Supporting Voice-Based Natural Language Interactions for Information Seeking Tasks of Various Complexity

by

Alexandra Vtyurina

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2021

© Alexandra Vtyurina 2021

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Max L. Wilson, Associate Professor, Mixed Reality Lab, University of Nottingham
Supervisor(s):	Charles L.A. Clarke, Professor, Cheriton School of Computer Science, University of Waterloo Edith Law, Professor, Cheriton School of Computer Science, University of Waterloo
Internal Members:	Ed Lank, Professor Cheriton School of Computer Science, University of Waterloo Mark Smucker, Associate Professor, Department of Management Sciences, University of Waterloo
Internal-External Member:	Oliver Schneider, Assistant Professor, Department of Management Sciences, University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This dissertation includes first-authored peer-reviewed material that has appeared in conference and journal proceedings published by the Association for Computing Machinery (ACM). The ACM's policy on reuse of published materials in a dissertation is as follows:[1]

> *\Authors can include partial or complete papers of their own (and no fee is expected) in a dissertation as long as citations and DOI pointers to the Versions of Record in the ACM Digital Library are included."*

The following list serves as a declaration of the Versions of Record for works included in this dissertation.

Portions of Chapter 4:
Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles L. A. Clarke. 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). Association for Computing Machinery, New York, NY, USA, 2187–2193.
DOI: https://doi.org/10.1145/3027063.3053175
Research presented in Chapter 4 was conducted at Emory University during the first author's stay there as a Visiting Scholar. Prof. Eugene Agichtein and Prof. Charles L.A. Clarke provided research guidance. Denis Savenkov provided assistance conducting the user study. All co-authors contributed in writing the manuscript.

Portions of Chapter 5:
Alexandra Vtyurina and Adam Fourney. 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). Association for Computing Machinery, New York, NY, USA, Paper 208, 1–7.
DOI: https://doi.org/10.1145/3173574.3173782
Research presented in Chapter 5 was conducted during the first author's internship at Microsoft Research, Redmond. Adam Fourney provided research guidance and participated in qualitative coding of participant utterances. All co-authors contributed in writing the manuscript.

[1] https://authors.acm.org/author-services/author-rights. Accessed in August 2020.

Portions of Chapter 6:
Alexandra Vtyurina, Charles L.A. Clarke, Edith Law, Johanne Trippas, and Horaţiu Bota. 2020. A Mixed-Method Analysis of Text and Audio Search Interfaces with Varying Task Complexity. In Proceedings of the 2015 International Conference on The Theory of Information Retrieval (ICTIR '20).
In press.
Research presented in Chapter 6 was conducted under research guidance of Prof. Charles L.A. Clarke and Prof. Edith Law in collaboration with Johanne Trippas and Horaţiu Bota. Horaţiu Bota provided his expertise on statistical analysis. All co-authors contributed in writing the manuscript.

Portions of Chapter 7:
Alexandra Vtyurina, Adam Fourney, Meredith Ringel Morris, Leah Findlater, and Ryen W. White. 2019. VERSE: Bridging Screen Readers and Voice Assistants for Enhanced Eyes-Free Web Search. In The 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19). Association for Computing Machinery, New York, NY, USA, 414–426.
DOI: https://doi.org/10.1145/3308561.3353773
Research presented in Chapter 7 was conducted during the first author's internship at Microsoft Research, Redmond. Adam Fourney, Meredith Ringel Morris, Leah Findlater, and Ryen White provided research guidance. Meredith Ringel Morris provided her assistance and expertise with hiring participants for the survey and the user study. Adam Fourney participated in coding survey responses. All co-authors contributed in writing the manuscript.

Abstract

Natural language interfaces have seen a steady increase in their popularity over the past decade leading to the ubiquity of digital assistants. Such digital assistants include voice-activated assistants, such as Amazon's Alexa, as well as text-based chat bots that can substitute for a human assistant in business settings (e.g., call centers, retail / banking websites) and at home. The main advantages of such systems are their ease of use and – in the case of voice-activated systems – hands-free interaction.

The majority of tasks undertaken by users of these commercially available voice-based digital assistants are simple in nature, where the responses of the agent are often determined using a rules-based approach. However, such systems have the potential to support users in completing more complex and involved tasks. In this dissertation, I describe experiments investigating user behaviours when interacting with natural language systems and how improvements in design of such systems can benefit the user experience.

Currently available commercial systems tend to be designed in a way to mimic superficial characteristics of a human-to-human conversation. However, the interaction with a digital assistant differs significantly from the interaction between two people, partly due to limitations of the underlying technology such as automatic speech recognition and natural language understanding. As computing technology evolves, it may make interactions with digital assistants resemble those between humans. The first part of this thesis explores how users will perceive the systems that are capable of human-level interaction, how users will behave while communicating with such systems, and new opportunities that may be opened by that behaviour.

Even in the absence of the technology that allows digital assistants to perform on a human level, the digital assistants that are widely adopted by people around the world are found to be beneficial for a number of use-cases. The second part of this thesis describes user studies aiming at enhancing the functionality of digital assistants using the existing level of technology. In particular, chapter 6 focuses on expanding the amount of information a digital assistant is able to deliver using a voice-only channel, and chapter 7 explores how expanded capabilities of voice-based digital assistants would benefit people with visual impairments.

The experiments presented throughout this dissertation produce a set of design guidelines for existing as well as potential future digital assistants. Experiments described in chapters 4, 6, and 7 focus on supporting the task of finding information online, while chapter 5 considers a case of guiding a user through a culinary recipe. The design recommendations provided by this thesis can be generalised in four categories: how naturally

a user can communicate their thoughts to the system, how understandable the system's responses are to the user, how flexible the system's parameters are, and how diverse the information delivered by the system is.

Acknowledgements

I would like to thank all people who helped bring this dissertation into existence. First and foremost, I would like to thank my PhD advisors Charlie Clarke and Edith Law for their support (work-related as well as emotional) and help navigating the waters of grad school. I would not be able to do this without you.

I would like to give my gratitude to the members of my examining committee: Max Wilson, Ed Lank, Mark Smucker, and Oliver Schneider. I look back on the day of my defence fondly – the deep discussions we had left me with a feeling of meaningfulness of my work, and the feedback you provided helped me significantly improve the dissertation. I am additionally thankful to Ed and Mark, whose advice and support helped me answer a find balance in life.

I am immensely greatful to my internship advisors who helped me view the world from a different perspective: to Eugene Agichtein – working together with you was an honour, and I will always think of my time in Atlanta with a smile; to Adam Fourney – developing VERSE together made me feel the rare sense of purpose and fulfillment, and working with you made my summers in Redmond a blast. I am greatful to all my co-authors who helped shape this work – Meredith Morris, Horaţiu Bota, Johanne Trippas, Denis Savenkov, Leah Findlater, and Ryen White. I feel very lucky to have had an opportunity to work side by side with you.

I would like to say thank you to all of the wonderful people I have had a pleasure to be around during these years. I trust that we helped keep each other's mental health in check and the caffeine levels consistently high. Bahareh Sarrafzadeh, my first co-author and teacher, thank you for leading me by example; Alex Williams, thank you for being an unwavering optimist and providing your guidance through the Starbucks drinks menu; Mike Schaekermann, thank you for all the laughter and beers we had together; Greg d'Eon and Blaine Lewis, thank you for the roads cycled, boulders climbed, and Rubik's cubes solved; Damien Masson, Nils Lukas, and Glaucia Melo, thank you for the most hilarious of discussions over coffee at C&D; Nathan Harms, thank you for maintaining my morale; David Maxwell and Johanne Trippas – your advice and encouragement were invaluable; Adam Roegiest and Kira Systems crew – thank you for the support and accommodations.

Thank you to my dear friend and co-author Horaţiu Bota for always being there for me. Thank you to my friends Marianna Rapoport and Abel Nieto for being with me through one of the darkest periods of my life and helping me emerge from it ever stronger. Thank you to Uzma Rehman for teaching me the art of self-acceptance. Thank you to Valentina Vaneeva and Alexey Kon for continuously inspiring me to climb higher, run longer, and

achieve more. Thank you to my wonderful friends Anya Sorokina, Katya Davydenko, Pasha Skornyakov, Kolya Shtanko, Dasha Mayorova, Anya TX, Evgeny Pasternaque for being the source of laughter and funny dog pictures.

Thank you to my family for believing in me throughout all these years. Thank you to my parents for teaching me perseverance and patience. Thank you to my sisters Masha and Ira for being my role models. Thank you to my grandmother Natalia who spent countless hours helping me with my school homework – I am happy to say it paid off.

Finally, to my loving partner Carmelo Mastrandrea – thank you for believing in me when I did not believe in myself.

Dedication

To Persei, the woof of my life.

Table of Contents

List of Figures xvi

List of Tables xviii

1 Introduction 1

1.1 Research Questions . . . 3

2 Background and Related Work 7

2.1 Conversation or Dialogue? . . . 7

2.2 Human Conversation . . . 8

2.3 Architecture of Dialogue Systems . . . 10

2.4 Adoption of Digital Assistants . . . 12

2.4.1 Perception of Digital Assistants . . . 14

2.5 Auditory Comprehension . . . 15

2.6 Voice Interfaces for Accessibility . . . 17

2.6.1 Accessing Web using a Screen Reader . . . 17

2.6.2 Novel Screen Reader Designs . . . 18

2.6.3 Voice-controlled Screen Readers . . . 19

2.6.4 Issues with Design of Voice Assistants . . . 20

2.7 Visual Interfaces for Web Search . . . 21

2.8 Search Task Complexity . . . 23

2.9 Chapter Summary . . . 24

3 Methodology 25
3.1 Wizard-of-Oz . . . 25
3.2 Workload Assessment . . . 27
3.3 Crowdsourcing . . . 29
3.4 Other Methods . . . 30
3.4.1 Controlled Experiment . . . 30
3.4.2 Usability Study . . . 30
3.4.3 Online Survey . . . 31
3.4.4 System Usability Scale (SUS) . . . 31

4 Exploring Conversational Search With Humans, Assistants, and Wizards 33
4.1 Motivation . . . 34
4.2 Study Design . . . 35
4.2.1 Wizard Agent . . . 36
4.2.2 Human Agent . . . 38
4.2.3 Automatic Agent . . . 38
4.3 Search Task Selection . . . 38
4.4 Results . . . 40
4.4.1 Overall Satisfaction . . . 40
4.4.2 Topical Quiz Success . . . 41
4.5 Qualitative ndings . . . 41
4.5.1 Conversational Context . . . 41
4.5.2 Trustworthiness of Information . . . 42
4.5.3 Social Acceptance . . . 42
4.6 Discussion and Design Implications . . . 42
4.7 Chapter Summary . . . 45

5 Exploring the Role of Conversational Cues in Guided Task Support with Virtual Assistants 46
5.1 Motivation . . . 46
5.2 Study design . . . 49
5.2.1 Apparatus . . . 49
5.2.2 Procedure . . . 49
5.3 Results . . . 51
5.3.1 General Impressions . . . 51
5.3.2 Types of User Utterances . . . 52
5.3.3 Explicit Requests . . . 54
5.3.4 Implicit Intents . . . 54
5.4 Implications, Limitations and Future Work . . . 57
5.5 Chapter Summary . . . 59

6 A Mixed-Method Analysis of Text and Audio Search Interfaces with Varying Task Complexity 60
6.1 Motivation . . . 61
6.2 Study Design . . . 62
6.2.1 Search Tasks . . . 63
6.2.2 Search Results . . . 65
6.2.3 Interfaces . . . 65
6.2.4 Procedure . . . 67
6.2.5 Participants . . . 69
6.3 Quantitative Findings . . . 69
6.3.1 Di erences in Ranking . . . 70
6.3.2 Perceived Workload . . . 73
6.4 Qualitative Findings . . . 74
6.4.1 Navigation Shortcuts . . . 74

6.4.2 Challenges with Audio Results Perception 75
6.4.3 Cognitive Load 77
6.5 Chapter Summary 78

7 VERSE: Bridging Screen Readers and Voice Assistants for Enhanced Eyes-Free Web Search 80
7.1 Motivation 80
7.2 Online Survey 81
7.2.1 Survey Design and Methodology 82
7.2.2 Participants 82
7.2.3 Findings 84
7.3 VERSE 89
7.3.1 Overview 89
7.3.2 Example Usage Scenario 90
7.3.3 VERSE Design Elements 92
7.4 Design Probe 95
7.4.1 Procedure 96
7.4.2 Participants 97
7.5 System Usability 97
7.6 Participant Feedback 98
7.7 Discussion 100
7.8 Chapter Summary 102

8 Discussion 103
8.1 Design Recommendations for Voice-based Dialogue Systems 104
8.2 Future work 106
8.2.1 Continuous Interaction 106
8.2.2 Customisation of Parameters 107

8.2.3 Comparison of Command-based with Conversation-like Approaches 107
8.2.4 Universal Interface 107
8.2.5 Parallels with Visual Iinterfaces for Search 108
8.2.6 Voice Interface for Driving 109
8.3 Chapter summary 113

9 Conclusion 114
9.1 Summary of Contributions 114

References 116

List of Figures

2.1 An example of Je erson transcription 9
2.2 Screenless smart speakers. 13
2.3 Smart speakers with a visual display. 14
2.4 Voice assistant Siri reverts to showing information on the screen instead of producing auditory output. 16

4.1 Three conversational agents used in the study. 37
4.2 Failure to maintain conversational context leads to the reformulation of the question. 42
4.3 Examples of user behaviour when interacting with Automatic and Human agents. 43

5.1 Example of implicit conversational cues used by a participant while interacting with a voice assistant. 48
5.2 Recipe used in the study. 50
5.3 Distribution of utterance categories across participants. 52
5.4 Distribution of implicit and explicit utterances across participants. 53
5.5 Example of a user showing readiness to proceed by repeating previous system response. 55
5.6 Example of acknowledgement by the user. 57
5.7 Example of a user repeating the response while completing the step. 57
5.8 Example of a user con rming an existing belief about a recipe step. 58
5.9 Example of a user asking for clari cation on the previous response 58

6.1 Text and audio interfaces used in the study. 65
6.2 Average di erence in the number of consistent result choices between text and audio interfaces. 71
6.3 Comparison of result selection in two studies. 72

8.1 Demonstration of the study setup proposed for the study. 112

List of Tables

1.1 Diagram of the thesis structure. 5

2.1 Task complexity from the Taxonomy of Learning Objectives [16]. 24

3.1 Diagram of methods used in this thesis. 31

4.1 Description of search tasks used in the study. 39
4.2 Comparison of three agents used in the study. 40

5.1 NASA TLX and SUS scores of culinary assistant. 51
5.2 Distribution of user utterance from di erent categories. 55

6.1 Search tasks used in the study 64
6.2 Audio search result is generated by concatenating parts of a corresponding text search result. 67
6.3 Participants characteristics in two studies. 69
6.4 Results of NASA-TLX questionnaire inLAB study. 74

7.1 Demographic characteristics of survey respondents. 83
7.2 General patterns of voice assistants use. 86
7.3 Mapping of voice commands and corresponding gestures in VERSE. 91
7.4 Example usage scenario of VERSE 94

Chapter 1

Introduction

Popular culture has thoroughly explored the idea of communication with digital assistants using natural language, with HAL 9000 of \2001: A Space Odyssey" and TARS of \Interstellar" providing just a few of many examples. Continued technological improvements have now made this idea a reality { we can nally communicate to computers much like we do with fellow humans. Increasingly, natural language interfaces, and voice interfaces in particular, are gaining popularity. For example, a survey conducted by Google in 2014 revealed that out of 1400 people, 55% of teenagers and 41% of adults used voice search at least once a day [78]. A 2016 study by Ido Guy showed that search queries submitted over voice resemble natural language questions more so than search queries submitted over text [84].

Advances in technology also promoted the popularity of digital assistants, such as Apple's Siri, Amazon's Alexa, Microsoft's Cortana, and Google Assistant. Digital assistants rst emerged with the release of Siri by Apple in 2011 as an iPhone-based application. At that point a mixed-modality paradigm was presented: Siri could be activated by voice by uttering a wake word, like \Hey Siri". It would attempt to answer user's request through voice, and in case of falling short of completing the request, Siri would revert to displaying visual information on the phone screen. Figure 2.4 illustrates an example of an exchange where a user's request is addressed by displaying a list of search results on the screen. The number of such digital assistants worldwide will grow from 2.5 billion today to 8 billion by 2023, according to estimates by Juniper Research [138].

Over time, digital assistants were programmed to ful l more and more user commands using voice only, until eventually the concept of a smart speaker was introduced. Amazon's Echo and Google Home are examples of these and are shown in Figure 2.2. A smart-speaker

based digital assistant is still paired with a screen-based device, such as a smart phone or a tablet, where visual information can be displayed, but the majority of the interactions are meant to be conducted using voice only. The popularity of smart speakers has been and is continuing to grow consistently. The Nielsen agency reported that in the second quarter of 2018, 24% of US households owned a smart speaker { a 2% increase compared to the rst quarter of 2018 [132]. In January 2020, Voicebot.AI presented the results of a survey con rming that over a third of US adult population { nearly 90 million people { owned a smart speaker [40].

In this thesis, I focus on digital assistants. Strictly speaking, programs like Siri and Alexa provide their users with an opportunity to formulate commands in natural language and issue them through voice, by speaking out loud, and receiving an audio, or sometimes text, response. This process often mimics normal human conversation, therefore a term \conversational assistants" has been coined to describe this technology. A trending example of application of digital assistants is \conversational search" { a process which presumes multiple exchanges between a system and a user in order to retrieve relevant information from the underlying document corpus (or in a more general scenario, the internet). However, it has been argued that the term \conversation" assumes a deeper interaction level than that presented by the current technology and that an interaction between a user and a digital assistant is \conversation-like". I agree with this argument, and consider the term \conversation" not applicable to currently available systems. Instead, I prefer to use the broader term \dialogue". The downside of this term is its breadth: many di erent processes may be considered a dialogue { issuing text commands in a computer console, using graphical user interfaces, as well as issuing voice commands to a digital assistant.

There are a number of ways to communicate with a digital assistant, with text (typing commands) and speech (speaking commands out loud) being the major ones. Throughout this thesis, I will use terms \text-based assistants" and "voice-based assistants" or \voice assistants" to indicate the primary modality in which an assistant operate, whereas the term \digital assistant" will be used to denote a modality-agnostic assistant. I will focus on studying interactions with voice-based assistants, except for the experiments reported in chapter 4 where a text-based exchange with an assistant was used to simplify the experimental setup.

Though currently available voice assistants are a long way away from the functionality (and malice) of HAL 9000, they are highly integrated with smart home devices, enabling their users to control lights, thermostats, door locks, etc. using voice-only commands. Multiple studies found that voice assistants are used for short simple tasks such as setting up alarms clocks, timers, looking up answers to factoid questions [123, 116] with a noticeable fraction of voice assistant users using them for the sake of entertainment and amusement.

Moreover, the integration of voice assistant with smart home devices and the voice-based interaction they provide, serves as an additional layer of accessibility and provide added independence to people with disabilities [5, 142].

Voice assistants are excellent tools made possible by mixing a number of recent technological advances. However, they are not without their limitations. Many marketing campaigns position voice assistants as a friend in a box that you can \just talk to". This creates in ated user expectations which lead to underwhelming experience [116]. After the rst period of curiosity and experimentation, the voice assistant usage tends to taper o . Users grow disillusioned by the lack of the assistants' ability to understand their intent, users identify language that works and stick with it to get reliable responses. Continued technological advancements are likely to improve problems like speech recognition and intent identi cation. My rst focus in this thesis is to investigate how user perception of digital assistants could change if they were not limited by technological capabilities.

However, even with in ated expectations, voice assistants have proven to be highly popular among various user groups. The leading reason for people using voice-based assistants is hands-free interaction and therefore the opportunity for multitasking they provide [116]. Some use cases include assistance during cooking process, lounging on the couch while choosing a TV channel, and others. A scenario mentioned above, where a user is redirected from voice-based interaction to a screen-based interaction terminates the voice-based exchange and breaks down the intended ow. Presenting information through an audio-only channel is essential for facilitating seamless and complete voice-based interaction. Therefore, the second goal of this thesis is to investigate ways in which voice assistants can expand their voice-only interaction using currently existing underlying technology.

1.1 Research Questions

This thesis poses two main research questions:

- RQ1 : How would users perceive digital agents that could understand them as well as their fellow humans?
- RQ2 : How can we improve interaction using currently available tools?

In chapters 4 and 5, I tackle RQ1 and explore \the systems that could be" by imitating them and studying how users react and communicate back to such agents. More speci cally, I ask:

- RQ1-a : Given equal performance, would people choose to communicate with another person or a digital assistant and why?
- RQ1-b : What opportunities are opened by the language people use?
- RQ1-c : What aspects of digital assistant design are important to consider?

I tackle RQ2 in chapters 7 and 6, where I experiment with systems built using available tools and address the question of how they can be improved. In particular, I examine how the process of web search can be conducted using primarily voice for general population and for people with visual impairments. Therefore I pose the following sub-questions:

- RQ2-a : Does the medium (text/audio) over which search results are delivered a ect the user's search result preference?
- RQ2-b : What aspects of audio-based search results are important for the accurate assessment of relevance by the user?
- RQ2-c : How might voice assistants and screen readers be merged to confer the unique advantages of each technology?

At each step of the investigation, I discovered aspects of system design that emerged repeatedly and played a role in how the interaction with a digital assistant is assessed by users. I consider these design elements essential in developing future digital assistants. By consolidating the results from di erent user studies, I developed the following thesis statement:

> A productive interaction with a dialogue system critically depends on how naturally a user can communicate their intent to the system, the understandability of the system's responses, the exibility of the system's parameters, and the diversity of information accessible through the system.

I describe each of the aspects in details below while table 1.1 indicates the thesis chapters in which each of the aspects is addressed.

	Chapter 4	Chapter 5	Chapter 6	Chapter 7
User's ability to communicate their intent to the system	x	x	x	x
Understandability of system's responses	x		x	
Flexibility of parameters	x		x	x
Diversity of information	x			x

Table 1.1: Diagram of the thesis structure. Each chapter covers at least one interaction aspect.

User's ability to communicate their intent to the system. Naturalness of interaction does not necessarily imply natural language interaction, rather, how an agent's functionality matches users' mental model of it. Here, I note that certain features used in a human-to-human conversation can be used to enhance the naturalness of the interaction between an agent and a user. One such feature is contextual awareness. In chapters 4, 6, and 7, I illustrate that the ability of an agent to keep the memory of the information mentioned previously (by both the user and the agent itself) is a positive aspect. For example, in chapter 4 persistent memory of past interactions allowed participants to forego formulating full self-su cient questions at every step and instead allowed them to use partial questions relying on previously appearing information. Furthermore, in chapter 7, participants expressed the desire to be able to use the content of an article for navigation, and in chapter 6, many participants implicitly echoed this sentiment by using content words to talk about search results they heard. I also argue that if presented with an opportunity, users will likely take advantage of the functionality borrowing from the human-to-human conversation. To this point, in chapter 5, I demonstrate that most participants did not exhibit reluctance to employ language characteristic of human-to-human communication when interacting with a digital agent.

Understandability of system's responses. Another part of designing a smooth interaction is providing the information back to the user in the form that is easily understood and interpreted by the user. The requirements for aspect may di er depending on the use case. For example, in chapter 4, information source was required to interpret the credibility of the information. However, in chapter 6, I present a di erent outlook, where the understandability of agent's responses implies incorporating pauses and tones into an audio

response.

Flexibility of parameters. To accommodate a variety of user groups as well as individual users, a system ought to o er a number of settings which can be changed as desired. Throughout the studies described in this thesis, I found that aspects such as the response length returned by the agent (chapter 4), the pitch and speed of voice used to generate auditory responses (chapters 6, 7) are based on preferences. Interestingly, while certain parameters such as answer length can be individual, others, such as speech rate, can be speci c to a user group { people with visual impairments will likely prefer audio played at a faster rate compared to sighted people.

Diversity of information. Finally, I point out that access to a variety of information is bene cial to the agent's design. In chapter 7, one of the agent's positive aspects was its ability to return information from multiple search parallels (e.g. news and Wikipedia). In a similar vein, participants from chapter 4 expressed a desire to have the agent provide results from opinion aggregating websites in addition to its regular search capabilities.

The remainder of this thesis is structured as follows. In chapter 2, I discuss prior work that impacted, inspired or otherwise in uenced my research. I also outline and provide background on the research methods used throughout this dissertation. In chapters 4 and 5, I describe the work focusing on future dialogue systems design. User studies in both chapters are based on simulated agents and are aimed at exploring the behaviour of the users interacting with the agents. In chapters 6 and 7, I explore the design of dialogue systems that can be built with currently available tools. In this exploration, I take two di erent angles. First, in 6, I investigate the caveats around designing a fully audio-based dialogue system for web search. I outline the key principles that should be followed when presenting search results in the absence of a screen and compare them with existing recommendations for visually displaying web search results. Afterwards, in chapter 7, I explore a similar problem { designing an audio-only dialogue system for web search, however, this time, I focus on designing speci cally for people with visual impairments. Finally, I discuss the ndings in chapter 8 and outline possible directions for future work in chapter 9.

Chapter 2

Background and Related Work

In this chapter, I discuss key prior and related work that informed, inspired, and impacted the research presented in this dissertation. I begin by discussing the terminology often applied to describe digital assistants, highlighing the di erences between human-to-human conversation and human-machine dialogue. I continue by describing standard architectures for digital assistants, followed by common use cases and attitude of users toward digital assistants. I continue by describing the role voice assistants play for people with disabilities in general, and visual impairments in particular. Finally, I outline the work done in the area of designing of speech and text interfaces for web exploration and search, as well as search tasks typically used in user studies.

2.1 Conversation or Dialogue?

Major commercially available systems operate based on prescribed scenarios and are able to function as long as the user's input is within certain prede ned boundaries but are otherwise quite fragile. However, in an attempt to appease the buyers, these commercial systems are designed to mimic certain super cial features of a human conversation. For example, a command for checking the weather can be phrased in a variety of di erent ways: from \What is the weather forecast?" to \Do I need an umbrella today?". Furthermore, the voices of digital assistants become increasingly similar to those of humans. The 2018 Google Duplex project demonstrated an assistant making a reservation at a restaurant by talking to a person [3]. Such imitation has downsides as well as merits. On one hand, this imitation produces high user expectations who may be quickly underwhelmed after the beginning of the interaction [116]. On the other hand, it leads to an anthropomorphisation

of such systems { a trait that has been found bene cial in certain medial applications, in particular for mental health [180, 144].

While digital assistants do mimic certain aspects of a human-to-human conversation, an interaction between a person and a computer system can be considered \conversation-like" but it is not a \conversation" in and of itself [140]. Language is something most of us learn from the young age and conversation is the go-to way of interaction using language. A conversation is considered to be one of the most seamless and intuitive ways to conduct an e cient interaction. Herbert Clark, one of the prominent researchers in the area of conversational analysis, suggests that a human conversation is an innately collaborative act without a prede ned outcome where each party contributes independently towards a mutual goal [54]. A human conversation encompasses a multitude of parameters making it di cult if not impossible to model with available tools [14].

In this thesis, I avoid the term \conversation" to describe an interaction between a human and a computer system. Instead, I use the term \dialogue" throughout this thesis to describe such interaction. The term \dialogue system" is well suited to describe the type of interaction we consider. However, it is also quite broad, encompassing a wide range of interactive system, such as most graphical user interfaces, a computer terminal, a telephone ight booking systems, and even an elevator. The terminology chosen in this thesis, does di er from that of a large body of research. However, even in prominent HCI and IR venues, it is acknowledged that the systems called \conversational" are currently operating in a limited capacity, for example, \conversational search" systems are mostly functioning as speech-based question-answering systems [61]. Perhaps the usage of the term conversation in this context comes from the desire to one day achieve a truly conversational experience with a computer, rather than re ecting the current capabilities of these systems. In the next sections, I outline the main traits of a human-to-human conversation and discuss architecture of many commercial digital assistants.

2.2 Human Conversation

Before we can teach a computer how to converse like a human, we should obtain a thorough understanding and formalization of how humans converse with each other. A eld of studies called conversational analysis is devoted to exploring the rst step of this process. The process of conversational analysis begins with collecting a large dataset of conversations (audio or video) and transcribing them in great detail, known as Je erson transcription [95] and shown in Figure 2.1. Such transcription includes such features as precise timestamps up to one tenth of a second, indications of pauses, intonation, volume, body language, etc.

Figure 2.1: An example of transcription made by Gail Je erson [95] outlining minute characteristics of speech including intonations, pauses, speech overlaps, etc.

After completing the transcription, researchers iteratively analyse the conversations noting the recurring patters and anomalies. Finally, scientists attempt to generalise and formalise the discovered patterns, leading way to understanding the structure of a conversation.

One of the pioneers of conversational analysis, Harvey Sacks, posited that in its basis, a conversation is an sequence of turns [155]. Sacks and colleagues proposed a model of a conversation having 14 characteristics, including:

- speaker change recurs, or at least occurs,
- turn order is not xed but varies,
- overwhelmingly, one party talks at a time,
- what parties say is not speci ed in advance,
- turn size is not xed but varies,
- length of conversation is not speci ed in advance,
- number of parties is not speci ed in advance,
- relative distribution of turns is not speci ed in advance.

A conversation contains a variety of other moving parts. As mentioned before, Herbert Clark postulated that grounding is necessary for a successful conversation, i.e. the speakers must agree on the common ground in order to continue the conversation [53]. Clark posited that a conversation is a collaborative act in which people must share information and mutual beliefs in order to continue the conversation. Furthermore, Marilyn Walker, investigated the shift in the initiative during a conversation { depending on the type of the conversation it may belong to a single speaker (e.g. an interview) or be mixed (e.g. a regular conversation) [182]. One of the additional complexities of spoken conversation analysis lays in the choice of units. Written text analysis can be conducted by splitting the text into sentences, a frequent unit of analysis is an utterance { in a dialogue, speakers usually exchange utterances. The complexity, however, is in the fact, that turns, and therefore utterances may overlap, one utterance can span multiple turns. In this thesis, the term \utterance" will be used to denote an uninterrupted block of speech from one partner, or interlocutor.

In the spirit of studying human-to-human conversations in order to model human-computer dialogue, Thomas et al. [175] and Trippas et al. [176] created spoken conversational search datasets, in which they observed how two individuals communicated over an audio-only channel to complete web search tasks. Both papers illustrate how people would approach web search through a conversation if the current technical limitations could be ignored. Prior to this work, Radlinsky and Craswell [147] presented a theoretical model for a conversational search system where they presented a detailed interaction flow and desired system's functionality to deliver a satisfactory user experience.

2.3 Architecture of Dialogue Systems

Digital assistants, and more broadly dialogue systems, can be implemented using two main approaches: corpus-based and rules-based. In a corpus-based approach, a system takes advantage of large datasets of prior conversations or exchanges, often between two people. Such datasets may include human-human phone conversations, movie scripts, exchanges between people on online board and forums, chains of tweets. Some such datasets may be created with a particular goal in mind, and crowdsourcing technique may be used to create these datasets.

Corpus-based systems often produce their responses based on the user's prior input either by finding the most similar message, or using encoder-decoder machine learning model to generate a response based on the underlying database. One may guess that such methods would perform poorly in completing a task, such as booking a table at a

restaurant for a particular day and time. Corpus-based systems may be best suited to chatbots { often text-based systems with the main goal of maintaining the interaction for as long as possible.Some chatbots are created for a speci c task, for example locating a speci c item in an e-commerce store, and in this case the corpus-based approach will not be be suitable. Another caveat for corpus-based systems is that they will re ect the data they have seen before. For example, a Twitter bot called Tay, released by Microsoft, learned from the responses of other Twitter users and had to be taken down after a 16 hours online. This was a result of Twitter users tweeting o ensive messages at Tay for the sake of entertainment. The bot soon picked up the unfortunate patterns. This phenomenon echoes a broader issue with training data used to produce machine learning models: Biased data produces biased models which in turn produce biased outcomes [43].

Another approach to modelling an interaction is rules-based. ELIZA [183] and PARRY [57] were the pioneer rules-based text-based systems. Both of them were used in the eld of clinical psychology and both were based on regular expression rules. Interestingly, PARRY was the rst chatbot to pass a Turing test { study subjects could not con dently distinguish transcripts of conversations between two people from transcripts of exchanges produced from an interaction with PARRY.

Modern digital assistants aim at assisting users with a variety of tasks, making it exceedingly di cult to model interactions with regular expressions. To make these interactions as stable as possible and to avoid breakdowns, people designing these digital assistants must make sure that an assistant is capable of handling various scenarios and user input within the boundaries of a given task. Upon receiving user input, the system rst attempts to classify the domain of user request, for example setting up an alarm or booking a ight. If a system is designed to operate within the boundaries of a single domain, this step is not completed. After identifying the general area of interest, the system attempts to identify a speci c user intent, for example looking up ight schedule or cancelling a previously booked ight. Each intent is usually paired with a set of slots { variables that are required to be lled in to complete the task. For example, if the system identi es that the user is looking to book a ight, the slots may include ight origin and destination, departure and return dates, preferred departure and return time and so on. Upon identifying the intent, the system attempts to extract some of the values from user's original input and then proceeds asking questions to ll the rest of the slots. For examples, if the user's original request was \I would like to book a ight from Toronto to Vancouver", the system may identify the intent to book a new ight and extract the \Toronto" as an origin and \Vancouver" as a destination. It may then proceed to ask the user about the preferred dates of the ight and whether the ight is meant to be one-way or return. The identi cation of user intents as well as slot value extraction is a product of machine learning models which are trained

to be able to process various phrasings.

2.4 Adoption of Digital Assistants

Digital assistants are known by a number of names in the literature: voice-activated personal assistants (VAPAs) [5], intelligent personal assistants (IPAs) [116], personal digital assistants (PDAs), and voice-activated digital assistants (VADAs). The wide adoption of digital assistants started with the release of Siri as a standalone iOS app in 2010. The application was acquired by Apple soon after, and in just over a year, Siri was integrated in the operating system of iPhone 4S. Since then, voice-activated digital assistants have been gaining popularity, and soon were given their own homes in specialized devices.

The year 2014 marked the beginning of the smart speaker era with the release of Amazon Echo { a smart speaker powered by Alexa { a digital assistant developed by Amazon. The trend was quickly followed by many major tech corporations: Google announced a Google Home [63], Microsoft partnered with Harman Kardon to release Invoke powered by Cortana [164], Apple relocated Siri from an iPhone to a Home Pod [136], Samsung has been expected to unveil their smart speaker powered by Bixby [49]. Yandex released Yandex station with assistant Alice [110, 194].

Figure 2.2 shows Amazon Echo and Google Home { two examples of screenless smart speaker powered by digital assistants. Though the majority of smart speakers do not have a screen, some, such as Amazon's Echo Show come with a display, presumably to overcome the challenge of presenting rich information in an audio-only format as well as to enable additional functionality such as video calls [163, 165]. Nielsen reports that in the second quarter of 2018, 24% of US households owned a smart speaker { a 2% increase compared to the rst quarter of 2018 [132].

According to an interview study of voice assistant users [116], one of its most appealing aspects is providing the ability to engage in hands-free interaction and multitasking, i.e., engage in another primary activity while keeping a conversation with a voice assistant as a secondary task. This use case comes to shine when a user's eyes and/or hands are engaged elsewhere (for example, walking or driving) [116]. A number of studies investigated what people use voice assistants for. It was found that voice assistants are often used to complete simple atomic tasks, such as controlling smart home appliances, music playback, setting timers and alarms, checking the weather are among the top categories [15, 26, 123, 179, 116, 106, 26].

Figure 2.2: Screenless smart speakers: Amazon Echo and Google Home

In designing voice assistants,companies use the state-of-the-art text-to-speech technology, and more generally, pursue the strategy of mimicking a human-to-human conversation. In an e ort to do so, a number of design tools o er a way to develop interactions based on slot- lling approach, whereby machine learning models are employed to categorise user inputs and synthesise appropriate system output[1]. This tactic may be a double-edged sword, as the increasingly natural speech bolsters users' expectations on the functionality of voice assistants, which at the moment is not nearing the level of human-to-human conversation [116].

Another major concern regarding voice assistants is related to Automatic Speech Recognition, or ASR [101, 124, 116]. In particular, problems may arise when recognizing names [170], accents [116], and children's speech [113]. Troubleshooting problematic interaction has been proven to be be challenging, if not impossible.

To illustrate the di culties in interaction, one can consider an information nding task as an example. Information requests are reported to constitute a major portion of user interactions with voice assistants [26, 106, 15, 123]. Requests for information can span a number of user intents: asking single-faceted factoid questions, such as \When was Albert Einstein born?". Such questions are typically answered by the assistants with high accuracy, producing a single phrase as an answer, e.g. \Albert Einstein was born on March 14 in 1879". However, multi-faceted queries, queries that are ambiguous, queries that can not be answered with a single sentence or phrase, or queries require deeper research are usually where voice assistants may all fall short. A common strategy is to read back a top-ranked paragraph, or an excerpt from Wikipedia, and when prompted for more information

[1] https://developers.google.com/assistant

Figure 2.3: Smart speakers with a display for convenient information visualisaion.

there is not much wiggle room. At the time of this writing, Alexa and Google Assistant o er to read more information from the source. When given a more complicated or ambiguous query many assistants will respond with \I cannot answer this right now" or \I searched the web for you, look for the results in the phone app". Both outcomes abruptly end the user interaction leaving a user with no recourse in the rst instance, or forcing them to engage with a visual interface in the second. Such experiences disrupt the user's voice experience, forces the user to switch modalities, and can understandably can cause user frustration.

2.4.1 Perception of Digital Assistants

Though increasingly popular, digital assistants have not yet become a social norm yet. Much work has been done in the area of comparing user interactions with a human and a computer. There are varying opinions on the subject. Edwards et al. [68] found no signi cant di erences in how Twitter users treated a social bot, whether it was perceived as a human or not. In turn, Cɐment and Guitton [56] report that the way bots are perceived varies with the role they play. They found that \invasive" Wikipedia bots received

more \polarizing" feedback { both positive and negative { compared to the bots that carried out \silent helper" functions. The similar result is reported by Murgia et al. [130] { Stackover ow bot receives more negative feedback for false answers when its identity as an automatic program is revealed. Another work by Aharoni and Fridlund [9] reports mixed results from participants who underwent a mock interview with a human and an automatic system. The authors report that there were no explicit di erences in the interviewer perception described by the participants, although the authors noticed signi cant di erences in people's behaviour { when talking to a human interviewer they made greater e ort to speak, smiled more, and were more a ected by a rejection.

2.5 Auditory Comprehension

Human auditory system is an extremely complex mechanism comprising many moving parts: echoic memory is used for a very short-term sound storage, ascending pathways from the ear to the brain [21]. Sound processing occurs over time and di erent processes occur at di erent timestamps. At 20ms a phoneme is recognised, syllabic stress requires 200ms (\melody" vs \melodic"), while 1-2 seconds are needed to recognise the intonation of the sentence, and determine whether a question is being asked.

Every sounds possesses a set of physical characteristics which can be accurately measured. These characteristics include frequency (or pitch) and intensity (or loudness). However, there is more to sound than these physical characteristics. For example, one qualitative feature of a sound istimbre, or a quality of sound. Timbre lets us distinguish between sounds produced by di erent means even if they have the same frequency and intensity (for example, the same note generated by a saxophone or a piano). Timbre is a quality that lets us distinguish between di erent voices. However, even the \physical" characteristics turn out to carry a qualitative avour to them, since the comprehension by two individuals of the same sound may vary depending on their training (an opera singer can distinguish slightest pitch variations, while some people have a higher tolerance to louder noises). As such, while sounds have objective measurable physical characteristics, it is important to remember that sound perception varies depending on the individual hearing it.

One important di erence between the auditory and visual perception is spatial locality. Vision is inherently spatial. While sound may convey spatial information about it's source | for example, the intensity di erence between the ears can signal direction of the sound { is not primarily a spatial signal, but rather a temporal signal. While visual signals inherently carries spatial information about objects around us, auditory signals is

inherently temporal, \with auditory input changing over time" [190], such that a person cannot take a \second look" as one can in reading [90].

While studies of speech perception started in the early 1940s, it is still a relatively new area of investigation. One thing is clear { speech perception and processing involves mechanisms di erent from regular sound processing. The scienti c community cannot come to an agreement on a basic \building block" of speech perception. Simple sound-to-phoneme mapping theory fails to describe this process, because sounds with di erent physical characteristics can be easily mapped to the same phoneme. People have no problem processing speech produced using di erent loudness (whisper), pitches (a man or a baby).

Shannon et al. [168] conducted a \temporal envelope" experiment, in which he and colleagues showed that deteriorated speech, stripped of the frequency values could be easily recognised by the participants. Following this discovery, Binder et al. [32], showed that similar brain activity was incurred by word and non-word sounds, making it likely that the processing of speech is a hierarchical process.

Figure 2.4: Voice assistant Siri reverts to showing information on the screen instead of producing auditory output.

Additionally, the plasticity of our brain allows to reuse visual cortex to process audio signals, thus increasing the characteristics of audio comprehension in people who consume large volumes of information through an auditory channel. Bragg et al. [37] conducted a large-scale study of human listening rates, nding that the mean of intelligible speech was 309 WPM (words per minute) { notably faster compared to the average speaking rate 200WPM [198] (the estimates vary quite a lot from 120 to over 200wpm). The study additionally noted that factors such as age, native language, and sightedness signi cantly impacted the listening rate. The study exposes a wide variability in the listening abilities throughout the population, suggesting that di erent experiences need to be tailored to suit each category.

Studies suggest that people consume information di erently depending on whether they

see it or hear it. After conducting a user study with a voice-based application for simple tasks such as checking email and retrieving a weather forecast, yankelovich1995designing concluded that information in voice interfaces should be organised di erently from visual ones. They found that vocabulary, information organization and ow may not translate well between two mediums.

2.6 Voice Interfaces for Accessibility

Though not originally designed for this purpose [129], voice assistants found a wide acceptance in the community of people with disabilities. They were found to provide an additional, or as Pradhan and colleagues put it, \accidental", layer of accessibility [142]. Voice assistants have been noted to support people with visual [7], motor [73], and cognitive impairments [23]. Additionally, prior research showed how they can be used to aid entertainment of young children [75, 114] and support older adults [141, 193, 167]. However, there are still unaddressed challenges in the design of commercial assistants. In this section, I will focus on the obstacles faced by people with visual impairments.

2.6.1 Accessing Web using a Screen Reader

To access visual information many people with visual impairments use screen readers { software that converts visual information to auditory using synthetic speech. Braille display is another tool that serves a similar purpose. It uses tactile feedback to duplicate the text on screen and is essential for deaf-blind population. Each operating system has a built-in screen reader: Narrator on Windows[2], VoiceOver on iOS[3] and OSX[4], TalkBack on Android[5]. Additionally, NVDA [6] and JAWS[7] are stand-alone screen readers that can be freely accessed or purchased. A number of studies exposed a variety of challenges faced by people with visual impairments on the Internet [27, 131, 107, 156, 128, 30].

While guidelines are in place for accessible web design (WCAG[8]), similar accessibility problems persist on the Internet as websites become increasingly more complex and adopt

[2] https://support.microsoft.com/en-ca/help/22798/windows-10-complete-guide-to-narrator
[3] https://www.apple.com/ca/accessibility/iphone/vision/
[4] https://help.apple.com/voiceover/mac/10.14/
[5] https://support.google.com/accessibility/android/answer/6283677
[6] https://www.nvaccess.org/
[7] https://www.freedomscienti c.com/Products/Blindness/JAWS
[8] https://www.w3.org/WAI/standards-guidelines/wcag/

new technologies [85]. To illustrate, in 1999 Jonathan Berry outlined found that a lack of text descriptions for images \excluded" screen reader users from accessing the Web and poor page design led to frustration and getting lost. Nearly two decades later, problems remain. Lack of text descriptions for pictures is a widespread issue [77, 31, 160] and one that causes the most user frustration [107].

The \exclusion" aspect is further emphasised in comparison studies between people with visual impairments using screen readers and sighted participants. People with visual impairments were found to spend more time and e ort on web-based tasks compared to sighted participants [156, 29]. To gain access to inaccessible content, expert screen reader users also employed advanced techniques [31, 29], including using multiple screen readers, accessing HTML source code, probing (clicking on the link to quickly return back), using search within a page to reach otherwise inaccessible content [156, 29, 31, 34].

Web search engines pose additional unique challenges to screen reader users. Sahib et al. [156] found that blind users may encounter problems at every step of information seeking, and showed lower levels of awareness of some search engine features such as query suggestions, spelling suggestions, and related searches, compared to sighted users. Although these features were accessible according to a technical de nition, using them was time consuming and cumbersome [139]. Likewise, Bigham et al. [29] found that blind participants spent signi cantly longer on search tasks compared to sighted participants, and exhibited more probing behaviour (i.e., \a user leaves and then quickly returns to a page" [29]) showing greater di culty in triaging search results. Assessing trustworthiness and credibility of search sources can also pose a problem. Abdolrahmani et al. [6, 4] found that blind users use signi cantly di erent web page features from sighted users to assess page credibility.

2.6.2 Novel Screen Reader Designs

Another problem with screen reader web access are pages overloaded with banners, ads, menus, as well as unclear structure of the content that makes the task of navigating the web time consuming and strenuous [74, 35, 117]. One approach to overcome this problem is to segment pages into semantically sound sections, and provide access to those sections as opposed to individual HTML elements. Such segmentation can be accomplished based on linguistic [117], visual [83], and hybrid features [92]. Such high-level segmentation reduces the e ort and time necessary to navigate web pages using screen readers, and is generally found appealing by users.

Traditional screen readers provide sequential access to web content. Stockman et

al. [172] explored how this linear representation can mismatch the document's spatial outline, contributing to high cognitive load for the user. To mitigate this issue, prior research has explored a variety of alternative screen reader designs [148], which is brie y outlined below.

One approach is to use concurrent speech, where several speech channels simultaneously vocalize information [82, 201]. For example, Zhu et al.'s [201] Sasayaki screen reader augments primary output by concurrently whispering meta information to the user.

A method for non-visual skimming presented by Ahmed et al. [10] attempts to emulate visual \glances" that sighted people use to roughly understand the contents of a page. Their results suggest that such non-visual skimming and summarization techniques can be useful for providing screen reader users with an overview of a page.

Khurana et al. [99] created SPRITEs { a system that uses a keyboard to map a spatial outline of the web page in an attempt to overcome the linear nature of screen reader output. All participants in a user evaluation completed tasks as fast as, or faster than, with their regular screen reader.

Another approach, employed by Gadde et al. [74], uses crowdsourcing methods to identify key semantic parts of a page. They developed DASX { a system that transported the users to the desired section using a single shortcut based on these semantic labels; as a result, they saw performance of screen reader users rise signi cantly. Islam et al. [92] used linguistic and visual features to segment web content into semantic parts. A pilot study showed such segmentation helped the user navigate quickly and skip irrelevant content. Semantic segmentation of web content allows clutter-free access, at the same time reducing the user's cognitive load.

2.6.3 Voice-controlled Screen Readers

Prior work has also explored the use of voice commands to control screen reader actions. Zhong et al. [200] created JustSpeak { a solution for voice control of an Android OS. JustSpeak accepts user voice input, interprets it in the context of metadata available on the screen, tries to identify the requested action, and nally executes this action. The authors outline potential bene ts of JustSpeak for blind and sighted users. Ahok et al. [18] implemented CaptiSpeak { a voice-enabled screen reader that is able to recognize commands like \click <name> link," \ nd <name> button," etc. Twenty participants with visual impairments used CaptiSpeak for the task of online shopping, lling out a university admissions form, nding an ad on Craigslist, and sending an email. CaptiSpeak was found to be more e cient than a regular screen reader. Both JustSpeak and CaptiSpeak reduce the

number of user actions needed to accomplish a task by building voice interaction into a screen reader. Chapter 7 investigates a complementary approach, which adds screen-reader-inspired capabilities to VAs, rather than adding voice control to screen readers.

Screen readers usually accept keyboard-based (or gesture-based on touch screens) input and provide audio feedback to the users. Screen readers supply the users with a number of shortcuts or gestures ranging from simple ones for basic features to more complicated ones for advanced features. As such, learning screen reader functionality has a steep learning curve. Adopting new complex technology such as a screen reader can be di cult for people who lost their sight later in their lives. To alleviate this issue, researchers have explored screen readers accepting voice-based input from users. For example, Capti-Speak [18] uses a dialogue model to convert natural language commands to keyboard shortcuts. It was found more e cient compared to a conventional screen reader. JustSpeak [200] and VoiceNavigator [58] enable universal access to Android devices through voice commands. Both were found superior in performance and preferable to conventional screen readers.

2.6.4 Issues with Design of Voice Assistants

Commercial voice assistants are modeled after a human-to-human conversation, striving to provide a \frictionless" user experience [38]. However, in the process of mimicking a human conversation, the needs of people with disabilities are not considered. The analyses of Mukkath et al. [129] and Branham et al. [38] discuss how striving for a natural conversational experience may harm the user experience of people with disabilities and in particular, people with visual impairments. For example, the commercial guidelines for voice assistant designers encourage short conversational turns in order to keep the cognitive complexity low. This con icts with the way people with visual impairments prefer to give commands { by providing complex and detailed commands [5]. Abdolrahmani et al. [5] additionally report that verbose feedback produced by voice assistants can be redundant and time-consuming. Cowan et al. [59] also discuss the push-back that may occur around the conversational nature of voice assistants, noting that while some people would prefer to personify Siri and chat with it, others would prefer to give it commands. The \natural" speed of a conversation is often considered frustratingly slow by people who are blind [5, 142] due to the fact that people who are blind can often comprehend speech at much higher rates compared to sighted people [37] as well as process multiple audio streams simultaneously [201].

Another concern is speech recognition errors, which is especially acute when dealing with children's speech, deaf or hard of hearing, and people with disabilities because of

variability in their speech patterns. Correcting speech recognition errors requires tedious work [142] and is often done using a visual interface. Such modality switch between audio interaction and touch-based interaction represents an signi cant interruption for voice assistant users for whom hands-free interaction can be essential [59, 116].

The issue of data privacy in interactions with a voice assistant can manifest itself in several ways. First, users might be concerned about the handling and misuse of their personal information by the company-developer of a voice assistant at hand [59]. Additionally, the design of voice assistants may make it di cult to use voice assistants in public without disclosing personal information or passers by overhearing the details the user's personal a airs [5].

2.7 Visual Interfaces for Web Search

As the number of indexed documents on the web is estimated to be in billions [181], web search becomes a ubiquitous and essential tool for navigating the web [157, 89, 189]. Over the years, researchers proposed several frameworks of information seeking process. The model proposed by Marchionini and White [119] consisted of the following stages:

- recognise, accept, and formulate the problem,
- express the problem to a search system in a form of a query,
- examine results,
- reformulate the problem,
- use results.

Current search system strive to support their users at each step of search process. In this section, I will focus on the process examining search results and, in particular, di erent solutions that have been proposed to aid this process.

A signi cant body of research was produced in an e ort to optimise web search. In their books, Marti Hearst [89] and Max Wilson [189] provide an detailed overview of the past e orts. Over the years of research, the design and the information displayed on the search engine results page, or SERP for short, have undergone a number of changes. However, its main component is a ranked list of results [89, 36]. Traditionally, a SERP consists of a list of documents, or hits. Each hit, is represented by a document surrogate { a combination

of the document's metadata, such as page title, url, and a snippet { a brief extract of the relevant part of the document.

Several researchers examined how the design of SERP and its parts can impact searcher preference. For example, Clarke et al. [55] used clickthrough inversions to identify caption features that make search engine result pages more attractive to searchers. They analysed pairs of adjacent search results, where the lower-ranked result received more clicks, to produce a set of guidelines for displaying search results. According to these guidelines, all query terms should appear in the title when possible, but if they appear in the title, they need not be present in the snippet. Additionally, URLs should be displayed in a manner that emphasizes their relevance to the query. Rose et al. [152] pursued a similar goal of identifying positive features of search results. They conducted an online survey, in which participants were asked preference questions about a set of displayed editor-generated captions. Their ndings include that users preferred full sentences in the snippets, rather than incomplete sentences, and that user trust was increased by the presence of genre cues (e.g., \o cial site").

Other researchers focused on manipulating parts of document surrogates. For example, Aula [19] found that presenting a document summary in a form of a bulletted list increased user performance, and boldedness decreased it. Special attention has been given to studying the desired snippet length for search results. Cutrell and Guan [62] varied snippet length for navigational and informational search tasks, nding that longer snippets are detrimental to the former and bene cial to the latter. They used an eye-tracking methodology to determine that the longer snippet tends to draw user attention to itself, whereas the URL, which plays an in uential role in navigational tasks, does not receive the same attention for informational tasks.

Following this work, Kaisser et al. [97] found that di erent answer lengths are preferred depending on the query type. They also found that crowdworkers could successfully predict the desired answer length given a query. Maxwell et al. [122] later investigated how the varying snippet length impacted user experience. They suggested that longer snippets were considered more informative and clear, as well as led participants to engage with the results more, though there was little change in objective accuracy measures. Paek et al. [134] experimented with interfaces presenting varying amount of information.

A number of researchers also explored alternative organisation of SERP. Dumais et al. [66] tried displaying the results grouped by theme instead of a list. White et al. [184] found that displaying a list of highly ranked document sentences rather than summaries can be bene cial. Among other techniques for information organization is Sarrafzadeh et al.'s work [158] that compared visualization of data through networks and hierarchically

organized data, concluding that networks led to user reading the underlying document signi cantly less. Many decisions in these areas are made based on the spatial locations on the page, and how the user's gaze is distributed on the page [44].

As the variety of types of information available on the internet grew to include web pages, maps, images, etc., so did the need to intelligently organise them. In his work, Horatiu Bota [36] investigated the aggregation of di erent data types and presenting them on a single page.

2.8 Search Task Complexity

Search tasks are a key component in the research and development of information retrieval systems. These tasks provide the goal that users need to achieve with their search and are often used by the researchers to investigate di erent interaction behaviours depending on the di culty of the search task [192, 98, 45].

Many researchers in IR have constructed tasks based on the Taxonomy of Learning [16], which allows for investigating tasks from the perspective of cognitive complexity. This taxonomy is traditionally used in educational settings but has more recently been adopted by information retrieval researchers [93]. The Taxonomy of Learning speci es six levels of cognitive complexity as: remember, understand, apply, analyse, evaluate, and create as seen in Table 2.1.

Prior research in visual text search has shown that more complex tasks lead to greater levels of search interactivity, for example through increased clicks, queries, and time on task [98, 192, 17]. Furthermore, research by Trippas et al. [177] showed that searchers may engage more in di erent parts of the search process depending on the task complexity.

Alternative taxonomies split search tasks into di erent categories. Broder [41] slice the types of search task along di erent axis according to their intent, presenting three classes: navigational (immediate intent to reach a particular site), informational (intent to acquire some information assumed to be present on one of more web pages), and transactional (the intent is to perform some web mediated activity). According to this taxonomy too user preferences may di er. To this end, Cutrell [62] showed that preferences for design of displayed search results may di er depending on the task at hand. They showed that longer snippets increased searcher e ciency for informational tasks, but had an opposite e ect on navigational tasks.

Table 2.1: Task complexity from the Taxonomy of Learning Objectives [16]. In the experiments described in this thesis, three levels were used: Remember, Understand, Analyse (highlighted in bold).

Dimension	De nition
Remember	Retrieving, recognising, and recalling relevant knowledge from long-term memory.
Understand	Constructing meaning from oral, written, and graphic messages through interpreting, exemplifying, classifying, summarising, inferring, comparing, and explaining.
Apply	Carrying out or using a procedure through executing or implementing.
Analyse	Breaking material into constituent parts, determining how the parts relate to one another and to an overall structure or purpose through di erentiating, organising, and attributing.
Evaluate	Making judgments based on criteria and standards through checking and critiquing.
Create	Putting elements together to form a coherent or functional whole; reorganising elements into a new pattern or structure through generating, planning, or producing.

2.9 Chapter Summary

In this chapter, I reviewed the research that impacted various aspects of this dissertation. The chapter began with an outline of the history of voice-activated digital assistants, their acceptance and usage. It continued with the examining speci cs of auditory comprehension and its di erences from visual comprehension. It then explored how voice assistants, and more generally, voice interfaces bene t people with visual impairments. Finally, the chapter provided a brief review of search task complexity taxonomies and visual search engine interfaces.

Chapter 3

Methodology

This thesis uses a variety of research methods to answer its research questions. This chapter explains and re ects upon the methods used in the experiments presented in the remaining chapters. I discuss reasons for choosing these speci c methods as well as merits and shortcomings of each.

3.1 Wizard-of-Oz

As with much of human behaviour, we tend to purposefully or unintentionally adapt the language we use to match that of our dialogue partner { we speak di erently with a toddler, a peer, and an elderly person [76, 51]. This phenomenon also occurs during human-computer dialogues. For example, Eva Luger and colleagues [116] discovered that users of Siri \learn" to phrase their commands in a way that is understood by the assistant and use this language repeatedly afterwards to arrive at the satisfactory results. Because of the limited functionality current digital agents o er, such dialogues are noticeably simpler than human-human ones. While the latter employ complex turn-taking techniques and use secondary communication channel to demonstrate engagement, the former are mostly limited to exchanging explicit commands and requested information. This renders extensive human-human dialogue datasets collected throughout decades of linguistics research unsuitable for the task of analysing human-computer dialogues.

As the research community pushes forward the frontiers of digital assistant research, a multitude of questions arises { would people be polite to computers when speaking to them? Would people use complex language if a dialogue system could correctly react to

it? How would children interact with a embodied robot? Many of these questions can be generalised into \How would people behave if a computer could do X". One way to study a question like this is by asking experiment participant to imagine the hypothetical situation and provide their feedback. However, what people say and what people do famously di er from each other [143]. Another way to approach a question like this is by implementing the system with capabilities in question and observing participants' interaction with it. This choice can prove to be time- and resource-consuming and in certain cases the hypothetical system can be beyond the technological state-of-the-art and simply cannot be implemented at the time.

To remedy situations described above and enable the research of hypothetical or implementation costly systems, a technique known as \Wizard-of-Oz", or WoZ for short, is used. During WoZ studies, subjects are told that they are interacting with an automated system, whereas in fact the system is partially or fully powered by a human operator, or a wizard. Such setup provides researchers with a full control over the information is delivered to the study participants as well the manner in which it is delivered. In a situation like this, participants are likely to exhibit a similar behaviour, for example, use a similar language, that they would when interacting with a fully automated system. While WoZ is popular in studies on natural language systems, for example digital assistants, it can be applied to conduct research of any intelligent interface where the \intelligence" is mimicked by the wizard.

Though the Wizard-of-Oz approach is a popular technique for studying future intelligent systems, a number of concerns should be addressed before choosing to use this methodology. One of the main limitations of WoZ, pointed out by Fraser and Gilbert [72], is participant deception that is frequently involved in the studies. In their WoZ implementation guidelines, Fraser and Gilbert bring up the issue of a potential embarrassment of the participants when the deception is uncovered. From an engineering perspective, Fraser and Gilbert consider an example of a voice dialogue system and discuss the possibility of controlling multiple aspects of the wizard behaviour. Such aspects include restricting the freedom of wizard's output to a set of prede ned choices, distorting the wizard's voice during communication with the user. In addition, the design of a WoZ study can be manipulated so that the wizard is in the participant's sight. An example of this setup can be found in the study by Yarosh et al. [196] where children participants interacted with a voice assistant that was manipulated by a person sitting across the table from them behind an opaque separator such that only their upper shoulders and face remained visible. In their 2012 review study, Riek found that WoZ studies of social robots were often used to emulate such aspects as verbal and non-verbal behaviour as well as navigation and mobility skills [151]. The same method is sometimes used in rapid prototyping to simulate parts of

a system that have not yet been implemented.

In this thesis, the WoZ approach is used in studies described in chapters 4 and 5 both of which study functionality that was not a part of digital assistants at the time of the studies. In chapter 4, an advanced search system is mimicked by the means of the wizard answering the study participants' requests and questions with the same accuracy that another human would. This technique allowed to isolate and study a single variable { how does the perception of a user change when a person is communicating with a computer compared to another person. In this case, the limitations included a learning e ect of the wizard, i.e. the wizard's knowledge of the topics may improved over the course of the experiment resulting in later subjects receiving higher quality of information or receiving it faster. In chapter 5, the WoZ approach was used to simulate a high accuracy digital cooking assistant and study the language study participants used when interacting with it. Whereas a human-level accuracy could not be guaranteed when employing a truly autonomous system. One of the limitations in this experiment was also a potential variability of the system's responses, however the preset answer options were used to reduce the potential variability and alleviate this limitation.

3.2 Workload Assessment

To quantify the amount of e ort required to complete a task using an interface, one may use the notion of a workload. Quantifying workload helps estimate how taxing a certain task or interacting with a system is for the user. One of the most popular ways to assess workload is by administering a questionnaire called NASA Task Load Index, or NASA-TLX for short [88]. It was originally developed in 1988 by Sandra G.Hart and colleagues and aimed at assessing the workload imposed by a variety of tasks in aircraft industry. Over years NASA-TLX became widely popular as a method of workload assessment during interaction with a variety of interfaces [87].

The goal of NASA-TLX is not only to provide an assessment of a perceived workload but also identify the factors that most contribute to it. To this end, NASA-TLX uses the scores of the following six scales to evaluate the overall workload:

- mental demand (MD),
- physical demand (PD),
- temporal demand (TD),

- performance (OP),
- e ort (EF),
- frustration (FR).

Because each of the six factors above can contribute di erently to the nal workload score, in her research, Hart proposes to weigh the scales in order of their perceived importance by the study subjects. As such, the NSA-TLX questionnaire is administered in two steps: (1) a subject chooses a score from 0 to 20 for each of the scales above, and (2) the subject conducts fteen pairwise comparisons of the scales above rank the scales in order of their importance. User scores are converted into a 0 - 100 point scale and weighted accordingly, yielding the nal workload score ranging between 0 and 100 points.

In the decades since its development, NASA-TLX has been widely used in a variety of areas and has sustained a number of modi cations. The most popular modi cation is Raw NASA-TLX, or RTLX, where the second step of pairwise scale comparison is omitted and scales are considered to be contributing to the workload equally. In the meta analysis conducted by Hart in 2006, it is noted that di erent studies have shown that RTLX is more, less, and equally as sensitive as the original NASA-TLX [87]. Other, less frequent modi cations, include reformulating the description of the scales according to the context of the task at hand. In her meta-analysis, Hart points out that such modi cations are undesirable without additional validation of the scales used.

One of the potential drawbacks of NASA-TLX is the ambiguity in interpretation of the workload scores. In other words, there are no anchor scores pointing to \acceptable" workload, or a so-called \red line" above which the workload is considered too high. To alleviate this issue, a Grier and colleagues [81] conducted a meta-analysis of research studies using any variant of NASA-TLX. The meta-analysis showed that the majority of reported workload scores ranged between 26.08 and 68 points. However, Grier [81] notes that subjects' boredom and frustration also contribute to the workload score. For example, in two studies where subjects had to do nothing but wait, the reported workload scores were 12.0 and 14.8. Whereas the lowest score in the analysis was 6.21 for an air-tra c control task.

In this thesis, RTLX is employed in chapters 6 and 5. The studies described in these chapter do not make an attempt to describe the workload in absolute terms but rather to compare the workload changes between di erent experimental conditions.

3.3 Crowdsourcing

Crowdsourcing is a practice where freelance workers, called crowdworkers, complete tasks posted by requesters on the crowdsourcing platform. It became especially popular when the rise of machine learning algorithms necessitated the need in large amounts of human-labelled data. Crowdsourcing is also a popular research tools for studies and experiments that require participation from a large number of people. For example, d'Eon in 2019 used crowdsourcing as a tool for identifying people's perception of fair pay in a collaboration task [64].

Crowdsourcing tasks can be voluntary or paid. Voluntary crowdsoucring tasks, often called citizen science, contribute to a bigger project and o er an opportunity to the crowd-workers to contribute their skill and knowledge in order to progress on the project. In this case, people are incentivesed by the project's success, or engaged by its gami ed nature. For example, Foldit project [], founded by David Baker is puzzle game aimed at analysing the possible protein folding algorithms submitted by the players. Zooniverse [169] is another example { a collection of citizen science projects covering areas including space, biology, medicine, and humanities. Both Foldit and Zooniverse have proven to be immensely successful, yielding results that would other require decades of a liated researchers' work to achieve.

On the other hand, paid crowdsourcing tasks di er in o ering monetary incentive to crowdwourkers. Platforms like Amazon Mechanical Turk (AMT) and Appen (former Figure Eight and Crowd ower) are the more popular ones. Such platform o er an easy way for requesters to connect with crowdoworkers. Requesters post tasks, called Human Intelligence Tasks, or HITs for short. Each HIT has a set of associated requirements for crowdworker eligibility that could include geographic region, native language, performance rating, and others. Once a crowdworker completes a HIT, the requester has an option to accept the work and pay the fee, or reject the work because of a low submission quality. A crowdworker's performance rating then is based on the percentage of accepted HITs they have completed. On the other hand, crowdworkers can choose HITs they would like to complete, for example based on the minimum payment [188].

Because of its low entrance threshold, crowdsourcing has become a full-time work for many. Social communities and forums contain a multitude of tips and automation scripts to streamline the work, list untrustworthy requesters, and share overall experiences. However, the payment workers receive per hour is often below minimum wage due to factors like requesters under-estimating their HIT's duration, time spent searching for HITs, working on rejected HITs [187, 86]. In this work, my colleagues and I aimed to provide fair payment based on a $15 per hour wage.

Due to the nature of crowdsoucring, there may be a signi cant chance of low quality submissions. To avoid this, requesters may set minimum conditions for people who are able to accept and complete their tasks, such as performance rating. In addition to the minimum requirements, a popular practice is to include a \golden task", or an attention check task, where there exists a single unambiguously correct answer. An example of such task can be asking to enter a speci c word in a text box. Submissions that fail attention check tasks help requesters leverage the nal quality of the collected data.

3.4 Other Methods

In addition to the methods described above, other methods were used to facilitate experiments which contributed to this thesis: online survey, usability study, controlled experiments, System Usability Scale (SUS).

3.4.1 Controlled Experiment

Often research questions are based around a hypothesis { a statement that proved or disproved by running an experiment. A hypothesis is usually centered around a speci c aspect, or an independent variable. In order to test the hypothesis, an experimenter should change that variable while keeping all other conditions xed. For example, in chapter 4 one of the hypotheses is as follows: people will rate the same system di erently depending on whether they think it is automatic or human-powered. In order to test whether this hypothesis is supported, only one variable should change { whether the subjects think they are interacting with an automatic or a human-powered system while keeping all other factors constant. By changing the value of the independent variable, experimental conditions are created. In order to avoid what is called a \carryover e ect" { a situation where the order of conditions in uences the outcome of the experiment, Latin square design is often used to counterbalance the order of experimental conditions. Latin square design provides a way to rotate experimental conditions to avoid order e ects, yet does not require to test all possible combinations of independent variables. the controlled experiments utilising Latin square design are used in chapters 4 and 6.

3.4.2 Usability Study

At times, when there is not a speci c hypothesis to be tested, a controlled experiment approach is not applicable. This is the case in chapters 5 and 7, where the goal of the

Table 3.1: Diagram of methods used in this thesis.

	Chapter 4	Chapter 5	Chapter 6	Chapter 7
Wizard-of-Oz	x	x		x
NASA-TLX		x	x	
Crowdsourcing			x	
System Usability Scale		x		x
Controlled experiment	x		x	
Usability study	x			x
Online survey				x

experiment is to evaluate the system, uncover its design aws, and study the behaviour of its users. In the case of a usability study, a study subject is asked to complete a proposed task using a system at hand, while the experimenter provides the instructions and observes the behaviour of the subject. During a usability study, the experimenter may collect quantitative as well as qualitative information about the interaction. While qualitative information may include insights for systems improvements, quantitative information may focus on measurable metrics, such as completion time, number of interactions, etc.

3.4.3 Online Survey

Controlled experiments and usability studies are often conducted in-person and require resources such as the experimenter's time, allocated space and tools. While on one hand, these factors contribute to collecting rich and in-depth information, they also limit the number of potential participants. A survey administered online provides a solution in cases when the sample of participants needs to be relatively large. Even though a survey does not allow for a deep dive into the answers provided by the respondents, a large number of respondents leads to capturing a wider spectrum of experiences and opinions. An online survey is used in chapter 7 to collect information about pros and cons of screen readers to inform the design of Verse.

3.4.4 System Usability Scale (SUS)

A System Usability Scale, or SUS for short, is a questionnaire combining ten statements. A participant is instructed to choose a response on a ve point scale ranging from 0 to 4,

or from \strongly agree" to \strongly disagree". After all responses have been selected, the scores are added and multiplied by 2.5 to produce the nal SUS score ranging from 0 to 100. In contrast with NASA-TLX scores, the absolute values of SUS have been interpreted to denote the quality of the system at hand. As such, any system scored 68 or higher is considered to have above average usability, conversely a system scoring below 68 is considered to be below average. In this thesis, SUS is used in chapters 5 and 7 to ensure that the design of the proposed systems was positively perceived by the study subjects.

Chapter 4

Exploring Conversational Search With Humans, Assistants, and Wizards

Digital assistants are often used for everyday tasks such as smart home controls and music playback. Finding information online and looking up answers to questions is another prominent use case [15, 26]. However, commercially available systems have limited capabilities and often respond with the highest ranking answer provided by the search engine. In cases when an answer is not available, the search results are shown on the screen of a paired device as shown in gure 2.4 and discussed in chapter 2. Researchers working in the area of conversational search are putting e orts into mitigating scenarios like this by designing systems that would be able to narrow down the area of user's interest by engaging them in a dialogue, much like a librarian would with a person with a broad interest in mind. Radlinsky and Craswell in [147] proposed a theoretical model of a search system, describing all scenarios and functions it should be able to process in order for users to achieve their goals. However, no such system has been developed yet, therefore it is users will see dialogue as a bene cial way to nding information online. This chapter describes an experiment with a simulated search system, providing information to the user via a conversation-like exchange.

4.1 Motivation

As discussed in chapters 1 and 2, voice-controlled and well as text-based assistants recently became a solid part of the market. Both technologies are becoming increasingly integrated into people's everyday lives. Voice-activated digital assistants are predominantly used for simple tasks such as controlling music, manipulating smart home appliances, and setting up a timer. On the other hand, text-based chat-bots are often designed to perform a single type of task [56, 68]. Nevertheless, both technologies are also being used for information seeking tasks. Web search is consistently found in the top few categories of use cases for voice assistants, while text-based chats are integrated into commercial websites to provide immediate support to the users [15, 26, 179].

Voice assistants are designed to mimic super cial aspects of human-to-human conversation, leading some people to perceive them as digital friends [145]. On the other hand, the process of designing a dialogue system is usually quite scripted and is based on intent recognition and slot- lling. Though the voice assistants' rules-based design does not impact their ability to successfully answer most factoid questions, the challenges begin when a users' intent becomes more complex and requires deeper engagement. The need to expand the abilities of a conversational system is re ected in the work of Braslavsky et al. [39] who underlined the need for asking clari cation questions in order to make an information-seeking conversation more e cient. The work by Radlinsky and Craswell [147] describing a theoretical framework for a dialogue-based search system makes a step towards designing dialogue systems capable of maintaining a conversation-like exchange.

While the development of text- and voice-based dialogue systems is underway, the users are making sure to motivate the progress. A growing proportion of web search queries are formulated as natural language questions [135, 112, 20] with an average length of a search query growing from 2.35 terms in 1998 [94] to 3.2 terms in 2016 [84]. This phenomenon can be partially explained by the increased usage of voice interfaces [185] and better question-answering technology embedded in search engines. As such, the demand for natural language interfaces for search is emerging. However, before jumping into implementing additional features for conversational-like search systems, it is important to gain a better understanding what the users' expectations are when interacting with a truly intelligent conversational search agent. It is equally important to anticipate how users might behave when faced with a conversational search system since user feedback is critical for system evaluation and improvements. In this chapter, I focus on text-based interaction and begin to explore the rst question of this thesis:RQ1: \How would users perceive digital agents that could understand them as well as their fellow humans?" In particular, I address the following more speci c questions:

- RQ1-a : Given equal performance, would people choose to communicate with another person or a digital assistant and why?
- RQ1-c : What aspects of digital assistant design are important to consider?

As a system capable of supporting dialogue-based interactions for search does not exist yet, user preferences and behaviours when interacting with such system can be studies by the means of a Wizard-of-Oz protocol. In this chapter, I describe a study comparing three conversational search systems: an existing commercial intelligent assistant, a human expert and a human disguised as an automatic system. A total of 21 participants were recruited for the study each of whom were faced with 3 complex information search tasks. Participants interacted with each system using a text-based messaging application. The results of the study suggest: (1) people do not have biases against automatic conversational systems, as long as their performance is acceptable; (2) existing digital assistants cannot be e ectively used for complex information search tasks; (3) by addressing requests from users, even current search systems might be able to improve their e ectiveness and usability, with feasible modi cations.

In the remainder of this chapter, I discuss the rationale behind choosing to use a Wizard-of-Oz protocol, as well as the search task selections for the study. I go on to describe each of the three conversational search systems and the evaluation metrics used in the study. Finally, I conclude with a discussion of the study's ndings.

4.2 Study Design

Three conversational agents were devised to address the research questions above: Wizard, Human, and Automatic. Each agent was assigned a photo to re ect their nature, as seen in Figure 4.1. A total of 21 participants were recruited { all graduate and undergraduate students at Emory University in Atlanta, GA (8 female, 13 male) { to complete three di erent search tasks selected from the TREC Session track 2014 [46] as seen in Table 4.1. Each participant completed three search tasks and interacted with all three agents, completing one task with one agent. The order of tasks as well as agents was rotated according to a Latin square design resulting in 9 groups (3 agents x 3 tasks). The participants were given a brief description of each agent (discussed below) but were not given any speci c instructions on how to communicate with any agent and therefore were free to interact with them in any way they chose. By omitting the instructions, the author and colleagues hoped to observe the way users would choose to interact with an \ideal" agent. Additionally,

the absence of speci c instructions served to elicit interaction patterns in agreement with users' expectations for each agent type. The participants were allowed to spend up to 10 minutes working on each task, after which they were asked to move on a topical quiz. Each quiz included three questions designed for a speci c topic. After seeing the topical quiz questions, the participants were not allowed to talk to the agent again. Doing so ensured that the participants did not have a set of prede ned points to cover and questions to ask the agent. After the study was completed, the answers to the topical quizzes submitted by the participants were analysed and evaluated on a scale from 0 to 2.

After completing the topical quiz, the participants lled out a questionnaire, where they were asked to rate their overall experience with the agent (on a scale 1 to 5). Participants were also asked what features in particular they liked and disliked about the agent and whether they were able to nd all the information they were looking for. Upon completing all the tasks, the participants were explained that one of the agents was powered by a human through a Wizard-of-Oz setup. After being debriefed, participants were asked to choose which system they liked best and why.

Although throughout much of this thesis, I focus on voice-based interaction with dialogue systems, these interactions can introduce unwanted error into an experiment. For example, automatic speech recognition (ASR) errors may contribute to a participant giving the system a lower score. The voice itself can be seen as a confounding variable, since prior research showed that preference for voices varies depending on the individual [149, 102]. To avoid introducing potential mistakes, the interaction in this study was implemented through a text-based exchanges on Facebook Messenger[1]. Participants used a Facebook account created speci cally for the purpose of the study. Message history was cleared prior to every experiment. Considering that the purpose of this experiment was to evaluate users' perception of the agents, the study's ndings are not speci c to the text-based interaction. However, care should be taken in expanding these ndings to voice-based dialogue systems, as factors such as response delay and automatic voice characteristics may have an e ect on the perception of an agent by the user.

4.2.1 Wizard Agent

This thesis explores human behaviour in human-computer conversational communication. However, there are currently no general purpose intelligent conversational search systems, that could be used for the purposes of this experiment. Therefore one was \faked" by substituting the backend with a person (the author of this thesis and a colleague played

[1] www.messenger.com

(a) Wizard agent (b) Human agent (c) Automatic agent

Figure 4.1: Three agents for conversational search: (4.1a) Wizard, (4.1b) Human, and (4.1c) Automatic.

the role of the wizard behind the curtain). However, the participants were told that it was an experimentalautomatic system, thus following a general Wizard-of-Oz setup. I will be further referring to this system as the Wizard agent, and the person in the backend as the Wizard. The Wizard had previously done the research about the topics of the three search tasks prior to the experiment and compiled a broad set of passages covering most of the aspects of each topic. At the time of the experiment, the Wizard tried to find the best passage to reply to the participant's question or comment. However, in cases where such passage could not be found, the Wizard would reply with a passage retrieved from web search, or write a new passage. In case the participant's question or comment was ambiguous, the Wizard was allowed to ask a clarification question to better identify the information need of the participant.

The Wizard agent was allowed to maintain the context of the conversation, respond to vague questions, understand implied concepts, and provide active feedback in form of clarification questions when needed (all of these capabilities do not yet exist in commercial systems). At the same time, by partially restricting the Wizard to a pre-compiled set of passages, it was possible to maintain the consistency of answers between participants, i.e., for the same question any participant would receive a similar answer. By analyzing the ways the participants communicated with the Wizard agent, the insights were gained

about strategies people use in a human-computer dialogue for solving complex tasks and look for design implications for automatic conversational systems thus addressing RQ1-b.

4.2.2 Human Agent

To explore the di erences between human-to-human and human-to-computer communication, a second conversational agent was devised { the Human agent. In this case, the Wizard from the previous setup was still serving as a backend, but the participants were explicitly informed that they were talking to a live person. Another di erence was that the Human agent was not restricted to the pre-compiled set of passages but was free to slightly reformulate or revise the passages to better respond to the question. By including both the Human and Wizard agents in the study, it was possible to maintain a constant level of intelligence for both agents, thus comparing not the accuracy of each agent, but rather the participants' attitude and expectations towards a perceived automatic agent compared to a known human. The results of this comparison served as evidence to address RQ1-a.

4.2.3 Automatic Agent

As a means of comparison to an existing conversational agent, a Google Assistant was used as a backend for the third agent. Every message sent by a participant was forwarded to the Google Assistant app, and the response was forwarded back to the participant. Most of the time, the response consisted of an URL and a text snippet. The participants were told that they were interacting with another experimental conversational search system, but were not given any speci c information about it. Using a system representative of the state-of-the-art technology made it possible to evaluate its drawbacks, and situations where it failed to respond properly. Collecting participants' feedback about each system resulted in a set of design changes that would improve commercial digital assistants for the task of information nding.

4.3 Search Task Selection

The goal in selecting search tasks was to nd ones that were likely to require interaction beyond a single question and answer. For this reason TREC Sessions track was selected. TREC is a yearly conference co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense. It consists of several tracks, each

Table 4.1: Description of the tasks used in the study. All the tasks were obtained from TREC Session track 2014 [46].

Topic ID	Topic description
10	Suppose you are writing an essay about a tax on \junk food". In your essay, you need to argue whether it's a good idea for a government to tax junk food and high-calorie snacks.
20	You have decided that you want to reduce the use of air conditioning in your house. You've thought that if you could protect the roof being overly hot due to sun exposure, you could keep the house temperature low without the excessive use of air conditioning.
21	Hydropower is considered one of the renewable sources of energy that could replace fossil fuels. Find information about the e ciency of hydropower, the technology behind it and any consequences building hydroelectric dams could have on the environment.

targeted to study a di erent application or aspect of information retrieval. Each year TREC organisers supply participants with a test set of documents and tasks. Participants in turn, submit the results of their retrieval systems which are then evaluated by NIST judges.

Sessions track has been a part of TREC since 2010 and is targeted towards evaluating retrieval systems over a search session rather than a single query. More concretely, the tasks included a cross between two facets of search tasks de ned by Li and Belkin [109] \product" and \goal quality". The \product" facet represented the end goal of the search task and could have be either \intellectual" { aimed at producing new ideas based on the information learned, or \factual" { locating existing information items. In turn, the \goal quality" item could be presented as \speci c goal" or \amorphous goal", or as Ingwersen and Järvelin [91] put it { \well de ned" and \ill de ned" information needs. By crossing the two facets, four task categories emerged: factual task with speci c goals { \known item" search, factual task with amorphous goals { \known subject search", intellectual task with speci c goals { \interpretive search", and intellectual task with amorphous goals { \exploratory search". It was not the aim of the study to compare the e ects of di erent search task categories, and therefore all topics from Sessions track were deemed suitable for the purpose. Topics with IDs 10, 20, and 21 were selected as search tasks. Table 4.1 demonstrates a detailed description of each topic, which was provided to the participants

Agent	Human	Wizard	Automatic
Overall satisfaction (max 5)	4.1	3.8	2.9
Able to find information (max 2)	1.5	1.3	1.0
Topical quiz success (max 2)	1.6	1.6	1.3

Table 4.2: Row 1: average satisfaction for each agent; row 2: average rate of success for finding desired information; row 3: average rate of success for answering topical quiz questions.

as a task prompt.

4.4 Results

After running the study, participants' preference and quiz scores were analysed. All of participants' verbal comments were qualitatively analysed to extract commonly occurring sentiment. This section describes the findings.

4.4.1 Overall Satisfaction

After completing each task, participants rated their overall experience of working with each agent on a 1 to 5 Likert scale. Average ratings for each agent are shown in the first row of Table 4.2. The scores were normally distributed for all three systems, making it possible to conduct a paired t-test. The differences in ratings between Human ($M = 4.1$, $SD = 0.8$) and Automatic ($M = 2.9$, $SD = 0.8$) systems as well as Wizard ($M = 3.8$, $SD = 0.7$) vs. Automatic systems were statistically significant with respect to t-statistic and p-values being $t(20) = 5.06$, $p < 0.01$ and $t(20) = 3.5$, $p < 0.001$. While the difference between the Human vs. Wizard systems was not significant with $t(20) = 1.6$; $p = 0.1$. The lack of significant difference in the scores between Wizard and Human systems does not imply that the systems are not different, rather it signifies that for the given context and population, the systems performed similarly. A number of variables could influence these results including the information need at hand, the relatively young age of participants (younger people may have different preferences than older generations), and the environment (laboratory experiments are rarely able to reconstruct real life scenarios).

In the final questionnaire, after completing all the tasks, participants were asked which system they liked the most. Out of 21 people, 8 people preferred the Human agent, 6 {

the Wizard agent, 4 { the Automatic agent, 2 people said they would use the Wizard or the Human depending on their goals, and 1 person said they would choose between the Human and the Automatic agent depending on the goal.

After completing each task participants were asked whether they were able to nd all the information they were looking for. Each answer was coded on a 0-2 scale (0 - no, I couldn't; 1 - partially; 2 - yes, I found everything I needed). Average results for each agent are shown in the middle row of Table 4.2.

4.4.2 Topical Quiz Success

After completing each task participants were asked three questions about the topic. Each of the answers was evaluated on a scale 0-2, where 0 meant no answer, 1 - poor answer, 2 - good answer. On average, participants showed a similar level of success with each agent. The average user ratings for each agent are shown in the bottom row of Table 4.2.

4.5 Qualitative ndings

I now turn to qualitative results, reporting the themes uncovered during the analysis of verbal comments participants provided after the study was completed. The participants' comments broke down into three areas: the ability of an agent to maintain the context between conversational turns, trustworthiness of the information provided by an agent, and the social acceptance of using a human agent for search tasks.

4.5.1 Conversational Context

Within a conversation, people expect that the main topic of the discussion is maintained, and they tend to ask short questions, omitting the subject, or referring to the subject using pronouns. Formulating a full question takes e ort and is unnatural. For the Automatic agent, anaphora resolution did not always work, which annoyed the participants (Figure 4.2 provides an example). For example, P19 had this feedback about the Automatic agent: "It didn't use contextual information so there was no way to expand on the previous answer it gave me." Conversely, when interacting with Human and Wizard agents, participants pointed out the ease of use, because their partially stated questions were understood and relevant answers were returned.

4.5.2 Trustworthiness of Information

Even though the Automatic agent did not always return a relevant result, it received approval from our participants for providing sources of its answers. P7 said "I [...] like to be able to verify the credibility of the sources used." Out of 21 participants, 13 people said that being able to access the URL allowed them to assess the trustworthiness of the source and therefore to accept or reject the answer. On the other hand, in spite the Human and Wizard agents returning more relevant results, they were both criticized for not providing the sources of their answers.

4.5.3 Social Acceptance

When dealing with the Human agent, four out of 21 participants reported feeling uncomfortable talking to a person, thought more about the social norms, were afraid to ask too many questions, were not sure how to start and end a conversation. P15 re ected on this aspect saying: \you have to think about social norms, asking too much, being too stupid, not giving them enough time to respond, troubling them."This additional burden of interacting with another person further motivates research in the area of automated digital agents.

4.6 Discussion and Design Implications

Figure 4.2: Automatic system (gray background) fails to maintain context, which causes the participant 15 (blue background) to reformulate his question twice.

Below I describe the list of design recommendations for a conversational search agent, that

(a) Explicit user feedback could be used to recover from failure. Part of a conversation between participant 12 (blue background) and Automatic system (gray background).

(b) A participant prefers web search to talking to a person. Part of a conversation between participant 7 (blue background) and Human agent (gray background).

Figure 4.3: Examples of user behaviour when interacting with Automatic and Human agents.

according to the study results will improve the
user experience. Additionally, I draw a connection between each of the design recommendations and the four interaction aspects constituting the statement of this thesis: how naturally a user can communicate their intent to the system, the understandability of the system's responses, the exibility of the system's parameters, and the diversity of information accessible through the system.

Context. Maintaining a context of the conversation to enable short questions and comments is crucial to user experience since formulating long sentences each time feels unnatural and takes longer.

Answer sources. Finding relevant and precise answers is important. But trustworthy information sources are equally important, and their absence may diminish the credibility of the system. While the Automatic agent supported each answer with an URL, Human and Wizard did not, unless speci cally asked. Answer source were of could be a deal breaker for some participants, as shown in Figure 4.3b.

Feedback. One crucial di erence between conversational search and conventional web search is an opportunity for the user to provide the system with explicit feedback. It is likely to contain essential information that may help the system to get back up from

failure and improve upon the previous result. For example, in Figure 4.3a a user says That's unhelpful. and rephrases her query. Feedback processing may also be of help in case a user decides to switch the focus of the search. It may also produce rich data for user satisfaction evaluation and can make a rival for the implicit relevance feedback used in web search engines.

Opinion aggregation. According to the participants, sometimes what is needed is the experience of other people in similar situations. A good conversational search system should be able to aggregate opinions and present them to the user in a short summary, perhaps explaining each one. P17 said "It would be nice if I could see a summarisation of different opinions that there exist – from different sources."

Direct answers vs. expanded information Regarding this aspect, our participants split into two camps: those who preferred getting direct answers to the question provided, and those who preferred also getting a broader context. Those expecting concise answers, were unhappy that the answers returned by the systems were too long (even for Wizard and Human agents), and preferred to have their questions answered directly with minimum extra information. On the other hand, those who favoured longer answers, said they preferred talking to a person, who would recognize their true information need (beyond the immediate question) and provide the relevant information.

The findings discussed above illustrate examples of the four interaction aspects outlined in the statement of this thesis. The desire to keep context and ability to manipulate the flow of the dialogue by providing feedback to the system follow directly from the fact that a conversation between two people is seen as a collaborative act where each participant contributes towards a mutual goal. The inability of automatic systems to engage in this behaviour causes the lack of naturalness in phrasing one's queries. Further, the desire to know the source of information helps users understand the provenance of the response and allows for an accurate evaluation of received information. This information may further impact whether or not the user finishes their search or continues to explore other sources. Participants' feedback also provided an example of personal preferences in the answer format – long and expanded with related information vs short and to the point. Importantly, these preferences may change depending on the information need or the environment. Finally, the study participants also expressed desire to have access to a variety of sources, perhaps even to a compilation of opinions about the topic at hand.

The above study was conducted using text-based messaging and future research may find it useful to confirm the findings during interactions with voice-based systems. Additionally, the participants' preferences with regards to the context should be explored. Prior research showed that people may not be comfortable with using voice-based digital

assistants in public [125]. It is also possible that voice or text may be preferred for personal search tasks.

4.7 Chapter Summary

This chapter described a user study with three text-based dialogue systems for web search. It compared participant behaviour when talking to a human expert, a commercial automatic system, and a perceived automatic system secretly controlled by a person (implemented through a Wizard-of-Oz protocol). The observations showed that people do not have biases against automatic systems and are glad to use them as long as their expectations about answer accuracy are met. Furthermore, automatic systems may be preferred in certain social contexts.

Chapter 5

Exploring the Role of Conversational Cues in Guided Task Support with Virtual Assistants

The previous chapter illustrated that people do not have prominent biases against interacting with text-based digital agents for the purposes of information nding. This chapter continues to investigate the idea of people's comfort when communicating with a digital agent but in a di erent context: the agent is voice-based instead of text-based, and the participants are asked to complete a culinary recipe instead of a web search task. The study goes beyond analysing the general sentiment about the agent and delivers a ne-grained analysis of the language used by the participants during their interactions with the voice-based digital agent.

5.1 Motivation

As discussed in chapter 2, a conversation between people is rich and full of subtle verbal and non-verbal cues. We use body language and tone to indicate that it's our partner's turn to talk and use facial expressions to show that we are not satis ed with the information we received. When we talk to computers, however, the protocol is much stricter and provides little space for behaviours that are commonplace in human-to-human conversation. In the majority of cases, commercially available agents support interactions that follow the <trigger word, question, answer> protocol. For example, a weather forecast request can go

as follows: \Hey Siri" followed by \What is the weather like today?" with a response: \It's currently cloudy and 4 degrees". Over time people adapt to this way of interaction and phrase their queries and questions in a way that an agent can understand best [116]. Such communication protocol di ers signi cantly from a human-to-human conversation and may cause trigger word fatigue { a phenomenon where the necessity to pronounce a trigger word before each request causes user frustration.

In the previous chapter, I described a study that focused on analysing user attitudes towards a digital agent for the task of information nding { one of the top use cases for commercial voice-based digital assistants. According to recent surveys, kitchens are in the top of the list of smart speaker placement. In their study, Graus et al. showed that users of smart speaker frequently employ them at mealtime to set timers or to manage short processes related to cooking [79]. Commercial voice-based digital agents, such as Google Home and Amazon's Echo, implement the functionality of walking a user through a recipe step by step [63, 171]. In this chapter, I describe a study which investigates how users would interact with a voice-based digital agent able to understand them as well as a human.

To tackle this problem, the author and colleagues ran a high- delity Wizard-of-Oz study in which people were asked to interact with a voice-based digital agent as they prepared a simple culinary recipe. This chapter explores potential interactions that occur in the moments following, or in lieu of, users' explicit \trigger word, question, answer" triples. The goal of the study was to analyse the language that was used by the participants and answer the question: RQ1-b: What opportunities are opened by the language people use?

In order to conduct such an investigation, a unit of analysis needs to be de ned. A challenge with analysing transcripts of verbal exchange between two or more parties is that, unlike in written text where words and sentences follow one another, in situations when two or more parties are involved, participants may interrupt each other and speak at the same time [154]. During the analysis conducted in this study, a unit of analysis was considered to be a continuous uninterrupted speech from one of the parties, called an utterance. Although the discussed study employs a high- delity prototype of a voice-based digital assistant, a dialogue between the users and the system had very well de ned turns and no interruptions occurred at any time.

As participants engaged with the agent, it was observed that their utterances divided into two groups: (1) explicit { requests that are clearly phrased as questions (e.g. \What do I do next?", \What else?"), or as imperative commands (\Read me the next step.", \Next step."), and (2) implicit { utterances that could not be de nitively interpreted in isolation, i.e. without the information about previous exchanges, intonation, or timing.

1. U: Alright. Uh... What ingredients do I need?
2. A: One-half teaspoon of chili powder.
3. U: Okay.
4. A: One-eighth teaspoon of dried oregano.
5. U: Oregano. Okay.
6. A: One pinch onion powder.
7. U: Okie-doke.
8. A: Cayenne pepper to taste.
9. U: Sounds good.
10. U: Alright, can you repeat that?
11. A: Cayenne pepper to taste.
12. U: Uh, I meant all the ingredients.

Figure 5.1: An example exchange between a user (U) and the agent (A) during the Wizard-of-Oz study. Italicised user utterances are implicit conversational cues { utterances that advance the conversation and move the user closer to their goal, without the user asking an explicit question nor giving an imperative command. An implicit cue cannot be de nitively interpreted without seeing the history of preceding exchanges.

For example \Okay" to signal one's readiness to proceed to the next step, or asking to con rm the correctness of an instruction by repeating it. For the remainder of the chapter, I will refer to these two groups as explicit and implicit cues. For example, in Figure 1, I consider highlighted utterances to be the implicit cues. In other words, the user intents for utterances 3, 5, 7, 9 could not be clearly identi ed if they were stripped of the surrounding context, and are therefore considered to be implicit. On the other hand, utterances like 10 leave no doubt as to what the user intent is, even if not provided with any additional context.

The remainder of the chapter is structured as follows. I describe the Wizard-of-Oz experiment and introduce the taxonomy of intents of verbal conversational cues for a task-oriented dialogue. I continue to describe the di erent purposes of short a rmative utterances (e.g.,\Okay."), as well as conversational cues that repeat the system's previous response. I conclude the chapter by presenting design implications for future voice-based dialogue systems.

5.2 Study design

In this work, a high- delity Wizard-of-Oz simulation was developed. A simulated voice-based agent was used to study the role of conversational cues in guided task scenarios. I describe the protocol and apparatus below.

5.2.1 Apparatus

To study user interactions with a culinary assistant, a simulated assistant was developed using a Wizard-of-Oz protocol. (The success of this experiment depended on the delity of the Wizard-of-Oz simulation.) The main goal of the study was to simulate the limited capabilities of commercial systems in terms of the answers they are able to provide. The purpose of the experiments was not to study a well-rounded conversationalist that could also guide users through cooking a recipe. Conversely, the simulation was designed so that the assistant could support one narrowly de ned task.

One of the main considerations in developing of the assistant was to minimize latency and ensure consistency of responses across participants. To this end, a preset list of computer-synthesized audio responses was developed from which an experimenter { the Wizard { could select the appropriate one. The response list included each of the recipe's ingredients, each sentence from the list of recipe instructions, as distinct candidate answers. Additionally, relevant culinary de nitions were included (e.g., \a pinch", \to taste", etc.), meta-information about the recipe (\What is the cooking time?", \number of servings"), as well as a \no answer" response to handle questions that fell out of scope of the current task. The Wizard was allowed to type a free-form response, in case an unexpected but related question was asked. However, a post-experiment analysis showed that this option was used mainly to produce \yes" and \no" responses. Overall the design of the simulated assistant was successful { no participants reported suspecting that they were interacting with a simulation.

5.2.2 Procedure

A total of 10 participants (6 male, 4 female, average age 30) were invited to engage with a simulated conversational assistant with the goal of preparing a simple culinary recipe. Out of the 10 participants, 2 reported having used an intelligent assistant earlier that day, 5 { earlier that week, and 1 each { earlier that month, more than a month ago and never. 8 people said they usually enjoyed cooking, and 6 said they cooked often.