing. Eva Hornecker

Submission date: September 20, 2018

Declaration

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

Weimar, September 20, 2018

.....
Kevin Lang

User: *Alexa, why is animal confinement bad?*

Alexa: *Do you want to hear arguments about the topic "Humans should stop eating animal meat"?*

User: *No... Alexa, can you tell me whether I should visit the zoo?*

Alexa: *Open topic "Zoos should be forbidden" ...*

— excerpt from the 2nd study

Abstract

The need for finding persuasive arguments can arise in a variety of domains such as politics, finance, marketing or personal entertainment. In these domains, there is a demand to make decisions by oneself or to convince somebody about a specific topic. To obtain a conclusion, one has to search thoroughly different sources in literature and on the web to compare various arguments. Voice interfaces, in form of smartphone applications or smart speakers, present the user with natural conversations in a comfortable way to make search requests in contrast to a traditional search interface with keyboard and display. Benefits and obstacles of such a new interface are analyzed by conducting two studies. The first one consists of a survey for analyzing the target group with questions about situations, motivations, and possible demanding features. The latter one is a wizard-of-oz experiment to investigate possible queries on how a user formulates requests to such a novel system. The results indicate that a search interface with conversational abilities can build a helpful assistant, but to satisfy the demands of a broader audience some additional information retrieval and visualization features need to be implemented.

Contents

Acknowledgements

My first gratitude goes to the Webis group and all staff members who I worked with for several years and who raised my interest in web technologies, machine learning, information resources, artificial intelligence, and other related topics.

I would like to thank Prof. Dr. Benno Stein and Prof. Dr. Ing. Eva Hornecker for accepting my work under their supervision and Dr. Jan Ehlers for the help in statistical analysis.

A special gratitude for Johannes Kiesel who was not only my advisor during my bachelor's thesis but also guided me in this thesis, so that I can be proud of my work.

I would like to also thank my friends Mark for reading my thesis, Jakob for being a nice colleague in several works, René for sharing many hobbies and many other people who accompanied my study and helped me to reach my goals.

And last but not least, my wonderful partner Nathalie. She supported me in many life situations and was always there when I needed her. I could not have done most of my experiments without her and she is the reason why I was able to continue studying and to finish my master's thesis. With her, I have found a soul mate and a person I want to spend the rest of my life with.

Chapter 1

Introduction

Arguments are needed everyday to make decisions for oneself or to convince somebody in situations without an obvious right answer. This can be discussions like: if it would be better to adopt a cat or a dog for the household, or if it would be better to buy a desktop computer or a notebook. Next to these low-impact discussions, arguments are used in domains with far-reaching consequences. In politics, arguments are needed that speak for or against a decision, which can influence the economy or the society. Economists have to make decisions about investments that need a careful consideration of pro or con arguments.

The tedious search for convincing arguments is particularly problematic. One has to go through multiple sources (e.g., books, newspapers or countless websites) to get an overview of the pro and con arguments to a specific topic. Moreover, many of the sources are biased (Ulrike Hahn, 2009), vary in their quality in terms of logic, rhetoric, and dialectic (Wachsmuth et al., 2017a), or are deliberately producing misinformation in form of fake news (Mustafaraj and Metaxas, 2017). For these reasons, researchers in information retrieval and natural language processing have started to show a big interest in argument mining. Wachsmuth et al. (2017b) dedicated themselves to mining and retrieval of arguments from different debate portals, and introduced the web-interface *args.me*, a search interface, to present and rank arguments. Based on their system, which is still in development and getting features and improvements down to the present day, this thesis aims to develop a novel voice interface which could improve the quality and convenience of an argument search system for the user.

With the introduction of smartphones, voice assistants became more frequently used in everyday life and a popular topic in science. A few years later, home devices were brought to market. Compared to smartphones, these new gadgets had the

(a)

(b)

Figure 1.1: (a) shows the conversational A.I. "HAL 9000" from *2001: A Space Odyssey* and (b) Scotty from *Star Trek IV: The Voyage Home* who is confused how to use a voice interface in the past with a mouse.

advantage of providing a better natural sounding voice, can be easily activated with a keyword at home, have features for the control of smart devices, and provide an app store to let third-party developers create their own applications for the system. It seems that the dream of voice assistants and conversational AI became eventually true. Science fiction movies like "2001: A Space Odyssey" from 1968 and "Star Trek IV" from 1986 (see figure 1.1) already predicted that mouse and keyboard input devices are not common anymore in the future and are replaced by the more flexible and convenient usable voice interface of artificial intelligence systems.

This thesis develops a voice interface for an argument search system. In the following, core questions are proposed and answered with different approaches. The first question is, *why people want to use a voice-based argument search system.* Therefore, a first study was conducted to ask participants about situations and motivations to use such a novel system (chapter 3). Subsequently, a second study was conducted with a prototype in form of a Wizard of Oz experiment (chapter 4). Here, the core questions are *how participants interact with the system and which responses they expected.* The results are presented and discussed in the evaluation part of the first study (section 3.4) and the second study (section 4.4) and summarized in the conclusion (chapter 5).

Chapter 2

Related Work

This chapter focuses on related work, starting with some essential definitions and research this thesis is based on, followed by research on the acceptance of voice interfaces in different situations and under different motivations. The major part will be about models and design guidelines of voice interfaces and how different studies analyzed their specific search engines. Many works use different terms for the voice assistant, e.g. voice activated personal assistant, conversational search system, conversational agent, dialogue interface system, spoken dialogue system, voice user interface, or simply agent as a general term. In this thesis, the more general term "agent" was chosen because an "activation" of the system is not always required, and the system does not have to be on a human "conversational" level completely when it follows simple search patterns.

First of all, this paragraph presents a few clarifications about argumentation terms and some words about the structure of argumentations. Walton et al. (2008) defined an argument as a conclusion which is supported or attacked by at least one premise. The conclusion itself is a claim with a stance towards a topic. An example for a conclusion which is also used in this thesis is the statement "Zoos should be forbidden". "Zoos" are the topic in this example and "should be forbidden" the negative stance to it. A supporting premise to this conclusion would be "Animals confined to zoos suffer negative psychological effects". This statement gives a reasoning to the conclusion and with this builds a valid argument. However, premises like "I don't like Zoos" convey a stance, too, but are missing a reasoning. These statements are usually classified as opinion. Because opinions do not have a persuasive nature, they are not used in the experimental-setups of the studies of this thesis. In addition, a premise can also be a claim which can again be supported or attacked by other premises. This leads to a complex argument

structure. In this regard, Stab and Gurevych (2016) introduced some elementary and fine-grained argument relations. A new web interface which supports an ongoing chain of arguments is the debate platform kialo[1]. It demonstrates how debates can have a complex structure of arguments which overlaps with multiple topics. However, the navigation of such a complex argument structure with a new voice interface without visual output, would strain the cognitive load of the user. For this reason, only simple one level argument structures were used in the studies of this thesis to give the user an easier view of the resources of the system.

Nevertheless, this thesis does not focus on the mining of valid arguments because it is a follow-up work to the project "args.me" from Wachsmuth et al. (2017b). They developed a web interface with a search engine to find arguments from different debate portals. The arguments are presented in a ranked list with a positive and negative stance column to the conclusion the user typed in as a query. To make this system more convenient in the everyday life, this thesis wants to enhance it with a new voice interface. The user should get the opportunity to use his or her agent in form of a smart gadget to obtain arguments on the fly without drawing on keyboard or touch interface of other devices. Enhancing an existing system with a voice interface for navigation and output is not a new idea. Rohde and Baumann (2016) developed a framework to navigate the *Spoken Wikipedia*. The main idea here is the temporary storage of key words and links, which the user can look up later. A current example is the project 'Scout' from Mozilla[2]. The developers of Firefox work on a purely voice based web browser. This means, they want to create a browser which can be fully used by voice and presents all results with synthesized speech.

One important topic concerning agents is the acceptability of their usage by the user with different motivations in various locations and situations. Easwara Moorthy and Vu (2014) conducted one of the first studies with smartphones and agents in 2014. They analyzed the differences in acceptance between speaking and texting somebody, talking about both private and non-private subjects, and using the smartphone at home or public facilities. Similar to findings from previous research, they found out that people do not like to talk about private information with agents but are willing to do so when they are in familiar places. They point out that people maybe behave differently using an agent when they are observed. This behavior is called the Hawthorne Effect(Carey, 1967) and could be an issue related to the public use of agents. One purpose of this thesis is to figure out if the search for arguments counts as private subject or if participants dissociate themselves from the topic when using the argument search engine. If the case

[1] https://www.kialo.com/explore, accessed 08.09.2018
[2] http://winfuture.de/news,103648.html, accessed 08.09.2018

should come up that most of the motivations to use an agent for argument search is treated private, then the field of application would be hardly limited. A possible implementation of the system could be called into question.

Efthymiou and Halvey (2016) conducted a more fine-grained study about the acceptance of locations in which agents can be used, namely driving a car, being at home, being in a metro as a passenger, walking on the pavement, sitting in a pub, or being at the workplace. In addition, they investigated the acceptance of audiences when the user is alone or together with colleagues, family, friends, partners, or strangers. The tasks presented to the participants were categorized in finding a direction, search for information, and entertainment. While the location "home" shows again the highest acceptance rate in combination with all three tasks, the task "search for information" has the lowest acceptance rates in every other location and audience situation. They explain this lower acceptance rates with the nature of the task of information search. Searching for directions is a task that everyone will encounter when he or she is in a new location or wants to be prepared for a trip. It is a demand one does not have to be embarrassed about. Additionally, entertainment has no hard constraints in locations and situations because no sensitive data is handled here. However, search for other information can disclose private issues in front of strangers or even closely related people. This insights could be confirmed with statistical values and also post evaluation interviews. Still, search for information has a higher acceptance rate in a working place compared to entertainment scenarios and finding directions. Even though unstated in the work of Efthymiou and Halvey (2016), it seems more appropriate to search for something when it is related to the employment which can be seen in their data. In the context of this thesis, a study was conducted about different situations of using a voice assistant for argument search. The results will show if they confirm the acceptance rates of the last two mentioned studies. Further on, several motivations are introduced which are adapted to argument search and provide insights in this special case of information search.

The following scientific works present different approaches on how to model an agent's behavior for information retrieval tasks. Radlinski and Craswell (2017) analyzed many former experimental set-ups for conversational search and formulated a definition for such systems:

> *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 about system feedback and expected responses from the user. This framework can be used to model different intent and response strategies on the server side of a voice assistant, but because this thesis pays attention to the acceptance and usability of the novel argument search system, an implementation and consideration of this framework remained on a low-level. Moreover, the proposed framework and upcoming mentioned scientific works have the problem that the requested results are mostly single items. In other words, the experimental set-up consists of a user with a search task and is fulfilled when the user found the best item, e.g. a phone number of a person or the cheapest flight to a country. The main goals of argument search, on the other hand, are to make a decision or to convince somebody. Both target states are quite subjective and can not be determined by the system when they are reached. Nevertheless, Radlinski and Craswell mention two other important considerations besides the framework. First, the system needs a memory to remember queries and insights the user delivered during the conversation to give more targeted answers to the user and to handle better future user requests. Secondly, a conversation with back-and-forth dialogue system is not needed when the user only expects a listing of items. Agents can be widely useful for information retrieval tasks without reflecting a human level conversation. The second study of this thesis is conducted with a Wizard of Oz experiment. The memory of the requests and presented arguments is handled by a human and the opportunity to navigate between arguments or to show specific information is given, but does not need to be utilized to accomplish the task. This makes the system easier to implement and less complex for the participants.

Carrying on the problem regarding memory of the system and the user, Azzopardi et al. (2018) defined "Current Information Needs", which change and evolve so that the agent needs to adapt them based on the "Past Information Needs" and their corresponding set of associated objects. For this, they defined classes of actions for the user and the agent. The actions for the agent consist of *inquire*, *reveal*, *traverse*, *suggest*, *explain*, *ending*, *error*, and *finalization actions*, while the user actions are classified as *inquire*, *navigate*, *interrupt*, *interrogate*, and *closing actions*. Based on these classes, actions for the second study of this thesis were designed.

Another point is the classification of information needs in precision-oriented or recall-oriented, introduced by Papangelis et al. (2017). They conducted a study about hotel search and reservations with agents. The precision-oriented information needs are requests by the user to locate and deliver single resources, while for the recall-oriented tasks the agent has to analyze and compare resources with their relations to make decisions. Because the argument search system

proposed in this thesis does not make decisions on its own (e.g., when the user asks "What is your opinion about the topic?" the system gives no answer and only redirects to the arguments), the information needed is only classified as precision-oriented. Nevertheless, the user is free to ask any question about combined information (e.g., "How many arguments are there about batteries?"), and the human agent answers them as well as possible.

Wildemuth and Freund (2012) identifies exploratory tasks by the following key attributes: they are associated with the goals of learning and/or investigation, are rather general than specific, can be open-ended, can target multiple items, can involve uncertainty, can arise through ill-structured information problems, are dynamic, lengthy, multi-faceted and complex, and are accompanied by other information and cognitive behaviors, e.g., sense-making. All this attributes are presented in the argument search system. Subsequently, Wildemuth and Freund gives advice on which preparation should be considered before implementing an exploratory search system. First, it is advisable to analyze the logs from existing systems. The only available query logs which were on hand are from the argument search web interface "args.me", presented above. Because the interface is quite new, the amount of user queries is very limited. Nevertheless, the log data shows that users are highly interested in topics like abortion, Donald Trump, universal health care, ban of the sale of video games to minors or the assassination of dictators. Other queries for simply testing the system were also quite common, like simply writing "test" or "google" into the search field. The logs show an interest of the users in controversial topics, but also an testing character to see how the system reacts to special text inputs. Another suggested method is the interview or observation of people completing the task of an exploratory search. In the time frame of this thesis it was not possible to make an additional study with people to see how they use an argument search system on the web. This would be a good topic for future work, to compare data from a web interface with a voice interface. This thesis limits itself to the evaluation of a voice interface. The insights gathered from these methods can be used to design use-cases for the first experiments with the system. For the studies of this thesis, mostly example situations were chosen, which are quite common in everyday life, and topics, which are often discussed on debate portals.

Rohde and Baumann (2016) suggest a framework to deal with a large scale of spoken texts. Their system uses fuzzy matching to attain the structural hierarchy, timing of all time-aligned words, sentence segmentation, and hyperlinks of an article. These features allow the application to leap over sentences, paragraphs or sections, and to navigate links and key words close to the current timing of the record. This temporary networking scheme could also be transferred to search

systems with longer texts from debates. However, this has to be supported from the server side of the agent and needs a proper implementation strategy. This task would be too comprehensive for this thesis, so it would be nice to consider it for future research.

Kaushik and Jones (2018) analyzed the behavior of users when confronted with search tasks. They used normal web search with text input and could identify four different types of behavior: type A enters one query and selects one document that delivers the information need, type B opens multiple documents to get a combined result that delivers the information need, type C performs an iteration of different queries and inspects the results of their worthiness to investigate them, and type 4 is not sure about expressing its information need and changes the behavior during the search. It would be interesting to see if these types of user behavior are also identifiable with a voice search interface. Still this analysis of a visual search in comparison with voice search is an issue for future work because it would be too labour-intensive to design a study for both search tasks by now.

The next paragraphs address different examples of research when people encounter voice interfaces. Porcheron et al. (2018) state that homes are multi-activity settings in which devices get recruited into and are regulated through the ongoing cooperative and collocated activities that take place there. In conclusion, voice assistants need coordinated actions from other members. The problem is that current available voice assistants can only listen to one user at a time and get confused when many sources of noise exist. Similar to Radlinski and Craswell (2017), their study points out that there should be a shift from conversation design to request/response design because of the lack of complexity in the search task. Porcheron et al. present some questions designers should consider when they build an agent such as:

- Is this response an interactional dead end?
- What resources does this response provide for a possible next request production?
- What might possibly be 'done' with this response?
- At which points might a user interrupt and take the next turn?
- How does the response design employ moments of silence?

The last question especially points out a big problem in voice interface interactions. The system could reach a dead end (e.g., all arguments were listed) and the user does not know how to handle the situation. The silence that follows from an uncooperative voice system is treated as troublesome in these moments for most of the users. The chosen design for the second study of this thesis assures that the user gets enough feedback at which position in the corpus of arguments he or she

is situated right now and how a continuation is possible.

Luger and Sellen (2016) made a qualitative study of Apple Siri, Google Now, and Microsoft Cortana users by observing their interactions with the voice assistant over a few weeks. Their main focus lied in the analysis of why people use voice assistants and how their expectations and experience influence the usage of such gadgets. The starting point for many voice assistant users is to play with the system and then to discover how easy it can be to manage small tasks compared to the usage of keyboard or touch screen: *"My feeling originally with Siri was that it was a toy... you'd ask it to do stupid stuff and then you start to do certain things with it and it starts to work, you know, like putting stuff in your calendar, and then it just becomes like an easier way of doing things"*. These playful interactions act as affordance to show the user the possibilities of a voice assistant and characterize the initial engagement with these systems. The first possibilities could be simple tasks like asking the voice assistant to check the weather or setting the alarm. The only problematic situation was when the system had a lack of feedback or transparency and failed to perform a task more than a few times: *"I gave it the benefit of the doubt... and then I thought no, you're always going to be rubbish"*. Such experiences can destroy the first contact of a user with an agent and are one of the reasons why many people reject voice applications.

Myers et al. (2018) made a detailed analysis of obstacles users of voice assistants are facing. They conducted a study with a total of 12 participants. Each of them had 3 tasks in which they could create, modify, delete, or invite other users to calender events. The study could record in total 146 obstacles users had to face, which makes in average around four obstacles per conversation with the agent. All obstacles were classified in four categories: NLP error (52,1%) when the request was misheard and mapped to a wrong intent, unfamiliar intent (20.5%) when the user tries to use an utterance the system cannot identify or the intent simply does not exist, system error (14.4%) which comes from a flaw from the system's architecture, and failed feedback (13.0%) when participants were observed to have ignored or misinterpreted the feedback of the system which caused further errors. Responding to these obstacles, ten classified tactics were used to overcome them: hyperarticulation, simplification, new utterance, use of more information, relying on GUI, settling, restarting, frustration attempts, quitting, or recall. Myers et al. state, that despite the fact that NLP errors had an occurrence of 52.1%, which was the highest compared to all other obstacles, the other categories seem to cause more frustration and confusion for the users. They needed a more distributed range of tactics for overcoming them, which indicates that the users do not have a correct mental model of the other obstacles and were less clear on how to solve them. The high number of possible errors which mostly occur

from natural language processing, missing feedback, and system errors are a big problem considering a proper evaluation of current voice assistants and their underlying technology. For this, the second study of this paper was conducted with a Wizard of Oz experiment to avoid most of the obstacles and to make a proper evaluation of a working system.

Wizard of Oz experiments became quite common within the research with voice assistants to fill the current gap in technology (Wolters et al. (2009), Trippas et al. (2017), Vtyurina et al. (2017), Vtyurina and Fourney (2018), Avula (2018)). Dubiel et al. (2018) had a similar study design and set-up to the second study conducted in this thesis. They used a Wizard of Oz experimental set-up to compare two different agents. Their goal was to explore how the people's search behavior changes when they are confronted with an agent which supports natural language interactions and one which does not. The first voice assistant acts like the prototype of the second study of this thesis. It gives the user the freedom to ask any question and to parse the information in an arbitrary order. The second prototype represents the current state of the art where the voice assistant gives the user an introduction how to use the system and the user has to provide the information in a specific pattern (e.g. "I need a flight from X to Y on the date Z.") so that the agent can parse and understand the query from the user. They collected data in the form of task completion time, task completion success, length of participant's turn, and more to get detailed statistics. The results showed that the conversational system was preferred more than the current voice based system. The superior system leads to significantly faster search task completion times and greater usability. Only a small number of people preferred the other system because they liked command control of gadgets and found it more predictable. As previously mentioned, the second study of this thesis uses an agent which supports natural language interactions, too, but distinguishes between two other modes: Does the user like to get a guideline at the beginning of the experiment or not? For this, the time was measured and the interactions of the participants with the system were recorded. Unfortunately, the data could not be evaluated in the same scope as Dubiel et al. did in their research because of time constraints and the amount of work, but further analysis can be done in future work.

Chapter 3

The 1st Study

This chapter takes a closer look on the first study of this thesis, which was conducted with an online survey. The goal of this study was to get insights about the situational acceptance of a novel voice argument search system and how users agree to the motivations presented for the system. Here, attention was paid to the users' needs and what discourages them to use the system. Furthermore, demographic data was collected and it was asked which features for voice assistants are preferred most. The results of this study can be aligned with the data of Easwara Moorthy and Vu (2014) and Efthymiou and Halvey (2016) about acceptance of voice assistants and can further be evaluated about the motivations of an argument search interface. While Wachsmuth et al. (2017b) invented a web interface for searching arguments, it has not been evaluated yet how people would use it and which use cases they would have. Hence, the public survey did not only attain data which can be used for modelling a voice assistant, it also gives insights on how the existing system can be improved. To evaluate the data from this study, some hypotheses are proposed which were derived from the previously mentioned works.

Hypotheses:

1. Mostly younger people use voice assistants.

It comes as no surprise that most of technology today is used by younger age groups because of the fact that they grew up using it [1]. On the other hand, smart home devices are advertised as gadgets which are convenient to use and that

[1] http://www.atechnologysociety.co.uk/how-young-generation-accepts-technology.html, accessed 08.09.2018

can assist older people at home (Morris et al., 2013). The distribution of the age groups of people who use voice assistants will be analyzed in this study and which anxieties or other reasons discourages them to not use them.

2. People prefer to use voice assistants for argument search at home instead of using them in public or at work.

3. People prefer to use voice assistants for argument search when they are alone.

This two hypotheses deal with the situations, in form of locations and audiences, in which people would like to use a voice assistant for argument search. The more detailed study by Efthymiou and Halvey (2016) could already show that people prefer to use voice assistants alone and at home regardless of the task. However, the acceptance rate of using a voice assistant for search tasks at work falls to 41% and to 16% when the user is in public. Likewise, the rates drop to 68% when a friend and 18% when a stranger is involved. The results of this study should have the same characteristics.

4. People would like to use a voice assistant for argument search more for convincing other people than for making their own decisions.

5. Entertainment and fun is an important aspect to use the voice assistant.

The hypotheses 4 and 5 are about the motivational aspects of using a voice assistant for argument search. It is assumed that people still prefer an extensive search for arguments on the web to build an opinion for a decision but will use a more faster and convenient voice assistant, when they want to convince a person nearby. The latter hypothesis is originated from other observations like Luger and Sellen (2016). People do not use voice assistant applications with their full range of functions from the beginning. They like to play with them and test their potentials for further usage. This study will show how much the participants appreciate entertainment motivations over the serious motivations stated in the fourth hypothesis.

6. People are interested in the source of arguments.

7. People want to get involved in the application, for example by adding new arguments or rating already existing ones.

In times in which “Fake News” became a serious issue in media (Mustafaraj and Metaxas, 2017), it is assumed that people are highly interested in the trustworthiness of information resources. A search engine for arguments can provide the feature for the lookup of argument sources which satisfies this desire. Additionally, it is assumed that people like to improve the quality of the argument corpus by adding new arguments or by rating the already existing ones. This desire is derived from the former hypothesis with the difference that people like to take control of the system by themselves to influence the quality of the arguments.

The further sections in this chapter will describe the structure of the online survey followed by the evaluation of the data to answer the hypotheses.

3.1 Experimental Set-Up

The online survey of the first study was structured in four pages that contain questions on demographics, the acceptance of situational and motivational use cases, the rating of possible functionalities of the novel system, and the possibility to sign up for a follow-up study, respectively. The distribution of the online survey will be described in subchapter 3.3.

3.1.1 Demographic Questions

The first study contains eight questions on the demographic background of the participants.

The aim of the first question was to ask for the gender of the participant. It is desirable to have an equal amount of participants of every gender to avoid false correlations. The online survey was forwarded to computer science and other media facilities with additional word-of-mouth recommendations, to span a preferably high and equal number of female and male participants. However, it is expected that more male people take part at the study because of the imbalance of genders in technically adept facilities Falkner et al. (2015). The number of the reached participants and their demographic background is reported in the subchapter 3.3 “Process”.

The second question is about the age of the participants. Five typical age groups were chosen from up to 17, 18 to 30, 31 to 49, 50 to 64 and 65 years or older. It is expected that most of the participants come from the second and third age group

because computer science is still considered to be a respectively “young” field, and also many students took part in the study.

The next question has the purpose to analyze how often the participants use computers or other intelligent systems. With this question, their affinity towards technical devices is estimated. People with a very low affinity towards technology should perhaps be excluded from the evaluation because they may not project themselves into the use cases presented in the next section.

The fourth question aims to ask about knowledge in specific fields of computer science, namely Human Computer Interaction, Information Retrieval and Natural Language Processing. All these fields are linked to the research of an argument search engine with a voice assistant. People with a broader background knowledge may have better insights to answer specific ratings or ranking questions.

The purpose of the fifth question is to get insights on how often the participants use voice assistants and how experienced and open they are for such a technology. When the participant is not interested in voice assistants, he or she can write a reasoning behind the answer in a text field. The qualitative data collected with this question will show which anxieties or other reasons discourage people to use voice assistants.

The sixth question should give some indication of how often the participants inform themselves about controversial topics on a daily, weekly, or less often scale. People with a low attraction to news and debates may have a small interest in a debating application.

The seventh question of the demographic page is about how engaged the participants are when it comes to debates on the web. It states if the participants visit debate portals, like debate.org, debatepedia.org, or idebate.org, before and if yes, whether they have already contributed arguments there. It is assumed that people, who are a member of debate portals, have a higher interest or maybe a different opinion about a new debate system.

A last question was added after many participants of the pilot study complained about the too specific examples for debate platforms of question seven. This will be reported in the subchapter 3.2 “Pilot”. The demographic part of the online survey got an additional question about social media platforms which is a more common case to post arguments.

3.1.2 Use Cases

The second page of the survey is the central part of the study and addresses the acceptance of different situations and motivations in which a voice assistant with argument search features can be used. The cases for situations were taken from Efthymiou and Halvey (2016). Here, they distinguished between the locations at home, while driving, in a pub, walking on a pavement, as passenger and during work. In order to have less use cases in the survey and to eliminate redundant cases, the situations were summarized to the more general ones at "home", in "public" and at "work". Their second investigation was about the audience when a voice assistant is used, which they distinguished by being alone, with a family member, a colleague, a partner, a friend, or a stranger. Again, close-up persons were summarized to the case "friend", unknown persons to "stranger", and a solitary situation as "alone". The combination of each place with each audience case results in a total number of nine different situations. Three of them were discarded of the following reasons. The case "home-stranger" is too unrealistic to create a meaningful use case with it. It would not make much sense to use the voice assistant for argument search when a foreign person is in the home of the user. The case "public-friend" appears to be quite similar to the case "public-alone" because there is already a differentiation with the cases "home-alone" and "home-friend" between a solitary and a friend condition and the results should be similar here. Finally, the case "work-alone" was also discarded because of similar reasons to the former situations in which "alone" and "friend" cases are already distinguished and no clear difference is expected to the case "work-friend".

Every use case is rated by the participants for the acceptance of the situation and the motivation. For the motivation, two obvious possibilities where one could imagine to use a search engine for argument finding were chosen, namely to "make a decision" and to "convince somebody". Other possible motivations would be to "form an opinion" or to "support media", e.g. when a person watches or reads something and he or she wants to get more arguments about a topic because the consumed media is too one-sided. The first of these optional motivations was discarded because it was too similar to the "make a decision" case, and although the second case would be interesting to analyze, it would be too much work to distinguish every possible media or topic. This could be done in a further study which gives more insights into the support of other media with an application. A third case is the "fun" motivation. The work of Luger and Sellen (2016) showed that people mostly learned how to use voice assistant applications by testing them out and have fun with them. Although an argument search engine can be used to get serious information about controversial topics, it can also be misused to get arguments about pointless debates. Two amusing motivations were added to the

survey to analyze the reaction of the participants.

		Situation			
Case	**Topic**	**Location**	**Audience**	**Example**	**Motivation**
A	political vote	home	alone	breakfast table	make decision
B	work uniform	work	stranger	counter of a bakery	convince s.b.
C	duck as pet	public	alone	park bench	have fun
D	electric cars	work	friend	office	convince s.b.
E	notebook or desktop	public	stranger	electronics store	make decision
F	pizza hawaii	home	friend	home in kitchen	have fun

Table 3.1: Use cases with topic and assigned motivation and situation with example.

Table 3.1 summarizes the use cases with their topics and attributes in situation and motivation. The complete written-out use cases can be looked up in the complete survey in the Appendix from page 75 to 82. The participants had to give a separate rating for the acceptance of the situation and the motivation. The rating scale goes from “convenient”, “plausible”, “unreasonable” to “inconceivable” and an optional “don’t know” field which rating was ignored in the analysis. Inside the use case texts the part about the situation was highlighted with a light yellow background color while the motivational part was highlighted with a light blue background color. It was decided to use this highlighting scheme to make it easier for the participants to distinguish between the two cases in the rating.

3.1.3 Functionalities

The third and last page of the survey has in total ten questions about how participants appreciate possible functionalities and different ranking criteria. The participants could rate the functionalities on a scale from “much appreciated”, “appreciated”, “nice to have” to “useless” with an optional “don’t know” field which was ignored in the analysis. Because some features are maybe difficult to understand when they are briefly explained, an example was added under each question in form of a possible request to the voice assistant. The intent of this questions was to determine which features should get a higher priority in building a prototype than other components.

The first question presented the main functionality of the new system which is to get pro and/or con arguments on a specific topic. It represents the baseline of this page. When people start to question this functionality then maybe an implementation of the system could be in vain because there is no real demand in such an application.

The second functionality is about getting the total number of arguments about a specific topic. It is already expected that this feature is not as appreciated as the first one, because the number of arguments does not have much meaning when they are mostly repeating or do not have much strength. It also does not make much sense to compare the total number of arguments across different topics. Nevertheless, it would be interesting to see how many people still appreciate this functionality.

The third question is about a navigational functionality. It stated that one can get pro or con arguments about a specific topic when adding a keyword. For example, when a user would like to know opposing arguments about electric cars in regard to "batteries", they can ask about arguments by adding this keyword. It is assumed that people prefer this feature more than the first one because arguments are not listed in an arbitrary order defined by the system but in favor of a specific subject.

The fourth and fifth questions are of similar nature and are about requesting the quality of the arguments. The first functionality gives the user the possibility to get more evidence about a stated claim in an argument. Maybe one premise is not enough for the user and he or she wants to know more facts which support or oppose the first stated claim. For example, this feature would be important if the user wants to disprove the statement of another person and needs specific arguments that speak against it. This functionality could be one of the most important components of the whole system if one really wants to hold a debate. The fifth functionality on the other hand gives the source of the argument, so from which person or website this citation is originated. The source can also be a big indicator for quality when it is from a reliable news website or only a post from a social network platform with no moderation or investigation.

The sixth and seventh questions concentrate on functionalities which involve the user himself or herself. The first one states that the user can rate already existing arguments in terms of how good they are and the seventh states that the user can add arguments to a topic by himself or herself. This questions should show how much the participants want to get involved into the system or how meaningful they think it is to rate or add arguments. Such extra functionalities can improve a system because arguments are not only crawled from the web but also can make it worse when a big part of the contributions are of vandalism nature. For this purpose, a description was added to the question that malevolent usage of those features will be prohibited.

The eighth question is about playing a game. The user tries to find the best argument to a topic and the system responds with a ranking score and how many

arguments exist which are better than the mentioned one. The ranking will be determined by the ranking system of the search engine. This functionality would be a first step to have a debate-like application against or with an artificial intelligence and is quite futuristic. The results will show if people like this idea, depreciate it or think it is not realizable now.

The ninth question has an open text field and participants can write suggestions for other functionalities they would like to have implemented in an argument search engine for a voice assistant. Although already eight functionalities about search, navigation, contribution and game characteristics were stated, the participants might have ideas for other functionalities which also make sense to include into the system.

The tenth and last question of the survey is a rating of ranking criteria how a search engine should order found arguments in the result list. In total, there are six different ranking criteria. The first is "Machine Learning" and states that the argument strength is rated by internal algorithms of the system. The second one is "User Rating" in which the strength of the arguments is determined by the rating of the users if it is available to the argument. The third one rates the arguments by the trustworthiness of the argument source scored by a community and is called "Source Reliability". On the other hand, there is "Source Coverage" which states that the listed arguments should come from different sources. Similar to the last one is "Aspect Coverage". Here the arguments should come with different aspects, comparable with the third introduced functionality of finding arguments with keywords, only in that distinct keywords are important here. The last ranking criteria is "Recency" and simply states that arguments are preferred which are most up-to-date. All the ranking criteria could be rated from "most important" to "least important" in a six-point scale with the additional "don't know" option if they can not imagine how the criteria works. It is assumed that "Aspect Coverage" and "Recency" will get the highest rating compared to the other criteria because arguments should be many-sided and up-to-date to persuade a bigger audience.

3.1.4 Invitation to the next Study

The fourth page does not contain any questions. It thanks for the participation of the survey and invites to the following study. When the participant wants to stay informed he or she can leave an e-mail address.

3.2 Pilot

Three students and three research assistant at the Bauhaus-Universität Weimar were asked to take part in a pilot-study for the survey. Completing times were measured and notes taken if the participant had a remark to any part of the study.

The time to complete the study was around 6 minutes in the fastest case and 13 minutes in the slowest. Besides a few comments about small writing mistakes or suggestions for rephrasing, there were other proposals which were incorporated in the survey.

First, some of the participants complained that the last question in the demographic part only refers to debate portals. Many people like to express their opinions not only on debate portals, but on social network platforms, news websites, or other message boards. Therefore, an eighth question was added to the first page which states if the participants write or read comments on such websites and if they show an affinity towards controversial topics there.

Another remark which was often stated was the differentiation between "Situation" and "Motivation" from the use cases on page two. The use cases were already highlighted with background colors if a part belongs to the rating of the situation or the motivation, but the participants still had problems to rate them separately. Hence, to make the use cases easier to rate, in front of each rating case an introduction phrase was added which summarizes the situation or motivation. For example, instead of only writing "*Rate Situation:*" or "*Rate Motivation:*" phrases were added like "*For me, using the assistant in a crowded store would be ...*" for a situational rating and "*For me, using the assistant to make a buying decision would be ...*" for a motivational rating. This new introductory phrases were highlighted with the colors yellow for the situation and blue for the motivation which were also used inside the text of the use cases.

One last comment was the unclear definitions of the ranking criteria from the last question of the functionalities page because they are too complex. It was said that criteria like "Source Coverage" or "Recency" have advantages and disadvantages that are difficult to compare with the other criteria and depend on the underlying data and topic. It was decided not to change this aspect because the participants should rate these criteria from their view regardless of how much insights they have about them. To make the criteria description a bit more understandable only small explanations were added, e.g. how specific ranking lists were created.

The complete online survey in its release state can be seen from page 75 to 82 in

the Appendix.

3.3 Process

In this part, the organisation of the survey is described and first insights of the demographic data are presented. The survey was accessible online from the 10th of January to the 24th of January 2018 on the platform *Umfrage Online* [2] in both English and German. The access to the survey was distributed via the mailing lists of the bachelor and master courses of Computer Science for Digital Media, the master course of Human-Computer Interaction at Bauhaus-University Weimar, research groups related to information retrieval and fake news detection, twitter, google groups, friends and family members. The recipients of the invitation e-mail were allowed to pass the mail on to other interested people.

The online survey was open for two weeks and achieved a total number of 97 participants. However, 30 participants were discarded of the following reasons: 20 participants did not finish the survey to the last page by leaving mostly after submitting the first page about the demographic questions, 9 participants used the lowest rating in over 70% of the cases and 1 participant used only the highest rating in 70% of the cases. For the analysis 67 participants remained with 39 completing the survey in English and 28 in German.

It was found that most of the participants were young adults. In numbers, 49 were between 18 and 30 years old, 11 between 31 and 49, 6 between 50 and 64, and a single one over 65 years old. It also seems that many participants already experienced strengths and limitations of voice assistants because 17 participants use them frequently and 33 participants used them rarely. 9 participants stated that they are interested in voice assistants and 8 stated that they are not interested in them at all. The non-interested participants could state reasons for their displeasure of voice assistants. The most mentioned problems are the inefficiency and bad voice recognition of the system and privacy issues. These problems are further analyzed in the section 3.4.2 of the qualitative data in the evaluation part.

The last questions on the demographic page address the interest and contribution to controversial topics. 40 participants stated that they inform themselves daily about debates, 19 do it weekly and 8 do it less often. Subsequently, the question if the participants contribute to debate portals, e.g. like debate.org, idebate.org, or debatepedia.org, followed. In total, only 2 participants stated that

[2] https://www.umfrageonline.com/, accessed 15.09.2018

they write posts on debate portals and at least 24 stated that they read discussions there. Everybody else were not aware of them. Regarding the contributions to other discussion platforms, the numbers are quite bigger. 18 said that they write comments on other social media or message board platforms and 35 stated that they at least read posts there. In conclusion, most of the participants are more consumer than contributor to controversial topics but stay informed at least once per week.

3.4 Evaluation

This section focuses on the question whether the hypotheses, which were presented at the beginning of the chapter, were confirmed by the results of this study and if further insights were found. It starts with the quantitative data which was collected from Likert-Scale ratings, followed by the qualitative data which was collected from comment sections.

The descriptive data was created with the open-source statistics program *JASP* (JASP Team, 2018). For cluster analysis, data management and to calculate the significance values with the Wilcoxon signed-rank test, the software package *SPSS* (IBM Corporation, 2013) was used. The corresponding effect size r of the datasets was calculated manually with the formula presented in equation 3.1.

$$r = \frac{Z}{\sqrt{n_x + n_y}} \tag{3.1}$$

The value Z is the absolute score of the distance to the mean in standard deviation units, while n_x and n_y are the valid numbers of measured values from the first and the second dataset in pairs. Pairs with values which were rated as “don’t know” were excluded in this evaluation. By Cohen (1988) the effect size r is indicated as small with a value of 0.1, medium with a value of 0.3 and large with a value of 0.5 or higher. The diagrams in the following sections were created with the free software environment for statistical computing and graphics *R* (R Development Core Team, 2008).

3.4.1 Quantitative Data

The quantitative data was collected to analyze the hypotheses stated at the start of Chapter 3. The following paragraphs describe the hypotheses and how the data underpins or negates them.

1. Mostly younger people use voice assistants.

The first hypothesis focuses on the age of the target audience. The results show that the target audience not only consists of people in the younger age groups.

Over 74.6% of the participants stated that they have already used a voice assistant and over 13.4% said that they are interested in using one in the future. Out of these people, 42 are in the age group of 18-30 years, 11 in age group of 31-50 years, 5 in the age group of 50-64 years and one person in the group of 65 years or older. Although the main targets audience, consisting of 71,20% of all participants, can be found around the age of 18-30 years, around 1/3 of the target audience can be categorizes into a higher age group.

It seems that people of every age group are interested in using voice assistants. A possible explanation for this could be the arising technology, and the advertising for smartphone assistants like Apple's Siri or Microsoft's Cortana and home devices like Amazon Echo or Google Home. The technology itself is becoming more user-friendly and reliable because one does not always need to use complicated interfaces. By simply pressing a button or using a keyword, one can already use voice assistant features. It can be concluded that one needs to address people of every age when it comes to implementing the voice assistant application. It should not be assumed that all technical terms are understandable to the user, yet every person should be able to interact with the application without prior knowledge.

2. People prefer to use voice assistants for argument search at home instead of using them in public or at work.

3. People prefer to use voice assistants for argument search when they are alone.

The second and third hypotheses are about the situations in which people would like to use voice assistants. Results show that the participants prefer using them at home and only with other people around them if they are familiar. Figure 3.1 shows an overview of the acceptance of all situational cases. In the Appendix are the descriptive statistics in table A.1 and the significance values of all dataset pairs in table A.2. The participants could rate on a scale from 1 with "convenient" to 4 with "inconceivable".

First, the "home" dedicated cases, which have an average score of 1.55 (sd=0.64) alone and 1.55 (sd=0.59) with a friend, have a significantly higher preference than

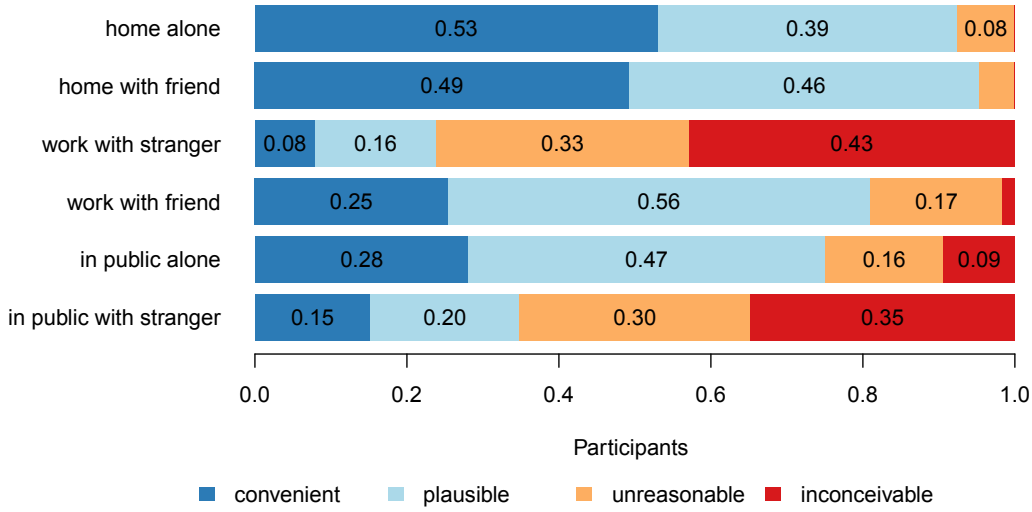

Figure 3.1: Convenience of using voice assistants for argument search for different situations.

their “public” or “work” counterparts, which supports the second hypothesis. Additionally, it confirms the results of Efthymiou and Halvey (2016). While “public” only has an average score of 2.06 (sd=0.91) with “alone” and 2.85 (sd=1.07) with “stranger”, “work” at least gains an average score of 1.95 (sd=0.71) with a “friend” and 3.11 (sd=0.95) with “stranger”. An explanation for this can be that many people are organizing their workspace as second home such that they feel comfortable enough to use their voice assistant there too.

However, there is no good evidence that supports the third hypothesis. Although the results show that there is a high gap in acceptance between the mean score of “alone” at “home” with 1.55 (sd=0.64) and in “public” with 1.95 (sd=0.71) compared to using the voice assistants in front of a “stranger” at “work” with 3.11 (sd=0.95) and in “public” with 2.85 (sd=1.07), the difference between “alone” and with “friends” cases are not always significantly different. The cases “home alone” with 1.55 (sd=0.64) and “home friend” with 1.55 (sd=0.59) show no significant inequality with a value of 0.861 and also the cases “public alone” with 2.06 (sd=0.91) and “work friend” with 1.95 (sd=0.71) are quite similar and have a bad inequality significance value of 0.721 with the Wilcoxon signed-rank test. It seems the participants show no constraint to use the voice assistant when they are together with familiar people.

Although, the introduction on the second page of the Pre-Study clarified that the application does not collect or request any private data, people still feel reserved about using a voice assistant in public or among strangers. A possible explanation

for this could be that people are ashamed of requesting specific arguments in public, or that they do not want to use their voice because it is still awkward to speak with the device when one is not alone and people around are listening. Regarding the implementation of the voice assistant, there should be an additional option to formulate the requests or get the results in form of text, so people can decide by themselves which is more appropriate in the respective situation.

4. People would like to use a voice assistant for argument search more for convincing other people than for making their own decisions.

5. Entertainment and fun is an important aspect to use the voice assistant.

The fourth and fifth hypotheses are dealing with the different motivations of using a voice assistant for argument search. The results show that the motivation "decision making" depends on the decision itself and with that can be rated worse than the "convince somebody" cases, while the "having fun" cases always do better than the other motivations. Figure 3.2 shows an overview of the acceptance of all motivational cases. In the Appendix are the descriptive statistics in table A.3 and the significance values of all dataset pairs in table A.4. The participants could rate again on a scale from 1 with "convenient" to 4 with "inconceivable".

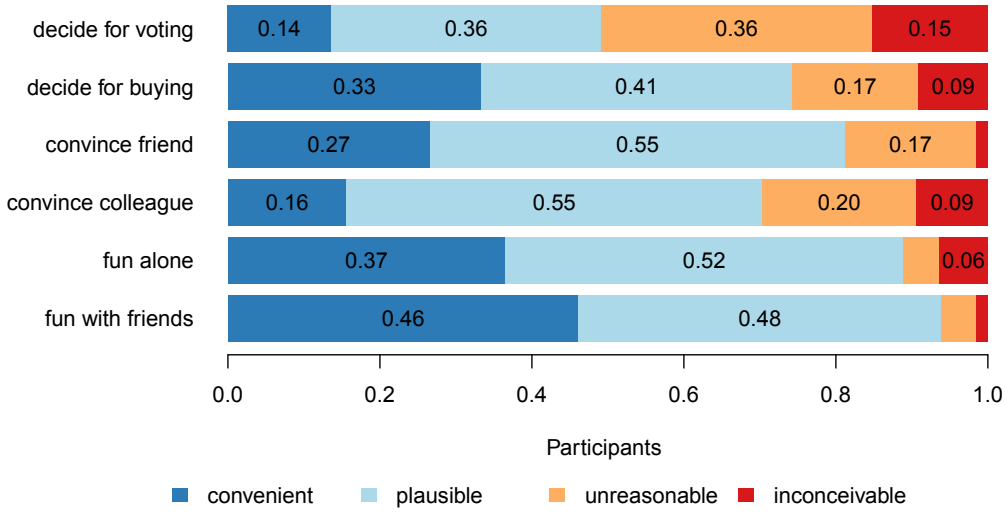

Figure 3.2: Acceptance of using voice assistants for argument search for different motivations.

For the fourth hypothesis, the results of the "convince somebody" cases were compared with the results of the "decision making" cases. While making a buying

decision has a mean acceptance value of 2.02 (sd=0.94), the acceptance of a voting decision is rated worse with a mean value of 2.53 (sd=0.92) and both datasets are unequal with a significance level of $p < 0.001$ and an effect size of Cohen's $r = 0.31$. This can be explained by the fact that a political voting decision has a bigger impact on the society and can not be compared to making a buying decision, where the impact affects only oneself and in case of doubt can be cancelled. For convincing, the case with "friend" has an acceptance value of 1.94 (sd=0.71) while with colleague has a slightly worse value of 2.23 (sd=0.83). Accordingly, a small gap is visible here which is not significantly unequal with a value of 0.007 by testing. It seems the participants like to use the voice assistant more in the presence of a friend because they can argue better with a friend than with a co-worker, towards whom the level of trust or the adequacy at work is not so high. The comparison of all four datasets shows that only the "voting decision" and "convince friend" cases have a strong inequality significance with a p-value of less than 0.001, while all the other pairs between decision and convincing cases are more equal. It seems that both kinds of motivations are not disparate in their acceptance, but it depends on the topic of the argument search how likely people would use the system. The hypothesis could not be proven true but the factor of high-stakes could be revealed.

The fifth hypothesis claims that people would also use a voice assistant to simply entertain themselves without a serious ulterior motive. Two cases were arranged with fun motivations for searching arguments. The results show that these cases, which have a mean acceptance score of 1.81 (sd=0.80) when the user is alone and 1.62 (sd=0.65) when he or she is with a friend, have a significant better score compared to their serious counterparts. In 5 of 8 cases the "fun" motivation cases have a significant better value of less than 0.05 than the other motivations in comparison. A possible explanation for this could be that people like to try out the limits of the voice application or simply want to be surprised about a topic, where no useful arguments can be expected. Similarities can be drawn to the website debate.org, which introduced the category "funny" to their debates[3]. Therefore, one should presume that people not only like to discuss about serious topics but also about nonsensical topics. The statements from the entertaining topics, even if they are not fulfilling the qualities of a real argument (because they are simply not true and have no valid premises), should not be discarded. Maybe, it is useful to separate them from the serious statements, but they still offer an amusement value, which can be seen in the results. Another explanation could be that the participants do not see the argument search application as a serious way to search for facts because they do not expect good results out of it. That would possibly not be a good sign because it could be an indicator for people not trusting the

[3]http://www.debate.org/opinions/funny/, accessed 13.09.2018

application and its main functionalities.

6. People are interested in the source of arguments.
7. People want to get involved in the application, for example by adding new arguments or rating already existing ones.

The last two hypotheses consider specific features which should be included in the application of an argument search system. The results show that users are highly interested in the source of the arguments but do not appreciate features where they get involved into the system. Figure 3.3 shows the appreciation of various possible features and figure 3.4 shows the ranking criteria. The Appendix contains the descriptive statistics in table A.5 and A.6 for the features and in table A.8 for the rankings. The significance values of all dataset pairs for the features are in table A.7 and for the rankings in table A.9. The participants could rate the features on a scale from 1 with "much appreciated" to 4 with "useless" and the ranking criteria from 1 with "most important" to 6 with "least important".

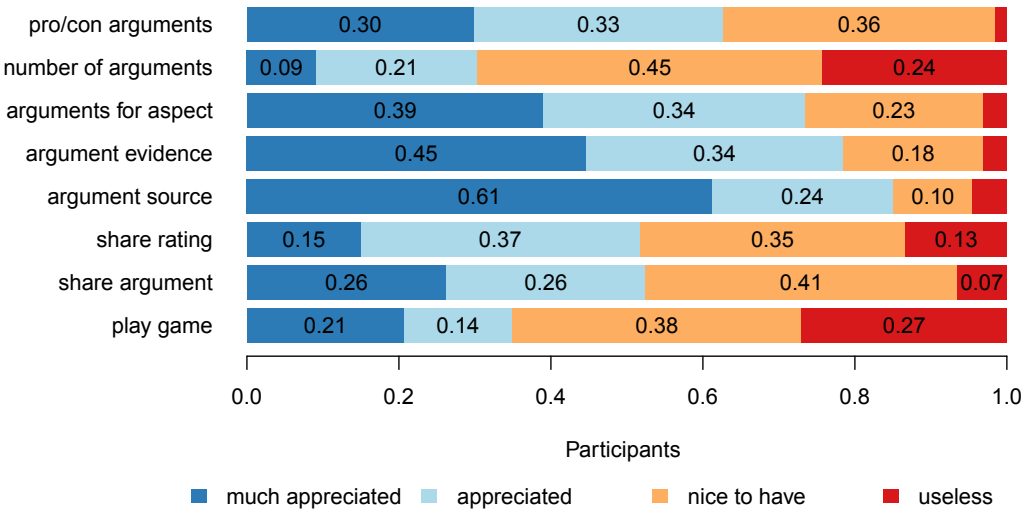

Figure 3.3: Appreciation for specific features of an argument search system.

Hypothesis six states that people are interested in the source of arguments. The appreciation results of question 5 about getting the source of an argument and the following source ranking criteria from question 10 of the feature page give evidence supporting this. In both cases, the mean ratings for including source related features is 1.58 (sd=0.86) in the feature ratings, and 1.53 (sd=0.93) in the ranking ratings, which is in comparison to all other features and criteria the

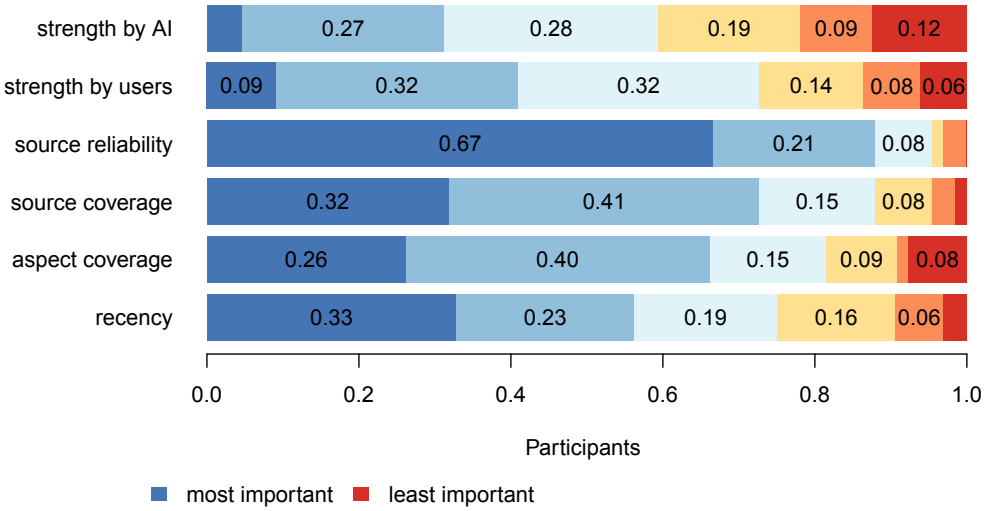

Figure 3.4: Rating of ranking criteria of an argument search system.

best value. The Wilcoxon signed-rank tests in the feature ratings show that the "argument source" is rated significantly better in 6 of 8 cases with a p-value of less than 0.001 and in 2 of 8 cases with less than 0.05 as significance value. In the ratings for the ranking criteria the "source reliability" is significantly better than any other criteria. Here, it can be seen that the majority of the participants have the strongest interest in the source of the arguments and its reliability. "Fake News" are a popular topic these days (Mustafaraj and Metaxas, 2017), and when the source of an argument cannot be verified, it seems that most of the participants do not rate it trustworthy. The feature of getting the source of an argument and ranking the argument result list by the source reliability should have an important focus in the implementation of the actual application.

The seventh hypothesis about contributing or rating of arguments for the system should be rejected. Question 3.6. dealing with "share rating" and question 3.7. dealing with "share arguments" to a topic achieved a result of 2.47 (sd=0.91) and 2.28 (sd=0.93), respectively. With these results, it can be seen that most of the participants think that these features are nice-to-have and some also appreciate them, but in comparison to the other features, they come in last. Furthermore, it can be seen in question 3.8. about "play game", which has a score of 2.7 (sd=1.08), and in the "strength by user" criteria from question 3.10., which has a rating of 2.97 (sd=1.30), that user integration does not have the same high priority like the other features or ranking criteria. The results show that people mostly want a good presentation and navigation through the argument search system but do not place much importance into contributing, interacting or challenging the system.

These features are not considered to be desirable for the system in the first place and should not get a high priority in the implementation of the voice application.

3.4.2 Qualitative Data

The survey gave the participants the opportunity to write down comments at specific parts in the study which are analyzed as qualitative data.

First, the participants could comment on why they are not interested in voice assistants if they selected this option. In total, eight people wrote a reasoning. The reason which was mentioned the most with three times was the concern about data privacy. The participants do not want to reveal their data with accounts to the companies and are afraid they are always wire-tapped by the device. This anxiety is not arbitrary. The problem is that it is also difficult to solve this issue at this time. Voice assistants mostly need an invocation phrase to start their work, but for this they also need to record their surrounding all the time. Big companies can use this data and are also justified to require accounts for their services because voice recognition and the market for applications is an expensive task. Another fact that was mentioned two times is the bad recognition of voices itself, especially when the voice has an accent. One of the biggest reasons why people stop working with voice assistants is the bad recognition of intents (Luger and Sellen (2016), Myers et al. (2018)). This problem can only be solved over time when technology becomes better in natural language processing. Two further comments referred to the missing added value of using such systems. Some people believe that they can do everything faster and more reliable with their traditional interfaces and are skeptical towards new technologies. It is hard to convince them of new gadgets that try to break new grounds. Other comments also included the slow processing of voice assistants compared to usage of keyboards and displays. At the very least, new insights in this domain show that voice recognition can beat keyboard on mobile devices in speed and accuracy of producing texts[4].

Question 9 of page 3 about possible functionalities of an argument search system offered the participants the chance to suggest other features they think were useful. In total, 13 people gave comments through the text field. A wish for a feature that was mentioned very often with a total of six times was the ordering of arguments by specific criteria, e.g. the political stance, statements of a favorite person, authors, or other sources. One participant in particular stated, that it

[4]https://www.npr.org/sections/alltechconsidered/2016/08/24/491156218/voice-recognition-software-finally-beats-humans-at-typing-study-finds?t=1537199386631&t=1537217320453, accessed 17.09.2018

would be useful to have an overview of categories or keywords that are connected to the controversial topic. The second study of this thesis tries to implement a guideline with categories in the search for arguments to help the user to navigate to specific arguments. Another often mentioned feature with three mentions in total was the request of facts, e.g. in form of reviews or prices of products or the explanation of a technical term. It is questionable if this really is a task for an argument search system or if the voice assistant should have this feature as a general functionality by its self. However, the results of the second study show that this is no trivial problem. Other less mentioned wishes were a feature for fact checking (which is quite similar to the request of evidence and source), the possibility to create own personal argument lists that are not only originated from one topic, the filtering of arguments that have less meaning in the local environment, and the additional output of arguments on a display besides the voice assistants. All those mentioned features can be considered to be useful for a voice assistant with argument search system but are too labor-intensive to include them in a first prototype. However, the conclusion of this thesis in chapter 5 will address these issues again.

Chapter 4

The 2nd Study

The second study of this thesis is about the implementation of a voice based argument search system mock-up and the evaluation of it. As mentioned previously in chapter 2 dealing with the related work, technology is not advanced enough to either have a voice assistant without natural language processing errors, nor is the args.me argument search engine reliable enough to deliver always the best arguments to every topic. Therefore, this study is conducted with a Wizard of Oz experimental set-up to test a system how it would work when most of the technological limits have been overcome. A real person, referred to as "agent", held a conversation with the participant, referred to as "user", by using prepared request and answer sheets. Requests from the user can be interpreted faster and are less error prone than in a normal voice recognition system, and the agent can access prepared data for the topics to deliver the most accurate and inquired results. Similar to the last study, some hypotheses are proposed which are evaluated with the results of this study.

Hypotheses:

1. When users want to convince somebody they are focused on getting arguments for their side.

It is assumed that there are only two serious motivations to use an argument search engine: to make a decision for oneself or to convince others that someone's opinion is valid. These motivations were taken from the first study and the results showed that they are legitimate reasons to search for arguments. Despite the fact that the user maybe formulates a request which includes this motivation explicitly, can the system recognize the motivation simply based on the arguments the

user wants to hear? The system can adjust its behaviour accordingly to the user's motivation to deliver more accurate results. It is assumed that users, who want to convince somebody, only want to hear arguments for the side they are pleading for. They do not want to hear the arguments which could make their statements futile.

2. Users felt the argument presentation was better organized when they had a category-guideline at the beginning.

This study has many similarities to the research of Dubiel et al. (2018). They compared a state-of-the-art voice agent with a futuristic one, which supports all kinds of natural language processing, and found out that the latter one is almost superior in every aspect. For this study, only the "superior" model was used, but with two different behaviors. In the first case, the user is completely free to browse through the arguments by him or herself and gets presented the arguments one after another without category-guideline. In the second case, the agent presents the user a list of categories based on the topic (e.g. economy, culture or aspects about morality) he or she can choose from. With this pre-selection, the user receives only the arguments he or she is interested in and has an overview and idea about the contents of this topic. This and similar search design recommendations can be found in the research of Hearst (2006) for creating a user search interface with faceted search. The wish for keyword suggestions and an overview of the topic was also often requested in the first study in the sub-section 3.4.2. The results at the end should show which behavior of the agent was more appreciated.

3. There is a correlation between satisfaction and expectation of the system.

The third hypothesis is originated from Luger and Sellen (2016) and Myers et al. (2018). Users feel most uncomfortable if their mental picture of a system does not match with the response of it. The participants will be asked how satisfied they were when they searched for arguments and if the system behaved as expected. Both results should correlate with each other.

4. Users think argument search with keyboard is still more efficient than with voice assistant.

Although the new argument search interface with voice assistant should make the usage more convenient for many people, the lack of visualization and navigation

tools will have a notable impact on the application. For the last hypothesis the qualitative data from every user is collected to receive information on how the interface was accepted, and the users get the chance to comment how they would improve the system.

The following sections explain the structure of this study and why specific variables were chosen, followed by the description of the process and the evaluation of the results.

4.1 Experimental Set-Up

This subchapter describes the structure of this study. The main parts of the study were the consent form with the demographic questions at the beginning, an introduction to the tasks, the tasks themselves and the evaluation sheets. The following subchapters 4.2 describe how the experimental-setup was refined and 4.3 how it was conducted. First a few words about the development of this study.

4.1.1 Development

The study was planed in an early alpha version for a pilot and a later beta version, which was used to conduct the study. In the alpha version a scenario was used which is not fully in line with a real Wizard of Oz experimental set-up. Although the agent was a real person who tried to behave like a voice assistant, both agent and user were in the same room. This had a bad influence on how the user behaves in front of the agent, because he or she knew it was a real person on the first sight. Two variables were defined for the topics: the motivation and the impact. The motivations were similar to the first study divided in *Decision Making, Convince Somebody* and *Entertainment*. For the impact, low- and high-stake topics were considered. The differentiation in these two categories should show how cautious the participants would use the system when there is less or more on the line. Table B.1 in the Appendix shows a distribution with some sample topics.

Participants had to select two of three topics of a category which makes in total eight tasks to search for arguments. This distribution of topics had some flaws in the experimental set-up. First, eight search tasks were too big to let one participant do them all one after another. Early tests showed that one task including the evaluation sheet takes around five to ten minutes. One participant would not only need around 90 minutes for the whole study, he or she would already have

signs of fatigue after a couple of tasks. The second issue was the motivation category "Entertainment". It is not possible to have high-stake topics here, when only having fun is the background of each scenario. Although entertainment is an important aspect (Luger and Sellen, 2016), it can not be compared to the other two motivation categories. A further experiment needs to be conducted to get insights on how playful interactions need to be handled, which is outside of the scope of this thesis. The third problem was the separation in low- and high-stake itself which is a highly subjective matter. To name a few examples: atheists might show less interest to inform themselves about the existence of god, female participants show more respect to the topic of legalizing abortion, a Linux operation system user would not care if Microsoft Windows or Apple Macintosh is better, or a vegetarian has no high-stake in the topic about stopping the consumption of animal meat because they already practice this lifestyle. On a final note, many of the listed topics in table B.1 have the problem that not many arguments exist to support them, or a majority of that arguments can be considered as opinions instead of real arguments because their statements can not be verified. This is surely a typical problem in argumentation mining, but the arguments presented in this study should be comprehensible and not speculations to make the user less confused.

In the end the presented scheme of variables and controversial topics was discarded and it was decided to use less variables and focus more on the behaviour of the system. The following subchapters describe the components of the experimental set-up.

4.1.2 Consent Form and Demographic Questions

At the beginning of the study, the participants received a paper with the consent form and some demographic questions. The consent form included subjects describing the recording of the conversation between the user and the agent and the permission to publish whole or parts of the transcripts of the recording in scientific publications. The participants stay anonymous and only the research team has access to all collected data. After signing this form, the participant had to fill out some demographic questions, which were pared-down compared to the former study.

1. What is your gender?
2. How old are you?

3. How often do you use voice assistants?
4. If you use a voice assistant, for what tasks do you use it?
5. How would you rate your English level?

The first two questions gather general information from the participants in the form of their gender divided in *female*, *male* or *others* and their age in the ranges of *17 or younger*, *18 to 30*, *31 to 49*, *50 to 64* and *65 or older*. The same options were used in the first study. In this study, there is again a higher number of male participants expected and in average a bigger user group in the age of 18 to 30.

The next question should classify how experienced the users already are with voice assistants. They could tick *Frequently*, *Rarely* or *Never* with the opportunity to state some application they used with such a system, in case of not ticking the last answer in the former question. Users with more experience in the usage of voice assistants have maybe a better mental image of the system and can formulate more direct requests to get the desired information.

The last question is about a self-rating of the English skills. To ensure that the participant can accomplish the tasks of the study, he or she should have an English level of *Proficient* or at least *Intermediate*. Participants who selected *Beginner* are mostly not suitable to hold a conversation with the agent and collected data would be not representative for this study. It is assumed that users speak in their first or advanced foreign language when they talk with a voice assistant. Because this study is designed with an English voice interface without any other additional languages, the speaking of that language is mandatory and potential participants who state that they are only on a *Beginner* level are not allowed to take part in the study.

4.1.3 Instructions

The introduction was the second sheet of paper which was handed over to the participants. It started with a short description of the study and stated that he or she has to select four of total six tasks and complete a questionnaire after each task.

Next, the participants received an overview of the features of the voice assistant which were available in this experiment:

- Get pro- and con-arguments on some topic
- Get arguments on a topic that relate to a specific keyword
- Get the total number of arguments on some topic
- Get argument categories for some topic
- Get an explanation or evidence for an argument already heard
- Get the source of an argument already heard
- Repeat what Alexa just said
- Say ”Stop” if you want to interrupt Alexa for a new command
- Say ”Close” to finish the current argument search of the topic
- Say ”Help” to get a list of possible commands

The first four options are the main features. Their main purpose is to get arguments and to navigate through them. The next two options deliver insights to the arguments in form of fictional statistics or publications. However, it should not be possible for the participants to distinguish between real and fictional evidence or sources. The last four provided requests are general commands that can be used in nearly every voice application.

In this context, it should also be mentioned that for this study it was decided to take “Alexa” as the default invocation name for the agent. Sales numbers from the third quarter of 2017 show that Amazon Echo devices with Alexa voice assistant had the biggest market share [1]. Although, shares are shifting more and more to Google devices recently because of the bigger influence in the last years, it is assumed that people are more used to the invocation name “Alexa”. Nevertheless, participants were free to talk with the agent with any invocation name they like and the features stated above only used the name “Alexa” as example.

Two further paragraphs on the introduction sheet stated that one task should not take longer than 5 to 10 minutes or it will eventually be aborted, and that some facts in this study are made-up and should not be relied on if a participant really wants to make a decision because of them. This is mostly referring to the evidence and sources of arguments, which were made-up on the fly by the agent. The text does not mention that these two features are the fake ones, so the participants maybe still have the demand to use them without knowing that they are not real.

After reading the sheet of paper the participants could ask questions related to the study. Afterwards, he or she had to return the instruction sheet back to the study organizer. The intention behind this was that the user does not simply

[1] https://ethority.de/2018/04/05/sind-die-verkaufszahlen-marktanteile-von-alexa-google-home-der-smartspeaker-markt-zahlen/, accessed 09.09.2018

execute all possible requests one after another but talks with the agent in a natural way. The list of possible requests only serves as an overview and an analysis of the actions in the evaluation subchapter 4.4 should show which features were used most frequently.

4.1.4 Tasks

The tasks were organized by the variables "motivation" and "category-guideline". After reading and understanding the task explanation in the instruction sheet, the user had to choose topics. From over 15 possible topics of the alpha version of the study (presented in subchapter 4.1.1), only 6 remained, which were formulated based on the motivation variable. Afterwards, they were rephrased for "making a decision" or "convincing somebody". Table 4.1 shows the selected topics with their version based on the motivation variable respectively.

topic	making a decision	convince somebody
A	Decide whether to buy an electric car	Convince your friend to buy an electric car
B	Decide whether to visit the Zoo	Convince your friend not to visit the Zoo
C	Decide whether to study abroad	Convince your friend to study abroad
D	Decide whether to stop eating meat	Convince your friend to stop eating meat
E	Decide whether conscious general AI should get fundamental rights	Convince your friend to support fundamental rights for conscious general AI
F	Decide whether to introduce a school uniform	Convince your friend to support a school uniform

Table 4.1: Topics of the second study formulated depending on the motivation variable

The participants got in total six slips of paper and on each was a topic with a background story to contextualize the tasks of searching for arguments. For example, table 4.2 shows a comparison of the motivational background stories to the topic "Zoos should be forbidden" next to each other. In the left case, the participant had to make a decision for him or herself, while on the right case, the participant had to convince a friend about the topic. Table B.15 in the Appendix shows all background stories to the "making a decision" cases and table B.16 to the "convince somebody" cases. The motivation "entertainment" was not considered in this study because of the lack of comparison to the other two possible motivations. In the first study of this thesis and in the research of Luger and Sellen, it was already perceived that entertainment is an important part of a

Decide whether to visit the Zoo.	Convince your friend not to visit the Zoo.
You want to go to the Zoo at the weekend to see wild animals in real life and to learn more about them. However, you saw recent protests in your town that convinced you that animal confinement is bad. You struggle if it is okay to visit and thus support such a facility and want to come to an informed decision in this regard.	Your friend wants to go to the Zoo at the weekend to see wild animals in real life and to learn more about them. However, you saw recent protests in your town that convinced you that animal confinement is bad. You therefore want to convince your friend it is not okay to visit and thus support such a facility. You are now looking for arguments that help you convince your friend.

Table 4.2: Examples for the topic "Zoos should be forbidden" with two different motivation background stories.

voice application to gain more experience with its usage. To distinguish between entertainment and serious motivations, a new study has to be designed, which extracts from different tasks how users learn about voice interfaces in the best way. This would be too labour-intensive to include it in this thesis but can be addressed in future work.

The variable "category-guideline" is originated from the question how people would like to interact with the voice assistant. Do they want to explore the interface completely by themselves with a natural conversation, or do they like to get an overview of the categories which moves them from one argument category to the next. Every participant was confronted with both modes and had to rate them afterwards. The behavior of the agent in these modes is described in subchapter 4.1.6 and the distribution of the variables and tasks in subchapter 4.1.

4.1.5 Questionnaires

After each task the participants received a questionnaire with eight statements they had to rate for themselves:

1. I was already well-informed about the topic
2. Alexa was helpful for the task
3. Using Alexa for the task was fast
4. Alexa was pleasant to use

5. Alexa behaved as I expected
6. The conversation with Alexa felt natural
7. Alexa's answers were well-structured
8. I would recommend Alexa, like I could use her for this task, to others

Every statement could be rated with a 5-point Likert scale and a "Don't know"-field, which was excluded in the evaluation.

The background to the first statement was to see if people, who already had a high knowledge about the arguments of a topic, rate the experience with Alexa worse because they expected a better presentation of the information. The other statements should give insights in respect to the helpfulness, efficiency, kindness, structure, naturalness and expectation of the system. The evaluation should show if there are any correlations between this dimensions. Luger and Sellen (2016) pointed out that a wrong expectation of the system can have a negative influence on all the other categories. The last statement is a final rating if the system is good enough to be recommended to other people. If this rating is unaffected by the former questions, the system could still be very recommendable because it is a novel approach to solve a problem which had no good solution until now.

There was a second additional questionnaire asking if the participant solved a task with the category-guideline. It recapitulates as a reminder to all categories of the last topic and contains the following statements:

(I selected those categories because...)

1. of their relevance to the task
2. of my interest in them
3. I felt I needed more or fewer arguments
4. I did not understand the name of the category

Again, every statement could be rated with a 5-point Likert scale and a "Don't know"-field.

The first statement should show how people get affected by the task and their background story when selecting the categories. If they give a bad rating on this point, perhaps they simply want a presentation of all arguments despite the

fact they were only considered one aspect in the background story. The second statement could be a reason for exploring many arguments, because they were interested in the whole argumentation. It can happen that the participants feel forced to select a category because they already dismissed nearly all of them in the category list. If that is the case, they can rate this in the third statement. The last point is about the understanding of the category titles. Maybe, they got a bad phrasing or need more than common knowledge to comprehend them.

At the end followed a post study questionnaire. It included only two questions:

1. Would you prefer argument search with keyboard input over a spoken one as in the study? Why?
2. How would you improve the Alexa system that was used in the study?

Both questions are of qualitative nature, so the participants could not rate them with a scale but had to answer in a few sentences.

The first question reflects one of the core questions of this thesis and the results provide insights to the third hypothesis of this study. After solving different tasks with the new system, participants could state which interface they prefer and give a reasoning to it. The statements give hints about for example efficiency, control, or memory problems.

The follow-up question is based on the first one. If the new system has problems compared to the traditional one, or if it has some flaws in a voice interface design itself, what would the participants suggest to fix them? Answers to this question could give good insights on how to implement a new prototype in future work.

Each of the previously mentioned questionnaires also had a comment part where the participants could write comments or explanations about facts they could not state in the questions.

4.1.6 Behavior of the agent

This study was conducted with an agent in a Wizard of Oz experiment. Similar to a server request-response architecture, there was a need to define some rules on how the agent has to behave in different situations. In the following, the rules are presented with a small caption and a description text.

Greeting and Farewell If the user only states a greeting or a farewell, the agent also responds in the same way without asking for further information.

Inquire If the user did not mention a topic or there is missing information (e.g. the user wants to hear about the source but did not state to which argument), the agent inquires them from him or her.

Suggestion If the user makes a longer pause, the agent asks the user what to do next or suggest other reasonable actions (e.g. list the con arguments after the pro ones or giving evidence if the user seems to be thoughtful).

Open without Categories If the user opens a topic and it is the mode without "category-guideline", the agent states the total number of all pro and con arguments and asks the user of which side he or she wants to hear arguments first.

Open with Categories If the user opens a topic and it is the mode with category-guideline, the agent first says the total number of categories and then starts to ask the user if he or she wants arguments to each category. After filtering out deselected categories, the agent starts with the first selected category.

List Arguments The agents reads up to three arguments of a list. If the list has more arguments left, the agent asks to continue the list.

End of List If there are no arguments of one side of the whole topic or a category left, the agent asks if now the other side should be listed. If both sides were already listed, the agent states that there are no arguments left for this category or topic. In case of the end of a category, the agent goes to the next category.

Open Category If a new category is opened, the agent first states the total number of pro and con arguments in the category and asks the user which side to present first.

Read Evidence If evidence is requested by the user, the agent states fake facts which are originated from predefined templates.

Read Source If the source is requested by the user, the agent states fitting or fictional websites the arguments could be originated from.

Based on the name conventions of Azzopardi et al. (2018), the actions of the agent were summarized into four classes. The descriptions of the Conversational Actions are shown in table B.2, of the Navigate Actions in table B.3, of the Inquire Actions in table B.4 and of the Reveal Actions in table B.5. The name of the actions in form of tags were also used in the transcription of the audio recordings in the evaluation part of this study.

4.1.7 Structure

This part is about the integration of the variables “motivation” and “category-guideline” plus the questionnaires. Figure 4.1 shows an overview of the structure of the study. Participants were divided in two groups: A and B. These groups are distinguished by the order of the motivation. While group A starts with “Making a decision” in the first part and “Convince somebody” in the second part, group B has this motivations contrariwise but is picking the topics for the tasks in the same order. For the first part, both groups can select the topics A, B and C and for the second part the topics D, E and F. This strategy is to make sure that the same topics will be chosen for both motivations and can be compared in the results. The description of the tasks is written in section 4.1.4.

The participants always had the chance to discard one of three topic for the tasks. The first study already showed that some people feel uncomfortable with specific topics. For this, the study was designed to give the participant at least one choice to discard a topic they do not like but to have two of three topics with enough intersection to get meaningful results at the end.

For the second variable “category-guideline” the first task in a part was without the help of a category list, while the second part was always with the help. The participants should first experience the system when they are totally free to ask whatever they want and then compare it to the second mode when they have to first select categories to narrow their desired arguments.

After each task the participant got a questionnaire to evaluate their acceptance of the system. A second questionnaire was forwarded if the task included the “category-guideline” and the participants got a final questionnaire at the end of the study. The content of the questionnaires are described in the section 4.1.5.

4.1.8 Rooms and Hardware

This section describes the locations and machines which were used to conduct the second study. The Wizard of Oz experiment required to have two separated rooms for the user and the agent. For the set-up, two rooms were chosen from the university building in the Karl-Haußknecht-Straße 7 in Weimar. The user was asked to take a seat on a sofa in one of the computer laboratories (see picture (a) of figure 4.2). The sofa was comfortable to sit on and the voice interface was laid down on the armrest of the chair. The interface consisted of a *Lenovo Netbook* and a *Jabra SPEAK 410*. The user could speak with the agent through the voice

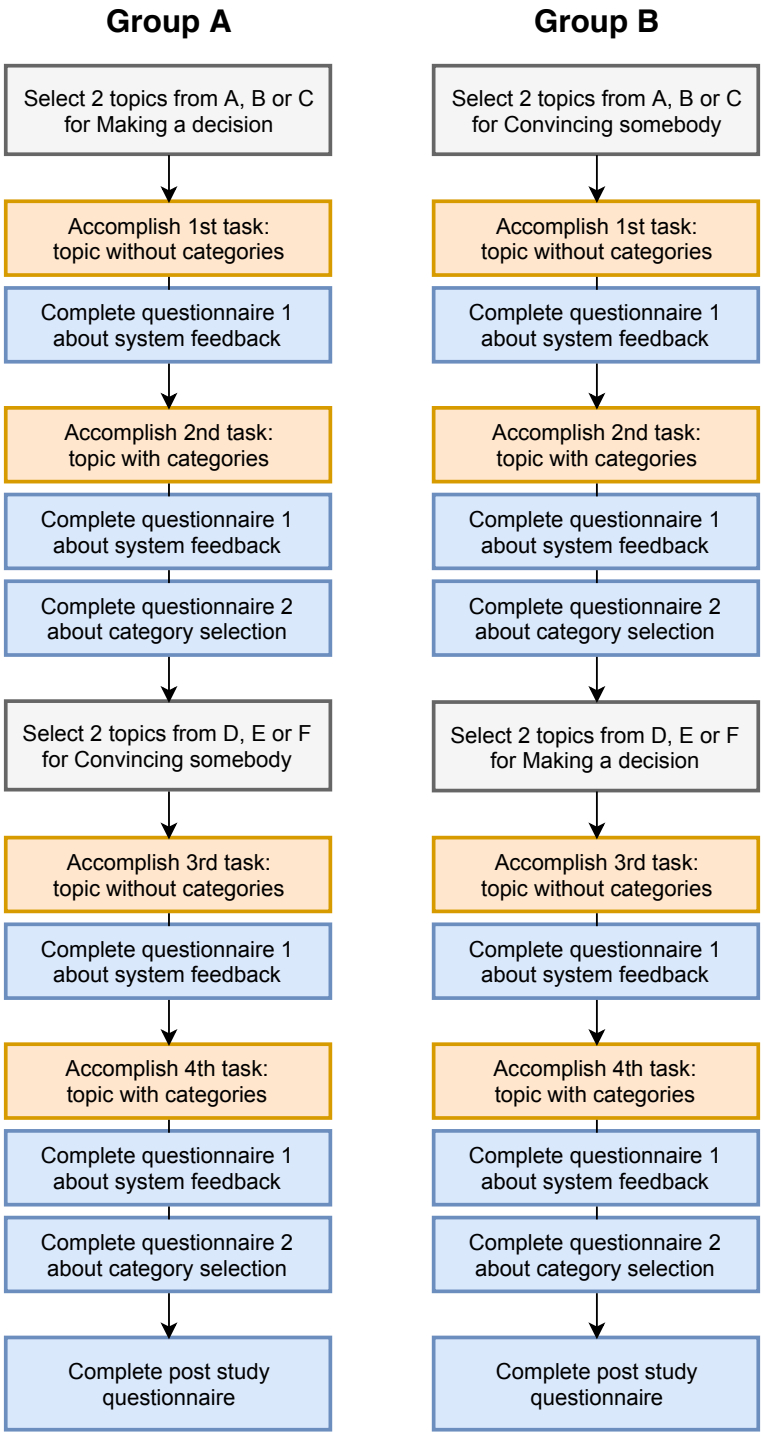

Figure 4.1: Structure of the second study split between two Groups A and B. Selections are coloured in grey, tasks in yellow and questionnaires in blue.

(a)

(b)

Figure 4.2: (a) shows the room with the voice interface the participant could use and (b) the workstation of the agent.

interface without using a keyword for invocation because the channel was always recording on both sides. However, the laptop screen was turned off so that the user could not see the software which was used to conduct the study.

The workplace of the human agent was positioned in a lobby before the laboratory (see picture b of figure 4.2). He sat at the workstation with scripts about how to answer different requests by the user. A *Samson Go Mic Clip-On USB microphone* was used to record the voice of the agent whenever he pushes a button. Furthermore, the agent had a headset so that only she could listen to the voice of the user. The set-up was designed in this way so that no additional noise was recorded on the agent side and that in problematic situations the agent could seek advice from the lead researcher without the user noticing it.

4.1.9 Software

This section is about the software which was used during the second study and in the post-processing of the data.

For having a communication channel between the user and the agent, the voice-over-Internet Protocol application *Discord*[2] was used. This freeware features a good sound quality with some settings in the recordings and the ability to run on different systems. This was important because the netbook on the user side only could use *Microsoft Windows 7*, while the workstation on the agent side had

[2]https://discordapp.com/, accessed 15.09.2018

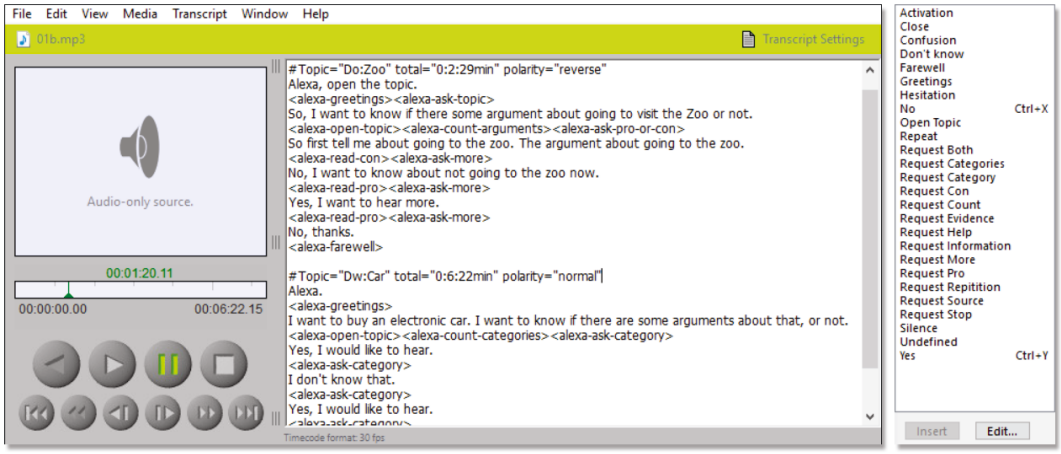

Figure 4.3: Software *InqScribe* with side panel of reoccurring tags.

Ubuntu 16.04 as operation system.

For the recording of the voices two methods were used. On the agent side the free software *Audio Recorder* [3] for Ubuntu was used to record multiple audio sources from the microphone and Discord application at the same time. Although this method was used to get a recording with the best quality directly over the agent microphone and the Discord output channel, a second audio record method was used in case the software or one of the gadgets fails or the audio channels interfere. A smartphone with a simple dictating machine application was put in the room of the user to have backup recording.

Transcribing the conversations between user and agent was done with the free version of the software *InqScribe* from Inquirium [4] (see figure 4.3). Each audio recording of the conversation was split up into the tasks and transcribed in text format with tags for the agent (see section 4.1.6 and the user (see section 4.4). The possibility to fast-forward or rewind the audio file with shortcuts and define often reoccurring tags with buttons in a side panel, made this software very convenient to use.

[3] https://launchpad.net/audio-recorder, accessed 09.09.2018
[4] https://www.inqscribe.com/, accessed 09.09.2018

4.2 Pilot

For the pilot study, a complete walkthrough of the second study was conducted with a participant who was excluded from the later iterations. Also, the data collected in this pilot was not integrated in the data of the real study.

The pilot study took around 45 minutes but the expected time was around 30 minutes because some questionnaires, tasks and other notes were not optimally organized. Through the process, the participant could identify a few typing errors and could ask for a clipboard because completing a questionnaire on a sofa without a pad was quite troublesome. Additionally, the sound from the speaker on the user side was a bit scratchy and on the agent-side it was quiet which was later adjusted in the settings of Discord. Further on, the participant was a bit confused about the arguments and the stance towards a topic, had problems to address specific arguments and did not like listening to categories at the beginning of a task when opening a topic. Despite the comments it was decided not to change the behavior of the agent because the mentioned problems were crucial aspects of the system which are open for analysis. Maybe, other participants have no problems with this behavior of the agent and only in this instance there was an exception.

Nevertheless, the participant enjoyed the study on a comfortable sofa and also liked the range and variety of the topics. In fact, they would still prefer the argument search with keyboard and screen because of navigation issues and faster typing.

4.3 Process

This subchapter focuses on the invitation and conduct of the second study. The study was executed without bigger problems, but a few changes had to be done during the process to adjust towards a proper outcome of the results.

First, an invitation email was written. It was send to the mailing lists of the bachelor and master courses of Computer Science and Digital Media, the master course of Human-Computer Interaction and the master course of Digital Engineering , all at Bauhaus-Universität Weimar. The content included a motivation part, information about time and place, and the note that participants get free ice cream, the chance to win a 25 € Amazon gift card, and the promise to sit on a comfortable sofa during the entire study. People could follow a link to the online

calendar tool Doodle[5] to make an appointment for a study session. It was possible to get a time slot over two weeks from the 25th of June to the 8th of July. The email mentioned that the study would take around 30 minutes and an one hour time slot could be reserved. For this study, it was decided that at least 10 participants with each four tasks are needed to have enough data for meaningful results. Other researches with a similar experimental set-up (Luger and Sellen (2016), Myers et al. (2018), Dubiel et al. (2018)) collected data from the same or less amount of participants. The total targeted number of participants was between 16 and 20 people to have a buffer for potentially invalid data that can not be used for the evaluation.

The process of the study itself went smoothly most of the time. In a few sessions the participants were a bit disturbed by noises outside the building or other rooms, but they were never loud enough to interrupt the study. Another small problem was the heat of the summer and the malfunction of the freezer compartment. The promised ice cream was not always available but the participants enjoyed the ice tea which was there as replacement. Small breaks between the tasks and enough to drink helped to keep up the good mood during the study.

Two matters of the structure of the study needed to be changed to get better results at the end. First, some participants were a bit puzzled when the agent behaved differently. In the even numbered tasks, the agent suddenly asked the user which categories he or she wants to hear instead of letting the participant browse freely through all the arguments. To make the user a bit more prepared for this situation, it was decided in an early stage of this study to give the user a hint before he or she starts the second task that the system behavior will now change in contrast to the first task. Another problem was, that the participant simply did not want to select categories at the beginning of the task, but the agent required it. When the user asked more than three times for other features of the system, the agent changed to the mode without category-guideline because it seems the user was not interested in this behavior anyway. This situation happened a few times and was counted as failure for this behavior of the system.

At the end, the study was conducted with 18 participants, of whom one third was female. 13 of the participants stated they are from the age group from 18 to 30 while the remaining part come from the age group 31 to 49. Nearly the half stated that they have never used any voice assistant before, seven use them rarely, and only 3 use them frequently. Typical stated use cases for voice assistants are setting of timers or alarms, information about weather, and search tasks in stores or maps. 10 people rated their English Level proficient and 8 people rated

[5] https://www.doodle.com, accessed 09.09.2018

themselves intermediate. No participants considered themselves as a beginner in English and the study showed that nobody had problems to formulate requests or to understand the agent. Only a few words inside the arguments lead to small confusion for the participants because they originated from higher scientific domains. Yet, it was always possible to ask about the meaning of any technical term.

4.4 Evaluation

This subchapter is about the insights gained from the second study and the evaluation of results with regard to the hypotheses. The data is subdivided into quantitative data, which is collected from ratings with Likert scales and measurements in the experiments, qualitative data, which was gained from comment sections and interviews with the participants, and observations, which were noted during the experiments.

Several programs were used to analyze the data. IBM's SPSS (IBM Corporation, 2013) had the purpose to organize and filter the data and to calculate the significances with Wilcoxon signed-rank tests. The same test was used by Dubiel et al. (2018) with a similar experimental set-up. The data is not normally distributed and neither are the ratings of the Likert scales metric. In conclusion, it is not possible to conduct a t-test or ANOVA to prove the data for significances. Nevertheless, the data is non-parametric, ordinal and paired in two groups which leads to the Wilcoxon signed-rank tests which has the null hypothesis that all median differences are equal. Equation 3.1 on page 21 showed how the effect size of the experiments is calculated. JASP (JASP Team, 2018) was the software to create the tables for the descriptive statistics.

4.4.1 Quantitative Data

In this study, quantitative data was collected in form of Likert scale ratings, measurements during the experiments, and by counting actions, tags and patterns in the transcripts of the interviews. First, the latter mentioned data sources are presented to get an overview of the study and insights of the first hypothesis. Then, the Likert scale ratings are analyzed to discuss the remaining hypotheses.

During the study, the time for each task was measured (see table 4.3). The fastest completion of a task was 96 seconds, while the slowest took 482 seconds. The

	making a decision		convince somebody	
	without cat.	with cat.	without cat.	with cat.
Valid	18	18	18	18
Missing	0	0	0	0
Mean	192.3	290.2	219.8	261.1
Median	175.5	278.5	190.5	238.5
Std. Deviation	68.81	72.67	85.48	88.81
Variance	4735	5281	7308	7887
Minimum	96.00	155.0	104.0	127.0
Maximum	346.0	438.0	403.0	482.0

Table 4.3: Descriptive statistics of completion time of the tasks in seconds. The tasks are subdivided if the user had to make decision or had to convince somebody and if the system offered a category-guideline or not.

mean values illustrate that the tasks that used a system with category-guideline took in the average case 41 to 98 seconds more than the tasks without category-guideline. This is explainable by the fact that the system first introduced the categories to the user before presenting arguments. Assuming that the user still wants to hear all arguments from all categories, the time needed for the category selection is a negative aspect which can be seen in the measured times. However, the results of the system feedback presented later will show that the participants did not perceive the system being slower with category-guideline.

topic	short	description	selected $\sum$
A	Car	Buying an electric car	16
B	Zoo	Zoos should be forbidden	12
C	Study	Studying abroad	8
D	Meat	Stop eating meat	11
E	A.I.	Fundamental rights for conscious AI	12
F	Uniform	Introducing school uniform	13

Table 4.4: Summary of the topics with short and long description and how often they were selected by the participants.

The next statistic is the amount of times specific topics were chosen. The participants had the chance to discard two topics over the whole study, left with four to handle for the tasks. Topic A with “Buying an electric car” was chosen 16 times, topic F with “Introducing School uniform” 13 times, topic B with “Zoos should be banned” 12 times, topic E with “Fundamental rights for conscious AI” 12 times, too, topic D with “Stop eating meat” 11 times and topic C with “Studying

abroad" only 8 times. These numbers are not surprising in consideration of the people who took part of the study. The participants got the instruction to select topics according to their interests and to prefer those they do not have much knowledge about. Many of them were students coming from abroad or they already did a semester abroad which is reason enough why topic C was selected less than any other topic. Similarly, topic D was also chosen less often than most of the other topics. Interviews during the tasks could reveal that sometimes topics were not chosen because the participant simply could not imagine him or herself in the role to make a decision, for example, to stop eating meat or to convince somebody else to stop it. The interest in topic A with 16 picks is explainable by the fact that every participant came from a facility with technical background. The topic about electric cars is current and draws interest to many people. However, it is odd why topic E, which has the same technical nature and is even more futuristic, was selected less than topic A. Participants stated here that they had problems understanding the words in the accompanying text and were afraid to get confronted with more of them when they open the task or ask for explanations. On the whole, no topic was excluded on a high rate which shows that the range of the topics for this study still was a good choice.

After completing all tests, the audio recordings were transcribed. For analysis purpose, the plaintext was later exchanged with pre-defined tags. The behavioral patterns and defined actions for the agent were already described in section 4.1.6 and the agent's action tags are shown in table B.2 to B.5 in the Appendix. During the transcription process, the user actions could be defined and are shown in the tables B.6 to B.8. Additionally, some indicators were defined for the plaintext of the transcript when the participant used any form of speech disfluency or special behavior, e.g., saying "hmm" when he or she is thinking or interrupting the agent. The full list of all defined indicators can be seen in table B.9.

Table 4.5 shows an example of a transcript. The task was to "convince somebody" and "without category-guideline". In other words, the user received a background story to convince a friend and the system did not offer any help at the beginning of the session for only selecting relevant categories. With 152 seconds the conversation was one of the shorter records between a user and the agent but it shows some insights on how the audio files were transcribed and translated into action tags. Usually, the user starts with an activation or greeting phrase followed by a request to open a topic for argument search. The agent confirms the topic and presents the total number of pro and con arguments. Subsequently, the user is asked which side of the arguments he or she wants to hear first. After a chain of requests, confirmations and presentations of arguments, the user closes the topic and both parties say farewell. This point should be accomplished when the user

#	Transcribed recording of user (white) and agent voice (gray)	Corresponding tags
1	Alexa?	<user-activate>
2	Hello.	<agent-greetings>
3	[ahh] Tell me about the [uhm] pros and cons of vegetarians.	<user-open-topic>
4	Open topic “Humas should stop eating animal meat”. There are in total 6 pro and 8 con arguments for this topic. Which one do you want to hear?	<agent-open-topic> <agent-count-arguments> <agent-ask-pro-or-con>
5	Six pro.	<user-request-pro>
6	Do you want to hear pro arguments?	<agent-ask-pro>
7	Yes.	<user-affirmation>
8	First, vegetarian diets generally lead to better health outcomes than omnivorous diets. Second, eating and producing meat heavily contributes to climate change. Third, harming animals has severe impacts on the local environment due to its heavy resource use and impact upon the landscape. Do you want to hear more pro arguments?	<agent-read-pro> <agent-ask-more>
9	Yes, I [uhm] I would.	<user-affirmation>
10	First, farming animals has severe impacts on the local environment due to its heavy resource use and impact upon the landscape. Second, Ceasing to produce meat will be necessary in order to meet the growing food needs of the world’s population. Third, the meat industry is particularly harmful...	<agent-read-pro>
11	[!] Stop, Alexa. Stop and [ahh] about cons, please. Tell me about cons of eating meat.	<user-request-stop> <user-request-con>
12	First, the taste of meat is delicious and brings many people pleasure in a manner that vegetarian food cannot fully imitate. Second, ceasing to eat meat would harm the economy. Third, The harms of meat are not inherent to all forms of animal farming, and can be ameliorated through better consumer choice and/or moderation rather than complete abstention. Do you want to hear more con arguments?	<agent-read-con> <agent-ask-more>
13	Yes, please.	<user-affirmation>
14	First, Consuming meat is an important part of many culture’s cuisines, religions, and practices. Second, Farming vegetarian products also directly or indirectly harms animals in the process of crop production and harvest. Third, as animals lack rights it is morally permissible to raise them for slaughter. Do you want to hear more con arguments?	<agent-read-con> <agent-ask-more>
15	No, close.	<user-negation> <user-close>
16	Good Bye.	<agent-farewell>
17	Bye.	<user-farewell>

Table 4.5: Example of transcript of a complete task with a conversation between user and agent. The plaintext of the user transcript in the left column is translated to user tags on the right column.

thinks he or she heard enough arguments for the task.

Some specific features can be observed in the dialogue in form of special behavior of the user. In turn three at the beginning of the conversation, the user has small problems to formulate the query to open the topic for argument search. The participant uses speech disfluencies like “ahh” and “uhm” which are non-lexical words that interrupt the flow of a request. These were quite common in many user requests and cause problems in the natural language processing. The agent can confuse them with wrong intents if it is not able to eliminate them. Table 4.6 shows a summary of the summated actions in their categories, the general number of all agent and user turns, and the number of all speech disfluencies by the users. All in all, it can be seen that on average in every third or fourth request by the user a filler word is included which can make the speech recognition for the agent difficult. Moreover, in 22 cases the agent was interrupted by the user which can also be seen in turn 11 of the example. This mostly happened when the user did not want to hear the remaining arguments from the list or already heard them before by accidentally navigating back to them. In a Wizard of Oz experiment, a real person acting as agent can handle such sudden behaviors in most of the cases, but a real system needs a good design of the states and possible intents to manage these requests. The study shows that an interruption by the user can occur in average in every fourth argument search session.

	total count
agent turns	936
user turns	956
conversational actions by agent	375
navigate actions by agent	196
inquire actions by agent	619
reveal actions by agent	618
conversational actions by user	487
navigate actions by user	203
inquire actions by user	343
speech disfluency “ahh”	133
speech disfluency “hmm”	24
speech disfluency “uhm”	121
interruption by user [!]	22
not understandable [?]	5

Table 4.6: Total number of actions and behaviors of user and agent.

The next analyzed data is about the explicit and implicit requests the participants used in the experiments. Statements can be classified as explicit, when a person formulates a sentence which can be understood without the context of the whole text, or as implicit, when information is missing and needs to be gathered from the phrases around the statement. These statements can also be found as requests between a user and an agent. Table 4.5 shows on turn 11 an explicit statement by the user, where he or she asks the agent for con arguments. On the contrary, turn 9 presents an implicit statement where the user only makes an affirmation to the agent. Turn 8 reveals that the agent read out some pro arguments and asked the user if he or she wants to hear more. Hence, the following statement by the user is an implicit request for pro arguments. Implicit statements, mostly in the form of affirmation or negation by the user, are easy to recognize by the agent, because of the manageable size of utterances, and are a typical human-computer design aspect in state-of-the-art voice assistants (Dubiel et al., 2018). The explicit statements, which can request intents outside of the current scope of the agent and are difficult to process as natural language, are problematic. Examples were shown in the last paragraph with speech disfluencies. While implicit statements are a typical problem in argument mining (Rajendran et al., 2016), explicit phrases are a bigger problem in the design of voice assistants. The issue could be managed in this study simply by conducting a Wizard of Oz experiment. Time will tell when systems are able to process explicit statements in form of complex requests without problems.

requests	explicit	implicit	more	total $\sum$
pro arguments	126	10	39	175
con arguments	62	58	25	145
evidence	20	4	0	24
source	4	4	0	8
information	65	0	0	65

Table 4.7: Total number of requests which were formulated explicit with a full phrase, implicit as “yes” or “no” answer or after a more request.

Table 4.7 shows an overview of all requests by the user to get information from resources managed by the agent. The requests are categorized as “explicit” statements, “implicit” statements and “more” statements, when the user affirmed to continue a list of arguments. The “more” category can also be counted as implicit, but is separated here from the normal implicit statements because it is triggered from a different question from the agent. These different questions can be seen in turn 6 and 8 of the example in table 4.5. The summary shows a total of 175 pro arguments requested and a total of 145 con arguments. The slight differences

in total numbers and also in comparison to explicit and implicit requests, can be explained by the behavior of the the system, also explained in section 4.1.6. The agent mostly asked the user at the beginning of a session which side of the arguments should be presented first. Because most of the participants started with the pro arguments, a high number of explicit pro argument requests can be seen. After telling all arguments of the pro side of a topic, the agent asks if the user now wants to hear the con side. The higher number of implicit con arguments are a logical conclusion. Nevertheless, the difference in the increase of requests can be explained by the fact, that there were slightly more pro than con arguments in total when all topics are summed up. When for example the pro side has 4 arguments in a category and the con side only 3, then the agent will ask for the pro side always one time for more, even when there is only one argument left. This one more request is missing on the con side, because the agent will read all 3 arguments in one turn.

Furthermore, there are special requests from the user for evidence, source, or other information. Very rarely, they were mentioned from the agent before, which explains the small implicit numbers for these three. The agent only asks if they were needed, when it tries to check the request of the users last turn. Surprisingly, the participants only requested the source of an argument in total 8 times in the whole study. This seems odd by the fact that the first study of this thesis showed that people have a high interest in the source of any statement. In addition, they were instructed that this functionality exists in this system, which excludes a possible lack of knowledge in the experiment. A possible explanation for this result could be the missing trust of the system to deliver any useful sources. The agent came up with possible sources when it was asked to reveal them, but to find out that all the sources were made up, the user should at least have requested them once. Only 8 source request in 72 sessions would not explain this behavior. Another explanation could be the opposite way around: the arguments were so comprehensible that a request for the source is unnecessary. This would speak for the collection of arguments which were chosen for the study. The requests for evidence are a bit higher but still in small numbers. Reasonings for this could be similar to the sources. Interesting and at the same time problematic, are the high numbers of additional information requests. This feature was not supported by the system and the agent mostly had to outwardly say that it does not had any results to the request. An information request could be *"How much does an electric car cost?"* or *"Give me good products for vegetarian diet."*. The agent had only the information on its paper sheets and it would be coincidence if one of the requested facts would be part of one argument. It could only react with general knowledge to some requests. The total number of 0 implicit requests also shows the lack of providing this feature by the agent. Relating to the interpretation of

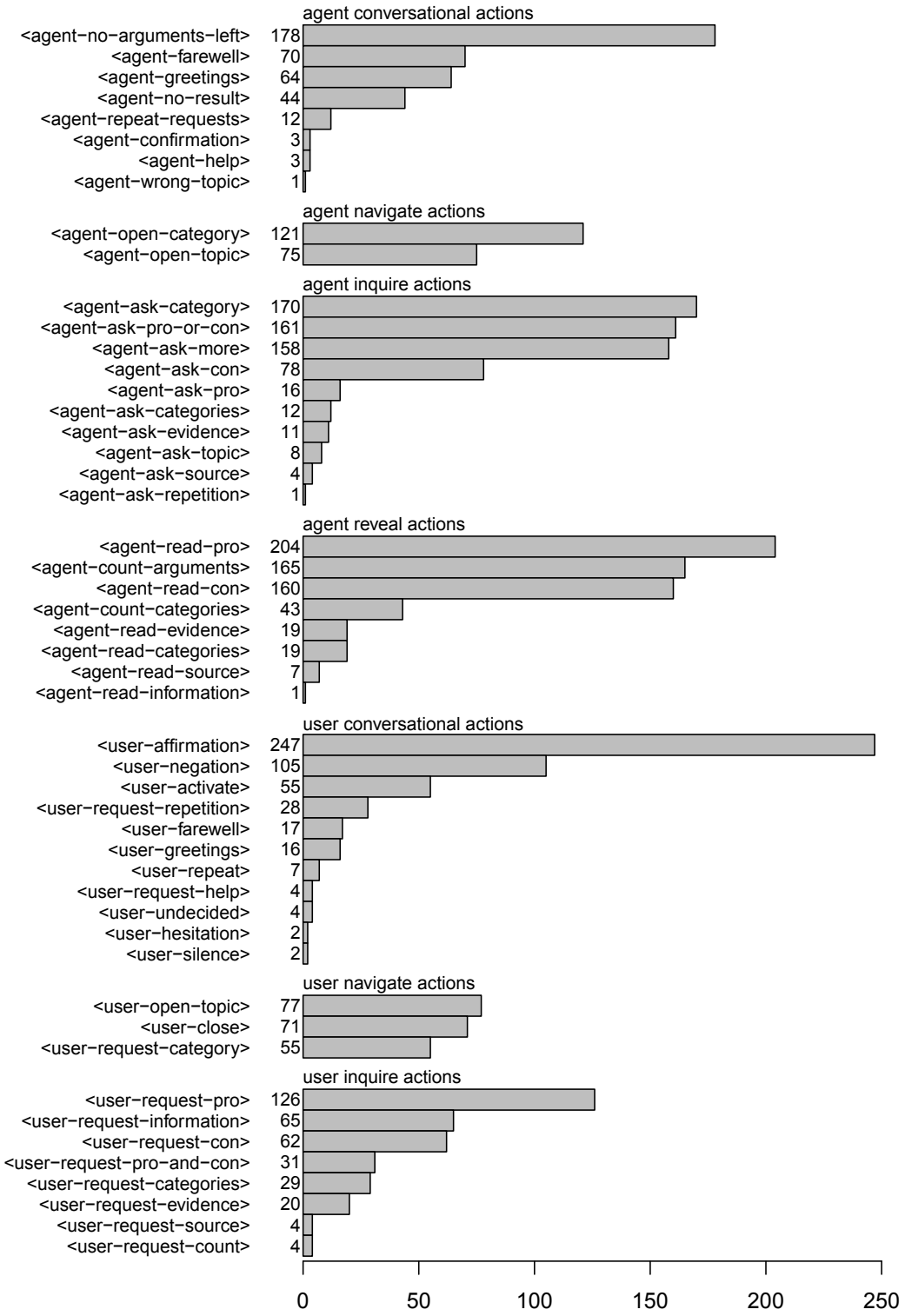

Figure 4.4: Distribution of all action by the user or agent which were expressed explicitly.

the total number of all actions, this topic will be discussed in a later part of this subchapter.

The last data presented independent from the hypotheses is the count of all actions grouped by their category (see figure 4.4). The action tags are sorted by their category and the amount of occurrences in the sessions. It should be pointed out that, the figure only shows the explicitly expressed statements by the user or the agent and not the implicit ones. These are indirectly included in the actions with the tag "<user-affirmation>", because the user only affirms implicitly an inquire action of the agent. The numbers of the implicit actions were presented in table 4.7. Nevertheless, figure 4.4 gives insights which actions should get the highest attention in the implementation of the system. Additionally, actions like "<agent-no-arguments-left>" with 178 occurrences need a proper implementation because the agent will use them in a real scenario very often. The agent also needs many different utterances for this action to appear more natural. Only one utterance would benefit the predictability of the system and the expectation of the user, but it would also make the agent appear more like a machine than a conversational partner. On the other hand, there are the user actions. The more occurrences the actions have there, the more different utterances the system has to learn to recognize the right intent of the user. The action "<user-affirmation>" has the highest occurrence number of 247, but it is also the easiest one to understand by the system. The user mostly only uses utterances like "yes" (153 occurrences) or "yes, please" (37 occurrences) for this kind of action. Accordingly, actions classified as "<user-negation>" are also easy to recognize by the system. The user navigate and require actions are more problematic. In 72 sessions, the users opened a topic 77 times. This number seems strange at first glance, but some user thought they closed the session accidentally because of a timeout and tried to reopen it. The utterances here are highly dependent on the topic, and the system had difficulties in understanding them. Following snippet from a transcript shows that even a human agent can misinterpret an intent of a user:

User:	*Alexa, why is animal confinement bad?*
Alexa:	*Do you want to hear arguments about the topic "Humans should stop eating animal meat"?*
User:	*No... Alexa, can you tell me whether I should visit the zoo?*
Alexa:	*Open topic "Zoos should be forbidden" ...*

The cruel treatment of animals could either be a hint to the topic "Stop eating meat" or to the topic "Zoos should be forbidden". In this case it was a try by the user to open the latter topic but the agent misinterpreted it. Far more problematic

#	information requests	resource
1	Definition “consciousness”.	encyclopedia
2	What does WWF stand for?	encyclopedia
3	How long does the battery of an electric car last?	product detail
4	What is [ahh] how much is [ahh] is the average cost of electric cars compared to the normal traditional ones?	shop comparison
5	Why should I [ahh] what should I eat [ahh] instead of meat?	health and food guide
6	[ahh] Is there any information or any research that [ahh] told me about what children themself feel about uniform?	scientific work
7	Which countries do have a school uniform?	school and country statistics
8	What is the number of animal farms in Germany with more than one million animals?	agriculture information
9	Okay. Do you know how many people will be at the zoo Erfurt tomorrow?	attendances statistics
10	What do you think of this topic?	decision-making ability

Table 4.8: Information requests by the user with information resources how to solve them.

are the request for additional information, classified with the tag “<user-request-information>”. They occurred 65 times and are the most difficult to recognize with predefined utterances and intents. Another issue is the resource of information. Table 4.8 shows some examples of information requests and the information resources to resolve them. The first two requests are simple and only about asking for definitions. The integration of some lexicons or using an API of an online encyclopedia can answer these requests. Likewise, the requests 3 and 4 are more complicated but still manageable in theory. The users asked for product details of one item or a comparison between multiple items. Although a challenging task, this problem can be solved with databases of product information. Only a good organization of product information is needed. The requests 5 to 9 represent the major challenge in this regard. The user requests information from several platforms which are either hard to retrieve or difficult to analyze. Health and food can be a highly subjective matter and many blogs and articles can write about different ideas of this topic. The number of attendances of a facility on the other hand are often concealed and not open for public. Google created recently a new search engine for public datasets[6] which could solve the problem. Still in development, this engine is limited to a few resources and can not find data to every possible topic. The last request represents a special issue. Instead of asking the agent which arguments can be presented, the user requests the opinion of

[6] https://toolbox.google.com/datasetsearch, accessed 08.09.2018

	Car		Zoo		Study		Meat		AI		Uniform	
	D	C	D	C	D	C	D	C	D	C	D	C
pro	28	30*	18	18*	5	7*	9	9*	7	12*	15	17*
con	19	16	18	7	8	10	13	8	9	11	13	13
ratio	.60	.65	.50	.72	.38	.41	.40	.53	.44	.52	.54	.67

Table 4.9: Distribution of pro and con arguments requests by topic and the motivations "making a decision" (D) or "convincing somebody" (C). The numbers with a * mark the polarity which was given in the background story to convince somebody.

the topic from the agent itself. Despite the fact that the agent has a database of many arguments at hand, it is difficult to judge which arguments are the most convincing ones. The assessment of argument quality is still a controversial topic (Wachsmuth et al., 2017a).

After collecting some data about the actions between user and agent, now the hypotheses are proceeded. It starts with the first thesis about the differentiation between the motivation of making a decision and to convince somebody.

1. When users want to convince somebody they are focused on getting arguments for their side.

It was assumed that users who try to convince somebody only want to hear one side of the topic which represents their own opinion. The opinion of the user was given in the background story of the task. An example text for convincing a friend for the pro side of the topic "Zoos should be forbidden" can be seen in figure 4.2 back on page 37. Table 4.9 shows a summary of how many requests were made for which side of a topic, subdivided in making a decision and convincing somebody. The question, if the polarity of the convincing motivations had an impact on the user requests, can be investigated with two methods. First, there is a look only on the "convincing somebody" cases. The given polarity was always on the pro side of the topic and the ratio to each topic shows that in 5 of 6 cases the participants asked for more pro than con arguments. The mean of all ratios is 0.5833 (sd=0.0852) in respect to the pro side. This is evidence which supports the hypothesis. However, the ratios for "making a decision" cases need to be included as well. If the ratios have the same characteristics here, a polarity to the pro side of the topics would be less expressive. The results show that participants requests more con arguments in the "making a decision" cases which leads to a more equal distribution of the requests for pro and con sides. Only 3 topics had a ratio bigger equal 50% and the other 3 had a stronger con polarity. The

mean value is only 0.4767 (sd=0.1622) for the positive ratio. The second method is to compare the ratios with each of the two motivations for every topic. The data reveals that in every case the polarity was a bit more for the pro side of the topics in the "convincing somebody" cases than in the "making a decision" cases. Therefore, all presented results indicate that the "convincing somebody" tasks lead to a higher polarity to one side than in the "making a decision" where the search for arguments is more equally distributed. The Wilcoxon signed-rank test shows significance of unequal datasets of 0.027 with -2.207 as standardized test statistic. The calculated effect size by the formula 3.1 from page 21 gives a value of 0.64 which indicates a large effect size by Cohen.

2. Users felt the argument presentation was better organized when they had a category-guideline at the beginning.

The second hypothesis states that users prefer the system with category-guideline more because the filtering of the categories leads to a better structure of the arguments. For this reason, the feedback data of the system in form of Likert scale ratings is examined. The participants had the chance to rate the system with and without category-guideline within the dimensions, asking if the system was helpful, fast, pleasant, natural, well-structured, behaved as expected, and if they would recommend it to others. Additionally, there was a rating if the participant was already well-informed about the topic, which could have a negative influence on the other dimensions. The descriptive statistic of all feedback dimensions of the system without category-guideline can be seen in table B.10 and with category-guideline in table B.11 in the Appendix. In the following table B.12 are the significance values with effect size if the datasets of the system feedback dimensions are unequal.

The first look on the descriptive values shows that the system without the category-guideline got a slightly better rating with a mean value of 1.889 (sd=0.75) in the dimension "well-structured" than the system with category-guideline with a mean value of 1.972 (sd=0.97). Because these values do not differ much, a second look is taken on the unequal significance between the two systems in this regard. The Wilcoxon signed-rank test gives a significance of 0.821 of unequal datasets. Both aspects show that their is no significant difference between the two ratings which declines the hypothesis. A possible error in this analysis could be the experimental-setup. As presented in figure 4.1 of page 42, the participants were always confronted first with the system without category-guideline and then with the new filtering feature. If the participants rated the system already with a perfect score in "well-structured" on the first task, it was not possible to gave it a better

score on the next task with the help of a category-guideline. The questionnaire should have been handed out after conducting the experiment with both systems. However, the collected qualitative data from the participants in the next section of this thesis will show that a greater part of the participants disliked the system with category-guideline. However, if the other feedback dimensions are considered, the system with category-guideline is shown in a better light. The participants felt the system being faster, despite the fact that the task completion times were higher (see table 4.3 on page 48), felt more natural, behaved as expected and was more pleasant to use. Especially the last mentioned feedback has a significance of unequal datasets of 0.001 with a medium effect size 0.39, which shows that the participants think the advanced system is more comfortable to use. The lesser significance of unequal datasets with a value of 0.419 to the dimension "already-well-informed" shows that in both systems the participants were more or less equally good informed about the topics which should have no influence on the evaluation of this data.

3. There is a correlation between satisfaction and expectation of the system.

The third hypothesis states that users of a voice assistant like the system more when it behaves as expected from the beginning. This behavior was already observed only on qualitative data in the works of Luger and Sellen (2016) and Myers et al. (2018). The summarized descriptive data of the feedback of all systems can be seen in table B.13 in the Appendix. The following table B.14 shows the significance of unequal datasets between all feedback dimensions.

The tests on unequal datasets show that there is not much correlation of the "expected" system feedback to any other dimension. A high correlation can be assumed if the p-value of the significance tests tends to 1 in the Wilcoxon ranked-test. The only connection which can be found is the relation to the feedback how pleasant the system felt. Both dimensions have a lesser significance value of 0.823 between unequal datasets which indicates a small correlation. However, the dimension "recommended" had the intent to give an overall impression of the system and with a significance of less than 0.05 and a small effect size of 0.17, the data shows a slightly difference between the two dimension. Hence, no clear evidence could be found that supports the idea that an system which behaves expected is essential for success of the system. An explanation for these insights could be the style in which this study was conducted. Luger and Sellen (2016) and Myers et al. (2018) performed their studies with real agents in form of smartphones or home devices, without conducting a Wizard of Oz experiment. This leads to many obstacles which were shown in the work of Myers et al. (2018).

The obstacles have a bad influence on the expectations of the system. But when they are eliminated, the user does not need full control of every aspect of the system as long as the system delivers the information which is needed. It seems the factor of "expected behavior of a system" is an issue only then, when it is related to natural language processing or wrong intent matching.

4.4.2 Qualitative Data

After the presentation of the quantitative data, this chapter discusses now the collected qualitative data in form of comments of the participants. The participants could write comments on a questionnaire after each task they completed. The qualitative questions at the end of the experiment were presented in section 4.1.5. Before the last hypothesis is verified, first a summary and presentation of the most frequently written comments by the participants.

Overall, 89 written comments were collected through the whole study. The bigger part of them was originated from small notes at the end of each task, 50 come from the the post-study questionnaire. To get an overview of the kind of comments the participants expressed at the end of each task, the comments there were categorized in the following groups: comments about the missing of additional information, the presentation of the arguments, the quality of the arguments, the memory problems of the users, the design of the category-guideline and the study itself. The percentage of each category can be seen in figure 4.5. In the following paragraphs all categories are explained and which comments the participants made there.

The **category-guideline** was the most commented issue in the qualitative data. The participants gave notes if the system behaved better with or without the help of a category filter. Only four out of ten users stated that they find the system with category-guideline more pleasant to use. The categories helped to get an overview of the topic and to structure the thoughts and gained knowledge of the system. Further, if two different topics have the same category (e.g., environmental aspects) it is possible to compare them in this category (e.g. has living without meat the same positive impact on the environment as having an electric car?). Nevertheless, the amount of negative feedback to the category-guideline was quite higher. Five of the participants stated that they were confused with the filtering of the arguments at the beginning of a task. They wanted to hear the arguments immediately after choosing a category without listening to the rest of the category list. Two participants complained about the unnecessary high or low number of categories in specific topics. While the topic "Buying an electric car" had 9

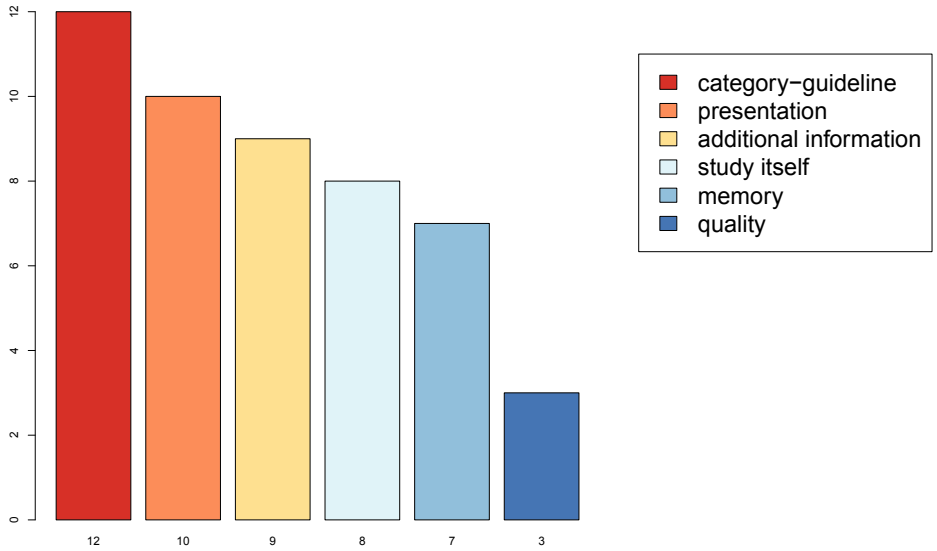

Figure 4.5: Distribution of comment categories.

categories with a small number of arguments for each, the topic "Fundamental rights for conscious AI" had only 2 categories with more arguments respectively. It seems a good advice to stay with a size of around 5 categories which are also meaningful for the whole context of the topic. Other less frequently mentioned comments were addressing the misinterpretation of the category titles, exploration difficulties with predefined categories, or that the selection of categories felt very machine-like and not like a natural conversation.

The **presentation** of the arguments was another often mentioned issue. Over the half of the comments were about the length or the too complicated content of the arguments. The statements should be short and precise without including technical terms which need to be looked up afterwards. Two participants note down that they would like to have paired pro and con arguments. So instead of first listing the pro arguments and then the con arguments, the agent can combine them to one phrase, like "It's true that ... however ...". This would help to make the system sound more natural but is also in reality a difficult topic in speech synthesis. Two participants criticized the restarting enumeration of arguments with every reveal action. They stated that the counter should be continuous through the whole list, which certainly makes sense when someone wants to refer to already mentioned arguments, but is difficult to realize in a system with a real human as agent who reads some arguments randomly from the list. Maybe, it would make sense to give the user the control of the enumeration scheme in a real implementation of the system. Another less mentioned comment was about

the constant request if the user wants to hear pro or con arguments after opening a topic. If the user only wants to hear one side for the whole topic, the system should remember not always to asks for the other side.

The **additional information** is a problem which was already discussed in the last section about the evaluation of the quantitative data. Figure 4.4 on page 54 shows the total numbers of explicit requests for additional information. Alone 65 out of 343 inquire requests by the user are for other resources which can not be found in the data of the arguments. These can be simple requests for the definition of words, up to complex information which need to be retrieved from arbitrary sources. Four participants stated that they simply wanted more information and felt limited by the question space. Another four commented that it should be possible to ask for simple definitions. One participant felt a break in the flow when the agent freely present its own data, but struggled when deeper information of the statements were requested. This information retrieval issue is a crucial problem for an argument search engine. In the web the user would simply use a search engine for retrieving the relevant information for him or herself, but one application cannot handle all possible information requests by its own, be it voice or web interface.

The **study itself** was mentioned a few times with several comments. First, two participants noted that they like synthesized voice of the agent. It remains questionable if the participants wrote this because they really liked the voice from the agent or because they knew there was a real human behind the system. Another two participants wrote that the background stories of the tasks were too general. It would be important in which city you like to buy an electric car and which countries are available in which to study abroad. To address as many people as possible, the background stories were formulated very general on purpose. One participant complained about the quiet voice of the agent which was suppressed by other noises outside the building. This was fixed in the hardware setup immediately. Another user revealed that there was a problem with the polarity of a topic. The topic “Zoos should be forbidden” had a positive polarity against zoos, but when the participants asked for pro arguments, he or she expected arguments in favour for the zoo. Maybe the stance of the conclusion should always be formulated in the way that it supports the main subject in the topic. There was also a surprising comment about the tasks. One participant wrote that the system did not help to convince a friend but to change the mind of the participant. It seems possible to have a change in motivations while listen to arguments for the contrary side.

The **memory** of the users was another issue which was noticeable through the whole study. Many participants complained that it is impossible to remember all arguments without taking notes. When a list of around 5 arguments was listed,

the first ones were already forgotten. Also some users asked for repetition of the arguments to gather keywords for further navigation. A few participants stated that a display is needed to bear the whole structure of the arguments in mind. As already mentioned in the presentation category, the arguments need to be shortly summarized so that the users can process them in the most efficient way without memory overload.

The **quality** of the arguments were the least mentioned comments. This is actually a good sign because it seems most of the participants were satisfied with the content of the arguments. Only a few people stated that there were too few arguments or they need to be more detailed. One participant also stated that the arguments for the topic "Fundamental rights for conscious AI" are not real arguments because they sound more like fears or predictions for the future without reasonable evidence from the present. This statement seems true but it is difficult to convince someone for a topic which lies far ahead in the future without the usage of a few speculations.

The qualitative data presented in the next paragraphs is collected from the comments of the post-study questionnaire. The insights there will help to validate the last hypothesis of this study which was mentioned at the beginning of this chapter.

4. Users think argument search with keyboard is still more efficient than with voice assistant.

To evaluate the hypothesis if people still prefer an argument search more with keyboard instead of a voice assistant, the comments of the first question of the post-study questionnaire are analyzed. At all 18 comments were collected, from which 4 had a complete positive attitude towards the new system, 6 which could take pleasure in a few aspects, and 8 which deny it almost completely.

The participants who prefer the voice assistant stated that they like to speak with a system in a natural way and it is very comfortable to do it hands-free in different situations like sitting on a sofa or maybe working in the kitchen. It is a fresh and new experience which they think is very handy in the future if such systems become reality.

Some participants had a divided opinion about the introduced system. They liked flexible voice input, which dominates the elaborate keyboard input in comparison, but the presentation of the arguments is the problem. The lack of visual feedback makes it hard to grasp the arguments and to get a picture of the whole

topic. Coupled with the missing interface to search for additional information, the output seems to be one of the major reasons why people would not like to switch to the new system.

The participants who still only trust in the old system stated similar reasons like the last group. They think a visual presentation is needed to manage the task of argument search in a proper way. It is easier to build a mental model of the topic when you have all arguments in one view and it is necessary to have an input controlled by hands to navigate fast between the resources. Especially arguments the participant already know are easier to skip in a text on a display as in a voice interface. Although, some participants stated that they are simply used to keyboards which has certainly a negative bias to new interfaces.

In conclusion, the hypothesis vindicated as true. Despite having an agent which has nearly no obstacles in natural language processing or wrong intent recognition, user still prefer an interface with keyboard which is easier to navigate and gives a better picture of the whole argument structure.

Participants still had the chance to write suggestions which they think would improve the novel system. Four people simply wrote that the abandonment of the category-guideline would already help to make the system less limited and more open to explore. Other four people mentioned the feature for asking additional information. It was already outlined why this is quite difficult in information retrieval. Furthermore, there was a suggestion for a better ranking scheme which three participants noted. It would be useful if the agent could explain why specific arguments are better than other and the user could choose by which criteria the arguments are sorted. Another often mentioned proposal was a screen to visualize the text. This thesis tried to implement a voice-only interface for the search of arguments. For future work, it would seem useful to include a visualization of the arguments on gadgets like home devices with screen or smartphones. They ensure that people can still use voice input with free hands. One last recommendation was the use of special sound effects to indicate the beginning of a new category or of an argument. This can be used to separate parts of the argument list not only by voice, but also by specific sound indicators. Maybe memory problems are solved when important parts of the audio are somehow highlighted.

4.4.3 Observations

This section is about observations which were noted during the experiments but which can not always be confirmed by data because the participants made no

comments about the issue. Nevertheless, they are meaningful to investigate and can be a topic for future work.

A few participants stated after the experiments that they expected a more natural conversation with the system. They said, the inquire requests by the agent did not have many utterances and it felt more like to speak with an automatic machine. This was true, because the agent was instructed to use the same script for every participant in purpose for comparison of the data. Likewise, the quantitative feedback data collected from the questionnaires shows divided results (see table B.13 in the Appendix). The "natural" dimension of the system feedback has a mean value of 2.403 (sd=1.109) in a scale from 1 to 5, which is only slightly better than the middle value of 3. It seems that, the participants had a torn opinion about this matter. The agent could understand more or less every request from the users without encountering any obstacles, which is a positive aspect of natural language processing, but the presentation of the arguments was less natural. Radlinski and Craswell (2017) analyzed that most of the information retrieval systems do not need a dialogue system which reflects a human level conversation and also in this study, a large group of participants could be satisfied with a simple listening of arguments. Yet, there should be methods which make a synthesized voice sound more appealing. Google Duplex[7] is an approach to train an agent in a very closed domain to sound very natural like a real person.

Another observation was the contact with the category-guideline. While most of the participants rejected the system completely, some could work fine with it after getting into the process. Maybe, the system would be preferred more if the participant got a better introduction for the category selection at the beginning of the session. Dubiel et al. (2018) also gave their participants an instruction to their more complex system. Anyway, it failed in comparison to the better conversational system which was similarly conducted in this study. Multifaceted search with categories is a useful method in most search interfaces on the web (Hearst, 2006), but it seems not to work in a voice only interface. The overview of all items, in this case arguments, is missing. The participants had problems to see which arguments were discarded by the filter. Maybe, it would have helped to tell the user how many arguments were associated with the category, before deselecting it. However, it is difficult to assess how much information the user can be presented without giving him or her an overload of memory. Already in this study it was exposed, that some users could not handle the amount of categories and facts they were confronted with.

[7] https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html, accessed 07.09.2018

Further on, there was an observation about the selection of topics. How often every topic was chosen was presented in table 4.4 on page 48. The instruction of the study was to select topics which the participant is not already well-informed about if possible. A few participants stated they deselected topics because of a special reason. They could not imagine to convince a friend about a topic, e.g., to stop eating meat. Instead, they picked another topic although they knew already more arguments about it. Table B.13 in the Appendix shows that most of the time the participants were already well informed about a topic when they selected one, with a mean value of 2.861 (sd=1.190) on a scale from 1 to 5. It is difficult to create background stories for the tasks which have a neutral polarity. In purpose to investigate the first hypothesis of the second study, a polarity needed to be set in order to have a comparison to the assumed equal distribution of argument requests from the "making a decision" cases. At the very least, the participants got the chance for two times to discard one topic, so that they were not totally forced to select topics they do not want to speak about.

A last stand out observation is about the politeness of the users towards the agent. The participants were free to say "Hello" to the agent to activate it, or simply to call it only by name. To say "Good Bye" was also nonbinding to close the session. In total 14 times the users greeted the agent, 17 times they said good bye and 82 times they used "please" in their requests. These numbers seem to be quite low in regard of 72 recorded sessions, but there is another interesting fact about them: they are mostly originated from female participants. Despite the fact, that only 6 of 18 participants were female, at all 7 greetings, 12 farewell and 36 mentions of "please" were from women. Normalized for comparison this means, that females used 2 times more greetings, 4.8 times more farewells and 1.6 times more the word "please". This is kind of surprising and indicates that men are more unfriendly than women towards an agent. Google is also aware of the problem, that agents mostly hear a commanding tone. For this they created a "Pretty Please" mode[8] to teach people, especially children, an appropriate tone for conversations.

[8]https://www.pocket-lint.com/apps/news/google/144422-google-assistant-pretty-please-mode-encourages-kids-to-say-please, accessed 07.09.2018

Chapter 5

Conclusion

This chapter is a recapitulation of the whole thesis, starting with a summary of the studies, followed by gained knowledge from the evaluations and finally future work about unresolved issues.

This thesis introduced a novel voice assistant application for argument search with preliminary analysis, development of a mock-up prototype and evaluation of the system. The core questions were: why would people want to use a voice assistant for argument search, how would they interact with such a system and what do they expect from it? Two studies were conducted for this purpose. The first study was an online survey which included questions on the acceptance of different situations, possible motivations and possible features for using a voice assistant for argument search. Additionally, participants could make comments about suggestions and concerns of such a system. The second study was about the design of a prototype and evaluation of the system in the form of a Wizard of Oz experiment. Arguments for six topics and a script with behavior rules were prepared for the human agent as the voice assistant. The participants had the task with the help of the voice assistant to make a decision or to convince somebody based on a background story. In addition, the system offered a free navigation through the arguments or a guideline with category filtering. Once the second study was conducted, the audio records were transcribed and later classified with tags for the agent and the user actions.

The results confirm the work of previous research on voice assistants and reveal insights for the implementation of an argument search engine for this interface. The first study confirmed that people see it as a good opportunity to use their voice assistant for the search of arguments. However, the acceptance of using the system depends on the privacy of the situation and the impact of the topic for which

arguments are requested. Most of the participants prefer to use the voice assistant in private places and only with familiar people around them. Moreover, they would rather use the voice assistant to make a buying decision, than for making a decision for a political vote. Apart from the serious usage, the participants showed a high interest to use the argument search engine for entertainment. This comes along with the data of Luger and Sellen (2016) which suggest that people need a playful interaction as an introduction to become acquainted with the application. The survey about the possibilities of the system showed that the participants have a high interest in the source of the arguments. Nevertheless, this could not be observed in the second study. Participants rarely requested the source but were often interested in additional information to the arguments. This additional information could not be provided in the experimental set-up of the study and represents a big issue in a real implementation of the system. The overview of the arguments and memory capacity of the participants was problematic, too. Although some participants stated that the category filter helped them to sort the arguments in their mind, most of them complained about the lack of getting a general view of the whole topic. This leads also to problems regarding the navigation to specific resources. Many participants suggested to include a screen to the system to handle that matter. Yet, over half of the participants enjoyed the usage of the voice assistant for argument search or at least saw benefits from it, in comparison to the traditional system with keyboard as input.

There are a few aspects which could not be investigated in this thesis and remain for future work. One case was the missing comparison in quantitative data to a traditional search interface. Many participants stated that they prefer the promptness and the overview of the data of a normal web interface with keyboard input, but no data is collected which shows the difference in completion time and satisfaction. It would be advisable to conduct a study which includes both interfaces. Another point was the vague definition of when a task was completed. In an exploratory search, it is always difficult to define when the goal is reached, and also the participants in the second study had to decide by themselves when they think they found enough arguments. Maybe, a predefined empty list which needs to be filled with arguments makes more sense as a goal for a following study. To help the user to get a better overview of the data, visualizations on screens can also be used. Most of the home devices have the ability to send text and graphics to the smartphone of the user or have a screen by themselves. This could become useful for future work. Further on, much data could be collected throughout the studies but not fully evaluated to all intents and purposes. One interesting investigation could be about the sentiment of the user transcripts and audio recordings. Did the participants show annoyance when they were confronted with obstacles, and how did they try to solve them? This leads to another analysis: a

Markov model to summarize the most used actions and states with their transition between agent and user. The structure of the intents of the system could get a good foundation with this model. Another analysis could be the selection of the categories. Did users always choose a specific category, or why did they decide to reactivate a category after filtering it out before? These are all questions which could not be answered in this thesis because of time constraints. Furthermore, the core question for a follow-up study should be, how long texts should be presented in form of audio for the user without him or her having memory problems, and how a playful interaction can be presented to the user so that he or she becomes acquainted with the novel system.

Appendix A

Study 1

	home alone	work with stranger	public alone	work with friend	public with stranger	home with friend
Valid	66	63	64	63	66	65
Missing	1	4	3	4	1	2
Mean	1.545	3.111	2.063	1.952	2.848	1.554
Median	1.000	3.000	2.000	2.000	3.000	2.000
Std. Deviation	0.6369	0.9523	0.9063	0.7055	1.070	0.5871
Variance	0.4056	0.9068	0.8214	0.4977	1.146	0.3447

Table A.1: Descriptive statistics of the situation ratings of the first study in an range of 1 to 4.

dataset 1		dataset 2	N	W	p	Z	r
home alone	-	work stranger	62	19.00	< .001	-6.262	.56
home alone	-	public alone	63	157.50	0.001	-3.235	.29
home alone	-	work friend	62	118.50	0.004	-2.855	.26
home alone	-	public stranger	65	93.50	< .001	-5.594	.49
home alone	-	home friend	64	289.50	0.861	-0.176	.02
work stranger	-	public alone	61	982.50	< .001	-5.357	.49
work stranger	-	work friend	60	995.00	< .001	-5.531	.51
work stranger	-	public stranger	62	402.50	.264	-1.117	.10
work stranger	-	home friend	61	1225.00	< .001	-6.253	.57
public alone	-	work friend	60	317.50	.721	-0.356	.03
public alone	-	public stranger	63	187.50	< .001	-3.939	.35
public alone	-	home friend	63	489.00	< .001	-3.893	.35
work friend	-	public stranger	62	68.50	< .001	-4.892	.44
work friend	-	home friend	62	468.00	< .001	-3.698	.33
public stranger	-	home friend	64	1420.00	< .001	-5.942	.56

Table A.2: Significance on dataset 1 ≠ dataset 2 of situations with Wilcoxon signed-rank test with N number of valid pairs, W the sum of ranks, p the significance value, Z the standardized test statistic and r the effect size (.1 = small, .3 = medium, .5 = large effect) .

	voting decision	buying decision	convince colleague	convince friend	fun alone	fun with friends
Valid	59	66	64	64	63	65
Missing	8	1	3	3	4	2
Mean	2.525	2.015	2.234	1.938	1.810	1.615
Median	3.000	2.000	2.000	2.000	2.000	2.000
Std. Deviation	0.9164	0.9363	0.8308	0.7099	0.8003	0.6541
Variance	0.8399	0.8767	0.6902	0.5040	0.6406	0.4279

Table A.3: Descriptive statistics of the motivation ratings of the first study in an range of 1 to 4.

dataset 1		dataset 2	N	W	p	Z	r
voting decision	-	buying decision	59	651	.001	-3.404	.31
voting decision	-	convince colleague	57	441	.010	-2.570	.24
voting decision	-	convince friend	59	470.50	< .001	-4.026	.37
voting decision	-	fun alone	55	790	< .001	-3.529	.34
voting decision	-	fun with friends	58	822	< .001	-4.780	.44
buying decision	-	convince colleague	63	324.5	.230	-1.202	.11
buying decision	-	convince friend	63	334	.517	-0.648	.06
buying decision	-	fun alone	62	413	.194	-1.300	.12
buying decision	-	fun with friends	65	576	.007	-2.710	.24
convince colleague	-	convince friend	62	253	.007	-2.694	.24
convince colleague	-	fun alone	61	572.5	.002	-3.089	.28
convince colleague	-	fun with friends	62	591	< .001	-4.214	.38
convince friend	-	fun alone	60	328	.196	-1.294	.12
convince friend	-	fun with friends	62	310	.002	-3.086	.28
fun alone	-	fun with friends	61	254	.035	-2.113	.19

Table A.4: Significance on dataset 1 ≠ dataset 2 of motivations with Wilcoxon signed-rank test with N number of valid pairs, W the sum of ranks, p the significance value, Z the standardized test statistic and r the effect size (.1 = small, .3 = medium, .5 = large effect) .

	pro/con arguments	number of arguments	arguments for aspect	argument evidence
Valid	67	66	64	65
Missing	0	1	3	2
Mean	2.090	2.848	1.906	1.800
Median	2.000	3.000	2.000	2.000
Std. Deviation	0.8480	0.8986	0.8677	0.8515
Variance	0.7191	0.8075	0.7530	0.7250

Table A.5: Descriptive statistics of the feature ratings of the first study in an range of 1 to 4 (a).

	argument source	share rating	share argument	play game
Valid	67	60	61	63
Missing	0	7	6	4
Mean	1.582	2.467	2.279	2.714
Median	2.000	2.000	2.000	3.000
Std. Deviation	0.8555	0.9107	0.9333	1.084
Variance	0.7318	0.8294	0.8710	1.175

Table A.6: Descriptive statistics of the feature ratings of the first study in an range of 1 to 4 (b).

dataset 1		dataset 2	N	W	p	Z	r
pro con arguments	-	number of arguments	66	134.50	< .001	-4.686	.41
pro con arguments	-	arguments for aspects	64	174.00	.105	-1.620	.14
pro con arguments	-	argument evidence	65	230.50	.003	-2.941	.26
pro con arguments	-	argument source	67	710.00	< .001	-3.785	.33
pro con arguments	-	share rating	60	256.00	.019	-2.353	.21
pro con arguments	-	share argument	61	262.50	.155	-1.421	.13
pro con arguments	-	play game	63	256.00	< .001	-3.501	.31
number of arguments	-	arguments for aspects	63	970.00	< .001	-5.253	.47
number of arguments	-	argument evidence	64	1033.00	< .001	-5.500	.49
number of arguments	-	argument source	66	1454.00	< .001	-6.279	.55
number of arguments	-	share rating	59	539.50	.010	-2.559	.24
number of arguments	-	share argument	61	785.00	< .001	-3.510	.32
number of arguments	-	play game	62	362.00	.426	-0.797	.07
arguments for aspect	-	argument evidence	62	132.00	.283	-1.075	.10
arguments for aspect	-	argument source	64	424.00	.024	-2.257	.20
arguments for aspect	-	share rating	58	140.00	< .001	-4.336	.40
arguments for aspect	-	share argument	59	187.00	< .001	-3.648	.34
arguments for aspect	-	play game	60	126.00	< .001	-4.557	.42
argument evidence	-	argument source	65	377.00	.024	-2.265	.20
argument evidence	-	share rating	59	108.00	< .001	-4.336	.40
argument evidence	-	share argument	59	112.00	< .001	-3.648	.34
argument evidence	-	play game	61	107.50	< .001	-4.725	.43
argument source	-	share rating	60	56.00	< .001	-5.376	.49
argument source	-	share argument	61	67.00	< .001	-4.299	.39
argument source	-	play game	63	95.00	< .001	-5.423	.48
share rating	-	share argument	56	340.50	.246	-1.160	.11
share rating	-	play game	56	228.00	.031	-2.156	.20
share argument	-	play game	56	76.00	.001	-3.190	.30

Table A.7: Significance on dataset 1 ≠ dataset 2 of features with Wilcoxon signed-rank test with N number of valid pairs, W the sum of ranks, p the significance value, Z the standardized test statistic and r the effect size (.1 = small, .3 = medium, .5 = large effect) .

	strength by AI	strength by users	source reliability	source coverage	aspect coverage	recency
Valid	64	66	66	66	65	64
Missing	3	1	1	1	2	3
Mean	3.391	2.970	1.530	2.136	2.431	2.484
Median	3.000	3.000	2.000	2.000	2.000	2.000
Std. Deviation	1.421	1.301	0.9318	1.135	1.414	1.414
Variance	2.020	1.691	0.8683	1.289	1.999	2.000

Table A.8: Descriptive statistics of the ranking ratings of the first study in an range of 1 to 6.

dataset 1		dataset 2	N	W	p	Z	r
strengh by AI	-	strengh by user	64	1025.00	.052	-1.928	.17
strengh by AI	-	source reliability	64	1618.00	< .001	-5.957	.53
strengh by AI	-	source coverage	64	1242.00	< .001	-4.721	.42
strengh by AI	-	aspect coverage	64	1026.50	< .001	-3.802	.34
strengh by AI	-	recency	63	1115.00	< .001	-3.942	.35
strengh by user	-	source reliability	66	1538.00	< .001	-5.740	.50
strengh by user	-	source coverage	66	1272.00	< .001	-3.625	.32
strengh by user	-	aspect coverage	65	975.00	.008	-2.667	.23
strengh by user	-	recency	64	984.00	.034	-2.123	.19
source reliability	-	source coverage	66	157.00	< .001	-3.826	.33
source reliability	-	aspect coverage	65	196.00	< .001	-3.822	.34
source reliability	-	recency	64	134.50	< .001	-4.142	.37
source coverage	-	aspect coverage	65	326.00	.107	-1.610	.14
source coverage	-	recency	64	474.00	.067	-1.830	0.16
aspect coverage	-	recency	63	583.50	.962	0.048	0.00

Table A.9: Significance on dataset 1 ≠ dataset 2 of ranking criteria with Wilcoxon signed-rank test with N number of valid pairs, W the sum of ranks, p the significance value, Z the standardized test statistic and r the effect size (.1 = small, .3 = medium, .5 = large effect) .

Argument Search with Voice Assistants

Page 1

First we would like to ask you a few general questions.

Page 2

1. What gender are you? *

- ◯ male
- ◯ female
- ◯ other

2. How old are you? *

- ◯ 17 years old or younger
- ◯ 18-30 years old
- ◯ 31-49 years old
- ◯ 50-64 years old
- ◯ 65 years or older

3. How often do you use computers or intelligent devices? *

	daily	weekly	rarely
On average, several hours...	◯	◯	◯

4. Do you have knowledge or experience in any of the following fields?

☐ Human Computer Interaction ☐ Information Retrieval ☐ Natural Language Processing

5. Do you use voice assistants? *

i.e., applications that you can ask through a microphone and that respond using a speaker.
e.g., Alexa, Cortana, Google Home, or Siri

- ◯ I use them frequently.
- ◯ I use them rarely.
- ◯ I've never used them before but I would like to try.
- ◯ What is a voice assistant?
- ◯ I am not interested in them, because: [ ]

6. How often do you inform yourself about controversial topics? *

e.g. politics, social issues, education, or technology

	daily	weekly	less often
I inform myself...	◯	◯	◯

7. Do you visit debate portals or even contribute to them? *

e.g. debate.org, debatepedia.org, or idebate.org

- ◯ I contribute to them.
- ◯ I only read them.
- ◯ I've never used them before.

8. Do you read or write comments on other websites? *

e.g. reddit, facebook, twitter, news sites, or message boards

- ◯ I write comments there.
- ◯ I only read comments there.
- ◯ I usually ignore comments there.
- ◯ I'm not on such websites.

Page 3

As part of our research, we want to find out when and why people would use a voice assistant for argument search. Imagine you have a voice assistant with you (e.g., in the smart phone), that you can ask for pro and con arguments for specific topics and which you trust to keep your data private.

Please rate how reasonable it would be for you to use a voice assistant for argument search in the following or similar scenarios. Give a separate rating for the situation in which the search is used and the motivation for using it.

1. You are sitting alone at the breakfast table at home and are reading in the newspaper about an important election in your country next week. You ask your voice assistant to give pro and con arguments for two specific parties to help you make a final voting decision. *

	convenient	plausible	unreasonable	inconceivable	don't know
For me, using the assistant alone at home would be ...	◯	◯	◯	◯	◯
For me, using the assistant to make a voting decision would be ...	◯	◯	◯	◯	◯

2. You are serving customers with a coworker behind the counter. Your coworker is annoyed because the staff have to start wearing specific work uniforms from tomorrow onwards. You ask your voice assistant to give pro arguments for work uniforms so that you can convince your colleague that they are actually practical. *

	convenient	plausible	unreasonable	inconceivable	don't know
For me, using the assistant at work in front of customers would be ...	◯	◯	◯	◯	◯
For me, using the assistant to convince my colleague would be ...	◯	◯	◯	◯	◯

3. You are sitting on a bench in the park in front of a lake with ducks. In the distance you see people walking their dogs. You are not a dog nor a cat person, but you want to know what it is like to have a duck as a pet. For fun and to entertain yourself, you ask your voice assistant to give you arguments for having a duck as pet. *

	convenient	plausible	unreasonable	inconceivable	don't know
For me, using the assistant alone in the park would be ...	◯	◯	◯	◯	◯
For me, using the assistant to entertain myself would be ...	◯	◯	◯	◯	◯

4. **A friend comes into your office at work. He mentions that there is a new charging station for electric cars in front of the building and he was also thinking of getting an electric car. You try to convince your friend that electric cars are not practical right now, so you ask your voice assistant for con arguments about this topic.** *

	convenient	plausible	unreasonable	inconceivable	don't know
For me, using the assistant at work with a friend would be ...	◯	◯	◯	◯	◯
For me, using the assistant to convince my friend would be ...	◯	◯	◯	◯	◯

5. **You are going into a very busy electronics store to buy a pre-configured PC for yourself. There are many Notebooks and Desktop PCs that fulfill your demands, but you cannot decide which kind of PC you want. To help you make a decision on which option is best, you ask your voice assistant for pro and con arguments of Notebooks and Desktop PCs.** *

	convenient	plausible	unreasonable	inconceivable	don't know
For me, using the assistant in a crowded store would be ...	◯	◯	◯	◯	◯
For me, using the assistant to make a buying decision would be ...	◯	◯	◯	◯	◯

6. **You are at home with a friend and you want to order pizza. You order a normal salami pizza and your friend orders a pizza hawaii. You talk about the sense or nonsense of pineapple on pizza, but can't come to a conclusion. To make a joke of the situation, you ask your voice assistant what negative arguments there are against pizza hawaii.** *

	convenient	plausible	unreasonable	inconceivable	don't know
For me, using the assistant at home with a friend would be ...	◯	◯	◯	◯	◯
For me, using the assistant to have fun with a friend would be ...	◯	◯	◯	◯	◯

Page 4

Please rate how much you would appreciate each functionality for a voice assistant. Assume that the functionality would work flawlessly.

1. **Get pro and/or con arguments on a specific topic.** *

e.g. Tell me supportive arguments to "electric cars".

	much appreciated	appreciated	nice to have	useless	don't know
For me, this functionality would be ...	◯	◯	◯	◯	◯

2. **Get the total number of arguments about a specific topic.** *

e.g. Tell me how many arguments exist on the topic "electric cars".

	much appreciated	appreciated	nice to have	useless	don't know
For me, this functionality would be ...	◯	◯	◯	◯	◯

3. **Get pro and/or con arguments about a specific topic for a related keyword.** *

e.g. Tell me supportive arguments to "electric cars" that contain "battery".

	much appreciated	appreciated	nice to have	useless	don't know
For me, this functionality would be ...	◯	◯	◯	◯	◯

4. **Get evidence about an argument mentioned beforehand.** *

e.g. Tell me evidence for the argument on "battery life".

	much appreciated	appreciated	nice to have	useless	don't know
For me, this functionality would be ...	◯	◯	◯	◯	◯

5. **Get the source of an argument mentioned beforehand (i.e., the news paper, forum, blog, ... that published it).** *

e.g. Tell me the source of the argument on "battery life".

	much appreciated	appreciated	nice to have	useless	don't know
For me, this functionality would be ...	◯	◯	◯	◯	◯

6. **Provide a rating for an argument mentioned beforehand (to help other users; malevolent ratings would be discarded).** *

e.g. Rate the argument on "battery life" with 3 out of 10.

	much appreciated	appreciated	nice to have	useless	don't know
For me, this functionality would be ...	◯	◯	◯	◯	◯

7. Provide an argument for the current topic (to help other users; malevolent contributions would be discarded). *

e.g. Add opposing argument "Electric vehicles have a limited range due to the low energy density of batteries."

	much appreciated	appreciated	nice to have	useless	don't know
For me, this functionality would be ...	◯	◯	◯	◯	◯

8. Play a debate game of finding the best arguments, as determined by the argument ranking of the search engine. *

e.g. You: Electric cars can slash greenhouse gas emissions.
Assistant: This argument has a score of 42.2. There remain two other high-scoring pro arguments to be found.

	much appreciated	appreciated	nice to have	useless	don't know
For me, this functionality would be ...	◯	◯	◯	◯	◯

9. Are there some other functionalities you would like to use? If yes, please describe them here briefly.

10. The voice assistant would provide you with arguments of a topic ranked by a mixture of different criteria. Please rate the criteria from most to least important. *

	most important					least important	don't know
Machine Rating *argument strength as rated by an algorithm*	◯	◯	◯	◯	◯	◯	◯
User Rating *argument strength as rated by the users (if available)*	◯	◯	◯	◯	◯	◯	◯
Source Reliability *trustworthiness of the argument source as rated by a community*	◯	◯	◯	◯	◯	◯	◯
Source Coverage *arguments from various sources are preferred*	◯	◯	◯	◯	◯	◯	◯
Aspect Coverage *arguments that cover various aspects of the topic are preferred*	◯	◯	◯	◯	◯	◯	◯
Recency *up-to-date arguments are preferred*	◯	◯	◯	◯	◯	◯	◯

Appendix B

Study 2

	Topics	
	low-stake	**high-stake**
Decision Making	• Is Windows better than Mac? • Should Zoos be forbidden? • Should I wear a bicycle helmet?	• Should I buy an electric car? • Should I study abroad? • Should I stop eating animal meat?
Convince Somebody	• Should we colonize Venus before Mars? • Do Aliens exist? • Should Fundamental rights be extended to conscious general AIs?	• Does god exist? • Should women have the right to choose abortion? • Should our school introduce school uniform?
Entertainment	• Is “The Last Jedi” one of the weakest Star Wars movies so far? • Is water wet? • Is the earth flat?	

Table B.1: Variables and topic distribution in the early alpha version of the second study.

action tag	description
<agent-greetings>	agent says 'Hello' to the user when he only calls her name without further information
<agent-farewell>	agent says 'Good bye' to the user, when he or she signals enough arguments were found
<agent-confirmation>	agent confirms a questions of a user which can simply be answered as 'yes', e.g. 'Did you say economy?' -> 'Yes.'
<agent-help>	agent lists all possible command requests, after the user asked for help
<agent-no-arguments-left>	agent is at the end of an argument list; can be end of topic, category or one side of pro or con
<agent-no-result>	agent couldn't find any results to the user request or has no information about it
<agent-repeat-requests>	agent didn't understand the user requests and asks to repeat it
<agent-wrong-topic>	response when user tried to open the next topic, without closing the last one and didn't complete the questionnaire

Table B.2: Conversational Actions, typical actions the agent provides to inform the user.

action tag	description
<agent-open-topic>	agent opens a topic after the user requested it
<agent-open-category>	agent opens a category after the user requested it

Table B.3: Navigate Actions, the agent opens a new resource with arguments for the user.

action tag	description
<agent-ask-category>	agent offers arguments of a specific category
<agent-ask-categories>	agent offers the full list of all categories
<agent-ask-pro-or-con>	agent offers the pro or con arguments of a topic or category
<agent-ask-pro>	agent offers the pro arguments of a topic or category
<agent-ask-con>	agent offers the con arguments of a topic or category
<agent-ask-more>	agent offers more arguments of the current side of a topic or category
<agent-ask-repetition>	agent offers to repeat something mentioned before
<agent-ask-source>	agent offers the source of a mentioned argument
<agent-ask-evidence>	agent offers to read the evidence of a mentioned argument
<agent-ask-topic>	agent offers to open a topic for the user

Table B.4: Inquire Actions, the agent asks the user if he or she wants to hear something.

action tag	description
<agent-read-pro>	agent reads pro arguments, up to 3 or end of list
<agent-read-con>	agent reads con arguments, up to 3 or end of list
<agent-read-source>	agent reads the source of a mentioned argument
<agent-read-evidence>	agent reads evidence of a mentioned argument
<agent-read-information>	agent gives more information because of a request
<agent-read-categories>	agent lists all available categories to a topic
<agent-count-arguments>	agent gives count of all arguments of a topic or category
<agent-count-categories>	agent gives count of all categories of a topic

Table B.5: Reveal Actions, the agent locates resources or attributes.

action tag	description
<user-activate>	user says only keyword to activate the agent
<user-greetings>	user says 'Hello' to start a conversation with the agent
<user-farewell>	user says 'Good Bye'
<user-affirmation>	user affirms a question by the agent
<user-negation>	user negates a question by the agent
<user-request-repetition>	user wants the repetition of last statement from the agent
<user-repeat>	user repeats one of his or her last statements
<user-request-help>	user wants a list of all possible commands
<user-hesitation>	user wavers and has problems to form next request
<user-undecided>	user cannot affirm or negate an inquire action from agent
<user-silence>	user did not respond after Alexa statement

Table B.6: Conversational Actions, typical actions the user performs towards an agent.

action tag	description
<user-open-topic>	user makes request to open topic
<user-request-category>	user wants arguments to specific category
<user-close>	user closes the topic

Table B.7: Navigate Actions, the user wants to open or close a new resource of arguments.

action tag	description
<user-request-pro>	user wants pro arguments
<user-request-con>	user wants con arguments
<user-request-pro-and-con>	user wants pro and con arguments
<user-request-categories>	user wants to hear the category list
<user-requests-information>	user requests specific information
<user-request-source>	user requests source of a mentioned argument
<user-request-evidence>	user requests evidence of a mentioned argument
<user-request-count>	user requests the total number of arguments

Table B.8: Inquire Actions, the user requests specific information from the agent.

shortcut	Description
[ahh]	speech disfluency for retention
[hmm]	speech disfluency for thinking
[uhm]	speech disfluency for doubtfulness
[!]	interruption of the agent by the user
[XXsec]	pause for XX seconds, minimum 2 seconds
[?]	not understandable

Table B.9: Indicators to mark specific events in the transcript of the user.

	already-well-informed	helpful	fast	pleasant	expected	natural	well-structured	recommended
Valid	36	36	36	36	36	36	36	36
Missing	0	0	0	0	0	0	0	0
Mean	2.972	2.194	2.333	2.306	2.222	2.528	1.889	2.444
Median	3.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000
Std. Deviation	1.055	1.009	1.014	1.091	1.072	1.158	0.7475	1.027
Variance	1.113	1.018	1.029	1.190	1.149	1.342	0.5587	1.054

Table B.10: Descriptive statistics of rating the system without category-guideline in a range of 1 to 5.

	already-well-informed	helpful	fast	pleasant	expected	natural	well-structured	recommended
Valid	36	36	36	36	36	36	36	36
Missing	0	0	0	0	0	0	0	0
Mean	2.778	2.139	2.111	2.167	2.111	2.361	1.972	2.472
Median	3.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000
Std. Deviation	1.333	0.9305	1.008	1.082	1.008	1.199	0.9706	1.082
Variance	1.778	0.8659	1.016	1.171	1.016	1.437	0.9421	1.171

Table B.11: Descriptive statistics of rating the system with category-guideline in a range of 1 to 5.

datasets	N	W	p	Z	r
already-well-informed	36	122.50	.419	-0.808	.10
helpful	36	54.00	.156	-1.419	.17
fast	36	24.00	.006	-2.763	.33
pleasant	36	6.00	.001	-3.337	.39
expected	36	85.50	.055	-1.916	.23
natural	36	32.50	.097	-1.661	.20
well-structured	36	80.50	.821	-0.226	.03
recommended	36	38.00	.092	-1.685	.20

Table B.12: Significance of unequal datasets from the feedback of the system without and with category-guideline. Conducted with Wilcoxon signed-rank test with N number of valid pairs, W the sum of ranks, p the significance value, Z the standardized test statistic and r the effect size (.1 = small, .3 = medium, .5 = large effect) .

	already-well-informed	helpful	fast	pleasant	expected	natural	well-structured	recommended
Valid	72	72	72	72	72	72	72	72
Missing	0	0	0	0	0	0	0	0
Mean	2.861	2.069	2.097	2.028	2.014	2.403	1.875	2.333
Median	3.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000
Std. Deviation	1.190	0.9243	1.050	0.9782	1.014	1.109	0.9029	1.075
Variance	1.417	0.8543	1.103	0.9570	1.028	1.230	0.8151	1.155

Table B.13: Descriptive statistics of rating the system in a range of 1 to 5.

dataset 1		dataset 2	N	W	p	Z	r
already-well-informed	-	helpful	72	1479.50	< .001	-4.258	.35
already-well-informed	-	fast	72	1232.50	< .001	-3.957	.33
already-well-informed	-	pleasant	72	1409.00	< .001	-4.394	.37
already-well-informed	-	expected	72	1387.50	< .001	-3.854	.32
already-well-informed	-	natural	72	913.50	.006	-2.723	.23
already-well-informed	-	well-structured	72	1405.50	< .001	-4.316	.36
already-well-informed	-	recommended	72	1321.00	.002	-3.118	.26
helpful	-	fast	72	308.00	.904	-0.120	.01
helpful	-	pleasant	72	332.50	.762	-0.302	.03
helpful	-	expected	72	454.00	.532	-0.624	.05
helpful	-	natural	72	336.00	.021	-2.302	.19
helpful	-	well-structured	72	590.00	.134	-1.499	.12
helpful	-	recommended	72	225.00	.045	-2.006	.17
fast	-	pleasant	72	341.00	.634	-0.476	.04
fast	-	expected	72	366.50	.585	-0.547	.05
fast	-	natural	72	261.50	.022	-2.285	.19
fast	-	well-structured	72	557.00	.088	-1.706	.14
fast	-	recommended	72	146.00	.004	-2.863	.24
pleasant	-	expected	72	275.50	.823	-0.223	.02
pleasant	-	natural	72	219.00	.004	-2.871	.24
pleasant	-	well-structured	72	507.00	.163	-1.393	.12
pleasant	-	recommended	72	124.00	.001	-3.245	.27
expected	-	natural	72	311.00	.014	-2.452	.20
expected	-	well-structured	72	514.00	.261	-1.124	.09
expected	-	recommended	72	235.50	.040	-2.049	.17
natural	-	well-structured	72	793.50	< .001	-3.620	.30
natural	-	recommended	72	713.00	.619	-0.497	.04
well-structured	-	recommended	72	193.00	.001	-3.176	.26

Table B.14: Significance on dataset 1 ≠ dataset 2 of system feedbacks with Wilcoxon signed-rank test with N number of valid pairs, W the sum of ranks, p the significance value, Z the standardized test statistic and r the effect size (.1 = small, .3 = medium, .5 = large effect) .

topic	background story
A	**Task: Decide whether to buy an electric car** You just finished your studies and got a job and a nice apartment. The apartment is 30 km away from your new job, so your bike will no longer do it for you. You care a lot for the environment, but you worry that electric cars are still not ready to be used on a daily basis and want to come to an informed decision whether you should buy an electric or a regular car.
B	**Task: Decide whether to visit the Zoo.** You want to go to the Zoo at the weekend to see wild animals in real life and to learn more about them. However, you saw recent protests in your town that convinced you that animal confinement is bad. You struggle if it is okay to visit and thus support such a facility and want to come to an informed decision in this regard.
C	**Task: Decide whether to study abroad.** You started to study and now got the chance to study in a different country for your Master's degree. However, you are afraid that you would loose the connection to the friends you made in Weimar if you would do so. On the other hand, you believe visiting other countries and experiencing their culture is very beneficial for self-development, and this would be a good chance to do so. You want to come to an informed decision in this regard.
D	**Task: Decide whether to stop eating meat.** You go to Mensa most days and observe that more and more students choose the vegetarian menu. You are concerned that not eating meat is unhealthy. However, you also heard the other students talk about the benefits of vegetarianism for the environment, which interests you. You wonder whether you should change your eating habits and want to come to an informed decision in this regard.
E	**Task: Decide whether conscious general AI should get fundamental rights.** You are researching in general artificial intelligence. You are asked to sign a petition which states that fundamental rights for conscious general artificial intelligence should be established now to avoid morality problems in the future. You are sceptical on defining consciousness legally, but you are also concerned about the morality problems. You want to come to an informed decision on whether to sign the petition.
F	**Task: Decide whether to introduce a school uniform.** You belong to the student council of your school and one of the upcoming topics is whether to introduce a school uniform. You are worrying about the acceptance of such a new school regulation. On the other hand, you heard it would help students to achieve better grades. You want to come to an informed decision on how to vote in the next council.

Table B.15: Background stories of the tasks when the motivation is to make a decision.

topic	background story
A	**Task: Convince your friend to buy an electric car.** Your friend just finished her studies and got a job and a nice apartment. The apartment is 30 km away from the job, so her bike will no longer do it for her. You see this as a chance to help the environment and to convince your friend of buying an electric car. However, you know she worries that electric cars are still not ready to be used on a daily basis. You are now looking for arguments that help you convince your friend.
B	**Task: Convince your friend not to visit the Zoo.** Your friend wants to go to the Zoo at the weekend to see wild animals in real life and to learn more about them. However, you saw recent protests in your town that convinced you that animal confinement is bad. You therefore want to convince your friend it is not okay to visit and thus support such a facility. You are now looking for arguments that help you convince your friend.
C	**Task: Convince your friend to study abroad.** Your friend started to study and now got the chance to study in a different country for his Master's degree. However, he is afraid that he would loose the connection to the friends he made in Weimar if he would do so. You believe visiting other countries and experiencing their culture is very beneficial for self-development, and this would be a good chance to do so. You therefore want to convince your friend to study abroad. You are now looking for arguments that help you convince your friend.
D	**Task: Convince your friend to stop eating meat.** Your friend goes to the Mensa most days and observed today that more and more students choose the vegetarian menu. She is still concerned that not eating meat is unhealthy. You yourself stopped eating meat recently as you heard other students talk about the benefits of vegetarianism for the environment. You think now is a good time to convince your friend to stop eating meat, as well. You are now looking for arguments that help you convince your friend.
E	**Task: Convince your friend to support fundamental rights for conscious general AI.** Your friend is researching in general artificial intelligence. You come across a petition which states that fundamental rights for conscious general artificial intelligence should be established now to avoid morality problems in the future. You know your friend is very sceptical on defining consciousness legally, but you are very concerned about the morality problems. You therefore want to convince your friend to sign the petition. You are now looking for arguments that help you convince your friend.
F	**Task: Convince your friend to support a school uniform.** Your friend belongs to the student council of your school and one of the upcoming topics is whether to introduce a school uniform. Your friend is worrying about the acceptance of such a new school regulation. However, You heard it would help students to achieve better grades. You therefore want to convince your friend to vote for introducing school uniforms. You are now looking for arguments that help you convince your friend.

Table B.16: Background stories of the tasks when the motivation is to convince somebody.

Bibliography

Avula, S. (2018). Wizard of oz: Protocols and challenges in studying searchbots to support collaborative search. 2

Azzopardi, L., Dubiel, M., Halvey, M., and Dalton, J. (2018). Conceptualizing agent-human interactions during the conversational search process. 2, 4.1.6

Carey, A. (1967). *The Hawthorne Studies: A Radical Criticism*. American sociological review. [Offprint]. Bobbs-Merrill. 2

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge. 3.4

Dubiel, M., Halvey, M., Azzopardi, L., and Daronnat, S. (2018). Investigating how conversational search agents affect user's behaviour, performance and search experience. In *The Second International Workshop on Conversational Approaches to Information Retrieval*. 2, 4, 4.3, 4.4, 4.4.1, 4.4.3

Easwara Moorthy, A. and Vu, K.-P. L. (2014). Voice activated personal assistant: Acceptability of use in the public space. In Yamamoto, S., editor, *Human Interface and the Management of Information. Information and Knowledge in Applications and Services*, pages 324–334, Cham. Springer International Publishing. 2, 3

Efthymiou, C. and Halvey, M. (2016). Evaluating the social acceptability of voice based smartwatch search. pages 267–278. 2, 3, 3.1.2, 3.4.1

Falkner, K., Szabo, C., Michell, D., Szorenyi, A., and Thyer, S. (2015). Gender gap in academia: Perceptions of female computer science academics. In *Proceedings of the 2015 ACM Conference on Innovation and Technology in Computer Science Education*, ITiCSE '15, pages 111–116, New York, NY, USA. ACM. 3.1.1

Hearst, M. A. (2006). Design recommendations for hierarchical faceted search interfaces. In *Proc. SIGIR 2006, Workshop on Faceted Search*, pages 26–30. 4, 4.4.3

IBM Corporation (2013). *IBM SPSS Statistics 22 Core System User's Guide*. Armonk, NY: IBM Corp. https://www-01.ibm.com/support/docview.wss?uid=swg27038407. 3.4, 4.4

JASP Team (2018). JASP (Version 0.9)[Computer software]. 3.4, 4.4

Kaushik, A. and Jones, G. J. F. (2018). Exploring current user web search behaviours in analysis tasks to be supported in conversational search. page 8. 2

Luger, E. and Sellen, A. (2016). "like having a really bad pa": The gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 5286–5297, New York, NY, USA. ACM. 2, 3, 3.1.2, 3.4.2, 4, 4.1.1, 4.1.4, 4.1.5, 4.3, 4.4.1, 5

Morris, M. E., Adair, B., Miller, K., Ozanne, E., Hampson, R., Pearce, A. J., Santamaria, N., Viegas, L., Long, M., and Said, C. M. (2013). Smart-home technologies to assist older people to live well at home. *Journal of aging science*, 1(1):1–9. 3

Mustafaraj, E. and Metaxas, P. T. (2017). The fake news spreading plague: Was it preventable? *CoRR*, abs/1703.06988. 1, 3, 3.4.1

Myers, C., Furqan, A., Nebolsky, J., Caro, K., and Zhu, J. (2018). Patterns for how users overcome obstacles in voice user interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 6:1–6:7, New York, NY, USA. ACM. 2, 3.4.2, 4, 4.3, 4.4.1

Papangelis, A., Papadakos, P., Kotti, M., Stylianou, Y., Tzitzikas, Y., and Plexousakis, D. (2017). LD-SDS: towards an expressive spoken dialogue system based on linked-data. *CoRR*, abs/1710.02973. 2

Porcheron, M., Fischer, J. E., Reeves, S., and Sharples, S. (2018). Voice interfaces in everyday life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 640:1–640:12, New York, NY, USA. ACM. 2

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. 3.4

Radlinski, F. and Craswell, N. (2017). A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, CHIIR 2017*. ACM. 2, 4.4.3

Rajendran, P., Bollegala, D., and Parsons, S. (2016). Contextual stance classification of opinions: A step towards enthymeme reconstruction in online reviews. In *Proceedings of the Third Workshop on Argument Mining, hosted by the 54th Annual Meeting of the Association for Computationa Linguistics, ArgMining@ACL 2016, August 12, Berlin, Germany*. 4.4.1

Rohde, M. and Baumann, T. (2016). Navigating the spoken wikipedia. 2

Stab, C. and Gurevych, I. (2016). Parsing argumentation structures in persuasive essays. *CoRR*, abs/1604.07370. 2

Trippas, J. R., Spina, D., Cavedon, L., and Sanderson, M. (2017). How do people interact in conversational speech-only search tasks: A preliminary analysis. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, pages 325–328, New York, NY, USA. ACM. 2

Ulrike Hahn, Adam JL Harris, A. C. (2009). *Argument content and argument source: An exploration*. 1

Vtyurina, A. and Fourney, A. (2018). Exploring the role of conversational cues in guided task support with virtual assistants. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 208:1–208:7, New York, NY, USA. ACM. 2

Vtyurina, A., Savenkov, D., Agichtein, E., and Clarke, C. L. A. (2017). Exploring conversational search with humans, assistants, and wizards. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, pages 2187–2193, New York, NY, USA. ACM. 2

Wachsmuth, H., Naderi, N., Habernal, I., Hou, Y., Hirst, G., Gurevych, I., and Stein, B. (2017a). Argumentation quality assessment: Theory vs. practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 250–255. 1, 4.4.1

Wachsmuth, H., Potthast, M., Al-Khatib, K., Ajjour, Y., Puschmann, J., Qu, J., Dorsch, J., Morari, V., Bevendorff, J., and Stein, B. (2017b). Building an Argument Search Engine for the Web. In *Proceedings of the Fourth Workshop on Argument Mining (ArgMining 17)*, pages 49–59. 1, 2, 3

Walton, D., Reed, C., and Macagno, F. (2008). *Argumentation Schemes*. Cambridge University Press. 2

Wildemuth, B. and Freund, L. (2012). Assigning search tasks designed to elicit exploratory search behaviors. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval, HCIR 2012*. 2

Wolters, M., Georgila, K., Moore, J. D., Logie, R. H., MacPherson, S. E., and Watson, M. (2009). Reducing working memory load in spoken dialogue systems. *Interact. Comput.*, 21(4):276–287. 2