



Spoken Conversational Search: Audio-only Interactive Information Retrieval

*A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy*

Johanne R. TRIPPAS

Master of Computer Science – RMIT University
Postgraduate Certificate in Gender and Cultural Studies – The University of Melbourne
Bachelor of Social Sciences – Sociale School Heverlee

School of Science
College of Science, Engineering, and Health
RMIT University

May, 2019

Declaration of Authorship

I certify that except where due acknowledgement has been made, the work is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program; any editorial work, paid or unpaid, carried out by a third party is acknowledged; and, ethics procedures and guidelines have been followed. I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Johanne R. TRIPPAS

[School of Science](#)

[RMIT University](#)

August 2019

“Ook de slak bereikt de ark.”

Johanne R. Trippas

Acknowledgements

I would like to thank my supervisors Professor Lawrence Cavedon, Professor Mark Sanderson, and Doctor Damiano Spina who have guided me through the process of conducting and writing research. I am thankful for all their advice and support. Finally, I would like to thank my parents, siblings, family, and friends for their unfailing support throughout my studies.

Publications

Publications Used in this Thesis

In this section we present the research papers that resulted from this Ph.D. study. For each paper, we refer to the corresponding chapter in which the content of the paper is included:

1. J. R. Trippas, D. Spina, M. Sanderson, and L. Cavedon. Towards understanding the impact of length in web search result summaries over a speech-only communication channel. In *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*, pages 991–994, 2015. The content of this paper is included in Chapter 4.
2. J. R. Trippas. Spoken conversational search: Information retrieval over a speech-only communication channel. In *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*, page 1067, 2015. The content of this paper is included in Chapter 2.
3. J. R. Trippas, D. Spina, M. Sanderson, and L. Cavedon. Results presentation methods for a spoken conversational search system. In *CIKM'15 First International Workshop on Novel Web Search Interfaces and Systems (NWSearch'15)*, pages 13–15, 2015. The content of this paper is included in Chapter 2.
4. J. R. Trippas. Spoken conversational search: Speech-only interactive information retrieval. In *Proceedings of Conference on Information Interaction and Retrieval (CHIIR)*, pages 373–375, 2016. The content of this paper is included in Chapter 2.
5. J. R. Trippas, D. Spina, L. Cavedon, and M. Sanderson. How do people interact in conversational speech-only search tasks: A preliminary analysis. In *Proceedings of Conference on Information Interaction and Retrieval (CHIIR)*, pages 325–328, 2017. The content of this paper is included in Chapter 8.
6. J. R. Trippas, D. Spina, L. Cavedon, and M. Sanderson. A conversational search transcription protocol and analysis. In *SIGIR 1st International Workshop on Conversational Approaches to Information Retrieval (CAIR'17)*, 2017. 5 pages. The content of this paper is included in Chapter 5.

7. J. R. Trippas, D. Spina, L. Cavedon, and M. Sanderson. Crowdsourcing user preferences and query judgements for speech-only search. In *SIGIR 1st International Workshop on Conversational Approaches to Information Retrieval (CAIR'17)*, 2017. 3 pages. The content of this paper is included in Chapter 5.
8. J. R. Trippas, D. Spina, L. Cavedon, H. Joho, and M. Sanderson. Informing the design of spoken conversational search. In *Proceedings of Conference on Information Interaction and Retrieval (CHIIR)*, pages 32–41, 2018. The content of this paper is included in Chapters 6 and 9.
9. J. R. Trippas and P. Thomas. Data sets for spoken conversational search. In *Proceedings of the CHIIR 2019 Workshop on Barriers to Interactive IR Resources Re-use (BIIRRR 2019)*. *CEUR-WS*, pages 14–18, 2019. The content of this paper is included in Chapter 7.
10. J. R. Trippas, D. Spina, P. Thomas, H. Joho, M. Sanderson, and L. Cavedon. Towards a model for spoken conversational search. *Information Processing & Management*, 2019. (Submitted). The content of this article is included in Chapters 7 and 9.

Additional Publications

During the course of this Ph.D., the following papers have been published, but are not included in this thesis as they are not directly connected to the research topic.

1. D. Spina, J. R. Trippas, L. Cavedon, and M. Sanderson. SpeakerLDA: Discovering topics in transcribed multi-speaker audio contents. In *Proceedings of the Third Edition Workshop on Speech, Language & Audio in Multimedia (SLAM)*, pages 7–10, 2015.
2. D. Spina, J. R. Trippas, L. Cavedon, and M. Sanderson. Extracting audio summaries to support effective spoken document search. *Journal of the Association for Information Science and Technology (JASIST)*, 68(9):2101–2115, 2017.
3. S. Shiga, H. Joho, R. Blanco, J. R. Trippas, and M. Sanderson. Modelling information needs in collaborative search conversations. In *Proceedings of Conference*

- on *Research and Development in Information Retrieval (SIGIR)*, pages 715–724, 2017.
4. M. Aliannejadi, M. Hasanain, J. Mao, J. Singh, J. R. Trippas, H. Zamani, and L. Dietz. ACM SIGIR student liaison program. *ACM SIGIR Forum*, 51(3):42–45, 2018.
 5. C. Qu, L. Yang, W. B. Croft, J. R. Trippas, Y. Zhang, and M. Qiu. Analyzing and characterizing user intent in information-seeking conversations. In *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*, pages 989–992, 2018.
 6. A. Chuklin, A. Severyn, J. R. Trippas, E. Alfonseca, H. Silen, and D. Spina. Prosody modifications for question-answering in voice-only settings. *arXiv preprint arXiv:1806.03957*, pages 1–5, 2018.
 7. C. Qu, L. Yang, W. B. Croft, Y. Zhang, J. R. Trippas, and M. Qiu. User intent prediction in information-seeking conversations. In *Proceedings of Conference on Information Interaction and Retrieval (CHIIR)*, pages 25–33, 2019.
 8. J. R. Trippas, D. Spina, F. Scholer, A. H. Awadallah, P. Bailey, P. N. Bennett, R. W. White, J. Liono, Y. Ren, F. D. Salim, and M. Sanderson. Learning about work tasks to inform intelligent assistant design. In *Proceedings of Conference on Information Interaction and Retrieval (CHIIR)*, pages 5–14, 2019.
 9. J. Liono, J. R. Trippas, D. Spina, M. S. Rahaman, Y. Ren, F. D. Salim, M. Sanderson, F. Scholer, and R. W. White. Building a benchmark for task progress in digital assistants. In *Proceedings of WSDM’19 Task Intelligence Workshop (TI@WSDM19)*, 2019. 6 pages.
 10. J. Kim, J. R. Trippas, M. Sanderson, Z. Bao, and W. B. Croft. How do computer scientists use Google Scholar?: A survey of user interest in elements on SERPs and author profile pages. In *Proceedings of the 8th Workshop on Bibliometric-enhanced Information Retrieval (BIR 2019)*. *CEUR-WS*, pages 64–75, 2019.

Contents

Declaration of Authorship

Acknowledgements	ii
Publications	iii
Contents	vi
List of Figures	xi
List of Tables	xii
Abbreviations	xiv

Abstract	1
----------	---

I Thesis Overview and Background 3

1 Introduction 5

1.1 Motivation	6
1.2 Challenges for Spoken Conversational Search	6
1.3 Contributions	8
1.4 Thesis Structure	10

2 Background 13

2.1 The Rise of the Spoken Conversational System	14
2.1.1 Spoken Conversational Search	14
2.2 Interactivity in Information Retrieval	16
2.2.1 Task and Task Complexity	17
2.2.1.1 Task Complexity and Search	18
2.2.1.2 Task Complexity and Discourse	18
2.3 Information Seeking Processes and Models	19
2.3.1 Modelling Information Seeking Through Dialogue	21
2.4 Fundamental Search Actions Through Audio	24

2.4.1	Spoken Queries	24
2.4.2	Results Presentation and Answer Organisation Through Audio	24
2.5	Speech User Interfaces	25
2.5.1	Spoken Dialogue Systems	26
2.5.2	Dialogue Analysis	27
2.6	Conclusion	28
2.7	Chapter Summary	28
II	User Preferences in Results Presentation and Access over an Audio-Only Communication Channel	29
3	Accessing Media Via an Audio-only Communication Channel	31
3.1	Introduction	32
3.1.1	RealSAM Application	33
3.1.2	Dataset	35
3.2	RealSAM Log Analysis	35
3.2.1	General and Session Descriptives	35
3.2.2	How People use RealSAM	36
3.2.2.1	One-Interaction Sessions	37
3.2.2.2	Search Sessions	38
3.2.3	Text-to-Speech Output	39
3.3	Discussion	40
3.3.1	Limitations	41
3.4	Conclusion	41
3.5	Chapter Summary	42
4	Results Presentation for Audio-only Communication	43
4.1	Introduction	45
4.1.1	Aims and Purpose	45
4.2	Methodology: Results Presentation	45
4.2.1	Crowdsourcing Method	46
4.2.2	Experimental Design	46
4.2.2.1	Tasks	46
4.2.2.2	Queries	46
4.2.2.3	Search Engine Results Summaries	47
4.2.2.4	Post-task and Exit Questionnaires	48
4.2.2.5	Using Text as Baseline for Audio	49
4.2.3	Participants	49
4.3	Results	51
4.3.1	Query Judgement Distribution	51
4.3.2	Preferred Length of Text Summaries	52
4.3.3	Preferred Length of Audio Summaries	52
4.4	Discussion	53
4.5	Conclusions	54
4.6	Chapter Summary	54

III	Towards a New Model of Spoken Conversational Search	57
	Introduction to Part III	59
	Aims and Purpose	60
	Natural Dialogue Study	60
	Exploratory Observational Analysis	61
	Thematic Analysis	61
	Validating Thematic Analysis	62
	Overall Approach and Setup	62
5	Methods	63
	5.1 Approach	64
	5.2 Definitions	65
	5.3 Study Design	66
	5.4 Recruitment and Sampling	67
	5.5 Data Collection Setup	67
	5.5.1 Task Design	67
	5.5.2 Procedure	68
	5.5.3 Questionnaires	70
	5.5.3.1 Pre-test Questionnaire	70
	5.5.3.2 Pre-task Questionnaire	71
	5.5.3.3 Post-task Questionnaire	72
	5.5.3.4 Exit Questionnaire	72
	5.5.4 Semi-structured Interview	72
	5.5.5 Apparatus	72
	5.6 Participants	73
	5.7 Transcription Methodology	74
	5.7.1 Transcription Principles	75
	5.7.2 Transcription Protocol	76
	5.7.3 Transcription Quality Assurances	77
	5.8 Data Analysis and Annotation Schema Creation	77
	5.8.1 Coding Transcriptions With Thematic Analysis to Develop SCoSAS	77
	5.8.2 Analysis of Coding	79
	5.8.3 Validation of Annotation Schema SCoSAS	79
	5.9 Chapter Summary	80
6	Observing Spoken Conversational Search Interaction Behaviour	81
	6.1 Search Interactions	82
	6.1.1 Query Formulation	82
	6.1.2 Search Results Exploration	84
	6.1.3 Query Reformulation	86
	6.1.4 Search Results Management	88
	6.2 Non-Search Interactions	89
	6.2.1 One Utterance Consists of Multiple Moves	89
	6.2.2 User and System Models and Memory	90
	6.2.3 Decision Offloading and Taking Control	91
	6.2.4 Effective Information Transfer	92

6.2.5	Linking Non-search Related Observations	92
6.3	Chapter Summary	93
7	Identifying, Classifying, and Validating the Interaction Space for Spoken Conversational Search	95
7.1	Aims	96
7.2	Utterance Classification: Themes for SCS	97
7.2.1	Theme 1: The Task Level	97
7.2.2	Theme 2: The Discourse Level	100
7.2.3	Theme 3: Other	103
7.2.4	SCoSAS Subtraction	103
7.3	Inter-rater Reliability and Code Overlap	103
7.4	Validation of SCoSAS	104
7.4.1	MISC Dataset	105
7.4.2	Validation of SCoSAS with MISC	106
7.4.2.1	MISC Data Statistics and Subset	106
7.4.3	Differences Between the SCSdata and MISC Datasets	107
7.4.3.1	Search Tasks	107
7.4.3.2	Setup of SCSdata and MISC	107
7.4.3.3	Transcription Differences	108
7.4.3.4	Utterance Labelling	108
7.4.4	Creating Comparable Datasets	109
7.4.5	Code Overlap and Coverage Between SCSdata and MISC Data	109
7.4.6	Descriptions of Code Set Differences	110
7.4.7	Discussion of SCoSAS Validation	112
7.5	Chapter Summary	113
8	Task Complexity and Interactivity for Spoken Conversational Search	115
8.1	Introduction	116
8.1.1	Research Questions	117
8.2	Methods	117
8.2.1	Interaction Behaviours	117
8.3	Results	118
8.3.1	Overall SCSdata Statistics	118
8.3.1.1	Utterance Length	119
8.3.1.2	One-word Turns	119
8.3.2	Data Analyses	120
8.3.2.1	Task Complexity and Search Interactions	120
8.3.2.2	Task Complexity and Discourse Utterances	121
8.3.2.3	Bigram Interactions	122
8.4	Discussion	123
8.5	Conclusion	124
8.6	Chapter Summary	125

IV Discussion	127
9 Recommendations for the Design of Spoken Conversational Search Systems	129
9.1 SCS Design Recommendations	130
9.1.1 Task Level Design Recommendations	130
9.1.2 Discourse Level Design Recommendations	134
9.2 Towards Models and Detectable Components of SCS	138
9.2.1 Schematic SCS Themes Model	138
9.2.2 Increased Complexity, Interactivity, and Pro-activity	140
9.2.3 Evaluating Existing Search Behaviour Models with SCoSAS	141
9.3 Expanding SCS Requirements	142
9.4 Chapter Summary	143
10 Conclusion and Future Work	145
10.1 Summary of Contributions	146
10.2 Extensions	149
10.3 Informing Future Experiments	149
10.3.1 Informing Wider Research Agendas	152
A Ethics Approvals and Participant Information Statement	153
A.1 Ethics Approval BSEHAPP 10-14	153
A.2 Ethics Approval ASEHAPP 08-16	155
A.3 Participant Information Statement	158
B Questionnaires and Semi-structured Observational Study Interview Questions	162
B.1 Pre-task questionnaire for the Seeker	163
B.2 Post-task questionnaire for the Seeker	164
B.3 Post-task questionnaire for the Intermediary	165
B.4 Exit questionnaire for the Seeker	166
B.5 Exit questionnaire for the Intermediary	167
B.6 Semi-structured Observational Study Interview Questions	168
C SCSdata	170
C.1 Provided Files	170
C.2 Acknowledgments	170
D Spoken Conversational Search Interaction Themes	171
D.1 Theme 1: Task Level	171
D.2 Theme 2: Discourse Level	173
D.3 Theme 4: Other Level	174
Bibliography	175

List of Figures

2.1	COR model by Sitter and Stein [168].	22
3.1	RealSAM device.	33
3.2	Normalised session frequency in 24 hours on weekday and weekend days.	36
3.3	Interaction frequency of super categories.	38
3.4	Speed of the output in the interactions.	39
4.1	Example CrowdFlower task.	50
5.1	Schematic overview of methodology.	65
5.2	Experimental setup.	66
5.3	Visual overview of experiment procedure.	66
5.4	Overview of questionnaires and their measures.	71
5.5	Participants' search engine usage per day ($N = 26$).	73
5.6	Frequency usage of intelligent personal assistants ($N = 26$).	74
5.7	Sample screenshot of ELAN transcription and analysis tool (anonymised). Annotations indicate (a) Seeker, (b) Intermediary, (1) Controlled vocab- ulary Seeker, (2) Transcription, (3) Query.	78
5.8	Example of coding utterances for Seeker and Intermediary.	80
6.1	Example of multiple moves in one utterance.	90
8.1	Number of words per turn for both actors ($N = 1044$).	119
9.1	Schematic model of SCoSAS themes and sub-themes.	139
9.2	Possible schematic inclusion of System Level function.	139
10.1	Future work includes creating models and systems.	150

List of Tables

2.1	Identified SCS requirements.	17
2.2	An overview of the actions and interactions hypothesised in Azzopardi et al. [19].	23
3.1	Top-five frequent input terms.	36
3.2	Most frequently used RealSAM interaction categories.	37
3.3	Most frequent one-interaction session categories.	38
3.4	Query characteristics.	39
3.5	Most frequent query terms.	39
4.1	Examples of queries and query descriptions for single-facet and multi-facet queries.	47
4.2	Examples of full and truncated summaries for occupational therapist query.	48
4.3	Post-task questionnaire questions.	49
4.4	Exit questionnaire questions.	49
4.5	Exit questionnaire results for preferences in the search engine result summaries.	52
5.1	Anderson and Krathwohl's Taxonomy of Learning objectives (cognitive process dimension) [9].	68
5.2	Example search tasks taken from Bailey et al. [20].	69
6.1	Example information request utterances.	83
7.1	Themes and sub-themes used by different actors.	103
7.2	Independent Assessors' code overlap.	104
7.3	Codes used by Assessor 1 and not by Assessor 2.	104
7.4	MISC search tasks.	107
7.5	SCSdata and MISC dataset descriptives.	109
7.6	Set difference between MISC and SCS datasets.	110
8.1	Research questions and hypotheses.	117
8.2	Interaction behaviour measures.	118
8.3	Interaction behaviours per task complexity.	120
8.4	Discourse behaviours per task complexity.	121
8.5	Interaction bigrams for task complexity.	122
9.1	SCS Design Recommendations (DR).	130
9.2	SCS grounding model components (or UMII).	136

9.3	Predefined SCS requirements.	142
9.4	Redefined SCS requirements.	143
B.1	Pre-task questionnaire for the Seeker.	163
B.2	Post-task questionnaire for the Seeker.	164
B.3	Post-task questionnaire for the Intermediary.	165
B.4	Exit questionnaire for the Seeker.	166
B.5	Exit questionnaire for the Intermediary.	167
D.1	Information Request (Seeker).	171
D.2	Results Presentation (Intermediary).	172
D.3	Search Assistance (Seeker and Intermediary).	172
D.4	Search Progression (Seeker).	172
D.5	Discourse Management (Seeker and Intermediary).	173
D.6	Grounding (Seeker).	173
D.7	Navigation (Seeker).	173
D.8	Visibility of System Status (Seeker and Intermediary).	173
D.9	Other Level (Seeker).	174

Abbreviations

ASK	A nomalous S tates of K nowledge
ASR	A utomatic S peech R ecognition
COR	C onversational R oles
DA	D ialogue A cts
ELAN	E UDICO L inguistic A nnotator
IR	I nformation R etrieval
IIR	I nteractive I nformation R etrieval
ISP	I nformation S eeking P rocess
ISU	I nformation S tate U date
MISC	M icrosoft I nformation- S eeking C onversation data
NDS	N atural D ialogue S tudy
QRFA	Q uery R equest F eedback A nswer
SCoSAS	S poken C onversational S earch A nnotation S chema
SCS	S poken C onversational S earch
SDS	S poken D ialogue S ystem
TREC	T ext R etrieval C onference
TTS	T ext- T o- S peech
WOZ	W izard of O z

RMIT UNIVERSITY

Abstract

School of Science
College of Science, Engineering, and Health

Doctor of Philosophy

Spoken Conversational Search: Audio-only Interactive Information Retrieval

by Johanne R. TRIPPAS

Speech-based web search where no keyboard or screens are available to present search engine results is becoming ubiquitous, mainly through the use of mobile devices and intelligent assistants such as Apple’s HomePod, Google Home, or Amazon Alexa. Currently, these intelligent assistants do not maintain a lengthy information exchange. They do not track context or present information suitable for an audio-only channel, and do not interact with the user in a multi-turn conversation. Understanding how users would interact with such an audio-only interaction system in multi-turn information seeking dialogues, and what users expect from these new systems, are unexplored in search settings. In particular, the knowledge on how to present search results over an audio-only channel and which interactions take place in this new search paradigm is crucial to incorporate while producing usable systems. Thus, constructing insight into the conversational structure of information seeking processes provides researchers and developers opportunities to build better systems while creating a research agenda and directions for future advancements in Spoken Conversational Search (SCS). Such insight has been identified as crucial in the growing SCS area.

At the moment, limited understanding has been acquired for SCS, for example how the components interact, how information should be presented, or how task complexity impacts the interactivity or discourse behaviours. We aim to address these knowledge gaps. This thesis outlines the breadth of SCS and forms a manifesto advancing this highly interactive search paradigm with new research directions including prescriptive notions for implementing identified challenges.

We investigate SCS through quantitative and qualitative designs: *(i)* log and crowd-sourcing experiments investigating different interaction and results presentation styles, and *(ii)* the creation and analysis of the first SCS dataset and annotation schema through

designing and conducting an observational study of information seeking dialogues. We propose new research directions and design recommendations based on the triangulation of three different datasets and methods: the log analysis to identify practical challenges and limitations of existing systems while informing our future observational study; the crowdsourcing experiment to validate a new experimental setup for future search engine results presentation investigations; and the observational study to establish the SCS dataset (SCSdata), form the first Spoken Conversational Search Annotation Schema (SCoSAS), and study interaction behaviours for different task complexities.

Our principle contributions are based on our observational study for which we developed a novel methodology utilising a qualitative design. We show that existing information seeking models may be insufficient for the new SCS search paradigm because they inadequately capture meta-discourse functions and the system’s role as an active agent. Thus, the results indicate that SCS systems have to support the user through discourse functions and be actively involved in the users’ search process. This suggests that interactivity between the user and system is necessary to overcome the increased complexity which has been imposed upon the user and system by the constraints of the audio-only communication channel. We then present the first schematic model for SCS which is derived from the SCoSAS through the qualitative analysis of the SCSdata. In addition, we demonstrate the applicability of our dataset by investigating the effect of task complexity on interaction and discourse behaviour. Lastly, we present SCS design recommendations and outline new research directions for SCS.

The implications of our work are practical, conceptual, and methodological. The practical implications include the development of the SCSdata, the SCoSAS, and SCS design recommendations. The conceptual implications include the development of a schematic SCS model which identifies the need for increased interactivity and pro-activity to overcome the audio-imposed complexity in SCS. The methodological implications include the development of the crowdsourcing framework, and techniques for developing and analysing SCS datasets. In summary, we believe that our findings can guide researchers and developers to help improve existing interactive systems which are less constrained, such as mobile search, as well as more constrained systems such as SCS systems.

Part I

Thesis Overview and Background

Chapter 1

Introduction

“The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it.” Weiser [210, p. 94]

Human speech and conversations are the most intuitive form of communication people use and yet interactions with computer systems, for example search, were historically primarily based on visual user-input (e.g., typed queries) and visual system-output (e.g., list of search results). Over the past decade, speech-based search applications have become more prominent and are increasingly accepted among the wider population. For example, Google reported in 2014 that 55% of people aged between 13–18 years old and 41% of adults use voice search more than once a day.¹ Research has been conducted into supporting search by voice user-input, identifying a number of difficulties in the narrow channel of speech [157]. Few studies, however, have focused on voice system-output. In addition, the conceptualisation of possible user–system interactions and the presentation of voice information have not been explored [212]. Given that conversation is the natural mode for information exchange in daily life, a conversational format for search input and output is logical and could overcome the difficulties inherent in the narrow channel of speech.

Searching in a more natural way over voice through conversation is a logical extension of the visual version, with the potential to transform how we interact with search systems while making searching more accessible and intuitive. The first step in achieving this is to narrow down and understand the expected possibilities of conversational moves in this audio-only communication channel. This thesis explores these conversational actions for the task of search.

¹<https://googleblog.blogspot.com/2014/10/omg-mobile-voice-survey-reveals-teens.html>

1.1 Motivation

Speech output is adequately used for single-turn factoid-style queries which only require one interaction (e.g., “Who is the prime minister of Australia?”) by systems such as Apple’s HomePod, Google Home, or Amazon Echo. However, when users seek answers to non-factoid or ambiguous style queries, which require an in-depth search results investigation, the system falls back on displaying the results list on the screen [159]. Nevertheless, there are many scenarios where an audio-only user interface is preferred, such as when operating machinery [67, 68]; when no screen or keyboard is available [56, 224]; when users are on the move [203]; or when using wearable devices [49]. More importantly, some user groups such as users with a visual impairment [156], people with dyslexia, or people with limited literacy skills are disadvantaged in accessing information on screen. Visually impaired users have been using screen reader software for many years, however, this software is still often difficult and frustrating to use because the content is mainly expressed visually [1].

Listening to complex search results over audio is cognitively taxing for users. This is because audio is a temporal medium and does not leave any traces to which the user may later refer, making speech a linear medium [117, 222, 223]. Thus, it is difficult to convey large amounts of information via audio without overloading the user’s short-term memory [117, 159, 203].

Conversational search has been identified as a critical new research area for Information Retrieval (IR) [6, 62]. The aim of this thesis is to explore this new interaction paradigm for effective and efficient Interactive Information Retrieval (IIR) over an audio-only channel: Spoken Conversational Search (SCS) enabling a conversational approach to defining user information needs, presenting results, and facilitating search reformulations.

1.2 Challenges for Spoken Conversational Search

This thesis is concerned with the exploration of two overarching challenges for SCS:

- How should search results be presented over an audio-only communication channel in order to support the user in their search exploration?
- How would people search in an audio-only interaction setting?

With respect to the first challenge, studies have investigated search results presentation with reference to the visual aspect of a summary or snippet [181], the number of search

engine results which should be displayed [107], or the effects of entity cards on search behaviour and perceived workload [33]. Research has also been undertaken to understand the optimal snippet presentation in a browser-based setting indicating the scale of the research problem [102, 103, 129]. Even though a plethora of research has been devoted to search results presentation in a browser-based setting, few studies have investigated effective search results presentation via an audio-only channel [68]. Furthermore, the focus in audio-only search has mainly been on spoken user input and little attention has been devoted to system output [212].

Presenting a search engine results page (SERP) with “ten blue links” over an audio-only communication channel presents a number of challenges; in particular, simply speaking the textual component of a standard browser-based search results list has been shown to be ineffectual [156]. For example, the structure, layout, and style of the webpage which is used to decide whether a document is relevant or not is more challenging to convey in an audio-only setting. The serial nature of the audio-only channel also makes it difficult for users to “skim” back and forth over a list of results (a standard process in browsing a visual list).

The length of a spoken search result summary plays a crucial role in the success or failure of presenting search results over audio. A short summary might not yield enough information to judge whether the retrieved document is relevant or not; in contrast, a more descriptive summary might take too long to be played and thereby diminish user experience. Thus a trade-off is necessary between a short summary and a longer, more descriptive summary. In particular, we seek a better understanding of how to present search results over audio while not overwhelming the users with information [203], nor leaving users uncertain as to whether they have covered the information space [206]. In this thesis, we hypothesise that interactivity through an audio-only channel may increase in order to overcome the complexity this narrow and limited bandwidth channel imposes. Thus, we believe that conveying information through interactions may alleviate some of the complexities which are associated with searching over an audio-only channel. Furthermore, investigating SCS may help us understand more effective ways to present search results than in the traditional search engine results page and thus transform how we fundamentally interact with search systems.

With respect to the second challenge, extracting SCS interactions is related to the complexity of multi-turn open domain information exchanges between two or more actors. Well-established conversational systems such as Spoken Dialogue Systems (SDS) are bound to a domain and are optimised for slot-filling in which possible interactions are pre-defined [132, 133]. However, since we are dealing with the open web, we need to study which actions users take to converse their information need while the system

supports the user in their document inspection, judgement, and query (re)formulation process. We capture the user's behaviour and define these processes. We also explore whether the complexity of the search process increases when interactions are completed over an audio-only channel. In order to study all the above, we created the first SCS dataset, *SCSdata*, on which further research was conducted to enable us to address these challenges.

In summary, this dissertation unpacks the breadth and complexity of SCS. We first explore how people access media with an existing but limited audio-only interaction system. This investigation helps us focus our research problem and highlights the importance of methodological rigour for SCS. Second, we explore results presentation preferences through manipulating the length of summaries. We propose a novel crowdsourcing methodology which can be used to investigate results presentation manipulations, including manipulations such as prosody and listenability. Thirdly, we define a methodology for creating conversational datasets, propose rigorous transcription and analysis protocols, and develop the *SCSdata*. The empirical observations from the *SCSdata* are used to understand how people behave in this new search paradigm: we demonstrate that the system needs to be actively involved to overcome the difficulties posed by the audio-only channel. We then continue to apply our qualitative methods to identify the range of atomic actions which take place in this highly interactive search process. We validate these actions with a different dataset. Furthermore, we use these actions to create the first annotation schema for SCS: the *SCoSAS* which allows us to investigate the interactivity in the dataset. Then, we use the *SCoSAS*-annotated *SCSdata* to investigate behavioural patterns in SCS. We study the impact of task complexity on interactivity and discourse utterances. Along with our findings, we propose new research avenues and design recommendations for SCS which are envisioned to also impact non-audio-only search interactions.

To study search results presentation, we developed a novel experimental design using a crowdsourcing framework which allowed us to obtain insight into users' preferences in the information exploration stage over an audio-only channel. We then undertook an observational study which was analysed using qualitative methods (thematic analysis) to determine the components or actions of an information-seeking process in a SCS setting. Thus, both quantitative and qualitative methods were used.

1.3 Contributions

This thesis explores and describes which facets or components are key in searching over an audio-only communication channel. It addressed two research questions: (i) How

should search results be presented in an audio-only communication channel? and (ii) How do people search in audio-only communication channel?

Our main contributions in this dissertation are the development of: a novel methodology using qualitative and quantitative methods which can be replicated in future research; a crowdsourcing framework to evaluate results presentation user preferences; an annotation schema for SCS; and design recommendations for SCS systems. In particular, our contributions are focused on three outcomes: (i) practical contributions which can have a direct impact on the development and research for SCS; (ii) conceptual contributions which extend the wider discussion on IIR by exploring how conversational assistants can support users; and (iii) methodological contributions to investigate SCS. These outcomes are the following:

- **Practical outcomes:**

- *Recommendations for logging audio-only interactions:* The analysis of an interaction log accentuates the need for extra interaction log guides. We present our recommendations in Chapter 3.
- *Dataset for SCS, SCSdata:* We released the SCSdata², and our publicly available dataset can be used for further evaluation and exploration by other researchers. The development of the dataset can be found in Chapter 5.
- *Annotation schema for SCSdata, SCoSAS:* We released an annotation schema based on the SCSdata, the SCoSAS, together with the annotated SCSdata. The development of the SCoSAS is explained in Chapter 7.
- *SCS design recommendations:* We introduce a novel set of design recommendations for SCS in Chapter 9.

- **Conceptual outcomes:**

- *Identification of increased complexity, interactivity, and pro-activity:* We establish that SCS needs to incorporate interactivity and pro-activity to overcome the complexity that the information seeking process in an audio-only channel poses in Chapters 6–9.
- *Recognition of discourse interactivity in SCS:* We formulate the need for discourse markers in audio-only search interactions to overcome communication breakdowns in complex tasks in Chapter 8.
- *Schematic SCS model:* We propose the first schematic model to abstract a complicated interaction process of SCS based on the SCoSAS in Chapter 9.

²http://bit.ly/SCSdata_thesis

- **Methodological outcomes:**

- *Novel crowdsourcing framework to investigate different results presentations:* We present our original crowdsourcing setup to investigate the impact of different summaries over an audio-only communication channel in Chapter 4.
- *Methodology for the development of SCS datasets including data collection setup, questionnaires, semi-structured interviews, and transcription methodology:* We developed a full methodological setup to create SCS datasets including all necessary tools which are explained in Chapter 5.
- *Methodology for analysing SCS data, annotation schema creation, and validation processes:* The process of analysis of a SCS through qualitative methods is established and presented in Chapter 5.

1.4 Thesis Structure

This thesis is organised in four parts and their corresponding chapters.

Part I – Thesis Overview and Background

Chapter 1 – Introduction: We outline and formalise the SCS problem, the challenges, and the scope of this research, as well as our practical, conceptual, and methodological contributions.

Chapter 2 – Background: We discuss prior research related to SCS, contextualising and combining different research fields such as linguistics and SDS to overcome the research gap in SCS.

Part II – User Preferences in Results Presentation and Access over an Audio-Only Communication Channel

Chapter 3 – Accessing Media Via an Audio-only Communication Channel: We conduct a log analysis from an audio-only interaction application. The analysis provides an initial examination of the communication and interaction behaviours in an audio-only environment. The study amplifies the challenges of analysing and designing such audio-only interactive systems.

Chapter 4 – Results Presentation for an Audio-only Communication Channel: We investigate the impact of search results summary length over an audio-only communication channel. We collect results presentation preferences for audio and text summaries, and show that users prefer longer, more informative summaries for text. However, this is not observed for audio-only

summaries. We also contribute by creating a reusable crowdsourcing framework to test search results presentation.

Part III – Towards a New Model of Spoken Conversational Search

Introduction to Part III: We provide the aims and purposes of our observational study and exploratory observational analysis. Then we present an overview of the qualitative method used, thematic analysis, and the validation steps. We conclude the introduction to Part III with an overview of the overall approach and setup of the following experiments.

Chapter 5 – Methods: We describe the methodology of our observational study including the experimental approach of the data collection to create the first SCS dataset, SCSdata. We specify the transcription methodology converting the audio and video recordings to text. We then describe data analysis and annotation methods used to create SCS annotation schemas.

Chapter 6 – Observing Spoken Conversational Search Interaction Behaviour: We discuss observational findings from the interactions of the SCSdata. The empirical evidence is described in relation to search interactions and non-search interactions which occurred between the participants in their information seeking conversations.

Chapter 7 – Identifying, Classifying, and Validating the Interaction Space for Spoken Conversational Search: We present the development of the annotation schema for SCS; the SCoSAS. This annotation schema reveals the different atomic actions or utterance functions and interactions taken by participants in an information seeking process. We then continue to validate our annotation schema with a similar dataset.

Chapter 8 – Task Complexity and Interactivity fo Spoken Conversational Search: We investigate the interactivity between the identified atomic actions in the SCS data in relation to different task complexities. We show that in more complex tasks a greater number of interaction behaviours are exhibited including an increase in discourse utterances.

Part IV – Discussion

Chapter 9 – Recommendations for the Design of Spoken Conversational Search Systems: This chapter triangulates and discusses the findings of the studies and presents schematic models of SCS while emphasising the increased complexity, interactivity, and pro-activity in this new search paradigm. We also provide SCS design recommendations.

Chapter 10 – Conclusion and Future Work: We summarise the main conclusions and contributions of the work. Additionally, we outline implications for both IR, IIR, and the wider research community. We conclude with suggested extensions to our work and recommendations for future research.

Finally, the thesis contains four appendices with complementary information about ethics approval and participant information statement (Appendix A), questionnaires and semi-structured interview questions for the observational study (Appendix B), SCSdata (Appendix C), and SCS interaction themes (Appendix D).

Chapter 2

Background

In this chapter, we provide the background to relevant previous work.¹ We first provide information on Spoken Conversational Systems and its search instance, Spoken Conversational Search (SCS) (Section 2.1). We then introduce the importance of interactivity in Information Retrieval (Section 2.2). We review interaction and discourse-behaviour based tasks and task complexity. We continue with a discussion of information seeking processes and models which are relevant to conversational interactions and conversations in information seeking (Section 2.3). Next, we outline two fundamental search actions for interacting with search systems, namely queries and results presentation concerning speech interactions (Section 2.4). Finally, we present advantages, disadvantages, and concerns related to speech user interfaces in general, including an introduction to SDS, dialogue analysis, and the interaction space in conversational search (Section 2.5). We conclude with a conclusion and summary in (Sections 2.6 and 2.7).

This background chapter illustrates the intersection of SCS with many different areas including SDS, IIR, and linguistics.

¹This chapter consists of the following publications J. R. Trippas. Spoken conversational search: Information retrieval over a speech-only communication channel. In *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*, page 1067, 2015, J. R. Trippas, D. Spina, M. Sanderson, and L. Cavedon. Results presentation methods for a spoken conversational search system. In *CIKM'15 First International Workshop on Novel Web Search Interfaces and Systems (NWSearch'15)*, pages 13–15, 2015, and J. R. Trippas. Spoken conversational search: Speech-only interactive information retrieval. In *Proceedings of Conference on Information Interaction and Retrieval (CHIIR)*, pages 373–375, 2016.

2.1 The Rise of the Spoken Conversational System

With a Spoken Conversational System, people may converse with their smart devices (e.g., smartphones, watches, or speakers) in a natural way to retrieve information, issue commands, or access services in which the system responds in an everyday spoken fashion. Thus, a Spoken Conversational System is a broad term for any system which enables users to interact over speech (i.e., voice) in a conversational manner.²

Researchers in areas such as speech technology and artificial intelligence have long anticipated and worked towards Spoken Conversational Systems. Until recently, the expected ease of using Spoken Conversational Systems was only accomplished in science fiction movies, such as *2001: A Space Odyssey*, *Star Wars*, or *Star Trek*. Apple broadcasted their concept video of the Knowledge Navigator³ in 1987, a software agent who assisted the user in tasks such as search, planning, or communication. This software agent included advanced text-to-speech (TTS), natural language processing, and speech understanding. More than ten years later in 2001, [Berners-Lee et al.](#) envisioned the future of the Semantic Web in which agents could take advantage of the hypertext link universality [29]. However, it was not until the introduction of Siri in 2011 that Spoken Conversational Systems received extensive attention.

Many technological improvements influenced the progress in Spoken Conversational Systems. For example, the recent advances in artificial intelligence powered the development in language technologies such as spoken dialogue management, natural language learning, and speech recognition [77, 177, 219]. Furthermore, our smart devices have increasingly become more capable, and we are connected to even more powerful processors through being continuously linked to the internet. Although we have advanced in many technological aspects for Spoken Conversational Systems, more work has to be completed before these systems are genuinely conversational.

2.1.1 Spoken Conversational Search

A search system aims to help users find relevant documents or information units for their expressed information need. Users formulate and express their information need for which a system will retrieve relevant documents or information units. The system then presents the retrieved information as representations of the documents or as surrogates. Users need to make choices and relevance judgements about documents by eliminating or keeping retrieved documents for further inspection. To have a search system help the

²In this thesis we use audio, speech, and voice interchangeably.

³http://bit.ly/know_nav

user, the system needs to determine which documents or information units may be of interest to the user.

Browser-based search interactions consist of two primary interactions, user-queries and system-results [212]. The most common approach for users to express an information need is through a query submitted to a browser-based system in a search box. Then, the system returns a ranked list with results for the user to inspect. This list is ordered by the results' calculated relevance to the query. The concept of "user-query" and "system-results" interactions dates back to when librarians acted as the intermediary to the documents and were able to elicit the users' information need. The idea of this basic query-results paradigm as primarily atomic actions is still used in many search applications [212]. Other actions include navigational or recommendation actions such as query and document suggestions.

In contrast to the browser-based query-results search paradigm, SCS supports spoken exchange as the mode of interactions. Thus, the users can ask the SCS system to help them through their search process. The ability of a system to converse with the users arguably increases the usefulness of the system to help the user with their information need. For example, an information seeking conversation may look like this if a user is looking for information on solid beeswax perfume:

USER: How do I make a block of beeswax into perfume?

SYSTEM: Would you like to make solid beeswax perfume or beeswax scented candles?

USER: Uhm, I would like to make beeswax into perfume blocks.

SYSTEM: OK, solid beeswax perfume blocks are made from beeswax, almond oil, and essential oil.

USER: Can I use them with like, normal perfume?

SYSTEM: ...

We illustrate that interactions from the query-results search paradigm, where results are presented in a ranked list, are unlike the SCS system. Instead, the interactions from a SCS system can be sequences with questions such as information requests, refinements, or elicitations. Providing answers to a user's information request is an alternative interaction form to the ranked results list. Through the process of expressing their information need (even if it is ill-formed) and receiving possible results, the user may be able

to clarify their information need. However, the possible atomic actions have not been mapped based on empirical data [19].

A distinction between conversational search in *text* and *audio* has to be made. Researchers have suggested that people express ideas differently when they talk than when they write [3, 60, 224]. In particular, written and conversational prose have differences in their lexical diversity [30]. Thus, the mode of information exchange is crucial when discussing or studying conversational search. We do not consider searchbots (i.e., chatbots that perform specific types of searches), conversations in forums over a visual domain, or sequential modelling of user–system interactions in this thesis [17, 145, 207, 228].

SCS is concerned with open domain multi-turn *verbal* natural language exchanges between user(s) and the system. Ultimately, the SCS multi-turn exchanges are mixed-initiative, meaning that systems also can take action or drive the conversation. The system also keeps track of what has been said avoiding asking the user to repeat previous statements. Thus the user’s information need can be expressed, formalised, or elicited through natural language conversational interactions. The system pro-actively supports the user’s search process, and responds with cognitively processable replies to the user which are relevant to their context.

Conversational search has been identified as an important new research direction at several meetings including the last two Strategic Workshops on IR [6, 62]. The new “Conversational” sub-area in IR and IIR has gained much interest. For example, there is a growing interest in SCS systems that go beyond “command and control” utterances from users and keep track of what has been said, in session and over multiple sessions, and thus, go further than one-turn exchanges in a multi-turn manner. At recent workshops⁴ it was indicated that there is a lack of understanding of search tasks, search result description, and evaluation of SCS [172]. More importantly, the IR community lacks a broader insight into how users will engage with these highly interactive search systems and which components may be involved.

We define SCS with the properties presented in Table 2.1.

2.2 Interactivity in Information Retrieval

Research which explores developing, evaluating, or indexing information is traditionally categorised as IR. This research mostly does not involve real people. IIR is concerned with the interaction between the system and the user, while IR is more system

⁴International Workshop on Conversational Approaches to Information Retrieval (CAIR) at SIGIR 2017 and 2018 (<https://sites.google.com/view/cair-ws/>)

TABLE 2.1: Identified SCS requirements.

		SCS
1	<i>Analogy</i>	Human intelligible dialogue-like, beyond command and control
2	<i>Language</i>	Spoken natural language, conversational
3	<i>System participation</i>	Pro-active, mixed-initiative (implies listening)
4	<i>Information request length</i>	Longer, more natural
5	<i>Results presentation mechanism</i>	Adaptive to users' need and context (ranked list is inadequate)
6	<i>Turn-taking</i>	Multi-turn
7	<i>History</i>	Over (multiple) sessions

focused [32, 106]. Indeed, “Interactive” signifies the involvement of a human in comparison to system-oriented approaches in other sub-fields of IR, often referred to as the Cranfield paradigm [55]. In particular, IIR’s core aim is to study how people use search systems to satisfy their information need [155] and which tasks play a fundamental role for evaluation [31].

2.2.1 Task and Task Complexity

When people interact with an IR system, they usually do so within the context of a task, defined as a piece of work which often needs to be completed in a specified length of time.⁵ Tasks, scenarios, simulated work tasks, or backstories are therefore widely used in different evaluation settings involving people, such as in human-computer interaction, strategic planning, or IIR [31, 44, 81, 213]. These scenarios provide a context for the participant to conduct the assigned task. A task or scenario often contains an actor, some background information on the actor, the goals or purpose of their action, and occasionally some sequences of actions the actor to perform [81]. Depending on the goal of the task or scenario, some of the mentioned components may be discarded.

Tasks are often used as a representation of the search goal or purpose and symbolise what the user wants to achieve with their search. The advantage of providing tasks to research participants is that tasks can be manipulated as part of the research design [213]. However, many tasks in IIR experiments are created by the researcher and therefore may not represent the searcher’s internal information need. Thus, creating tasks which can be widely used, naturalistic, and applicable to the participants is challenging and time-consuming [108].

⁵<https://www.merriam-webster.com/dictionary/task>

Tasks or scenarios enable researchers to observe the current interaction behaviours between a system and real users [81]. In this thesis, tasks will be used to assist the evaluation of SCS interactions. The outcome of an evaluation study, and thus the observed behaviours, can be affected by the characteristics of the task, such as task complexity.

2.2.1.1 Task Complexity and Search

Much research has been devoted to developing tasks and exploring their impact on search behaviour [44, 95, 105]. Task complexity is often used to indicate which cognitive resources are needed to fulfil the task. Jansen et al. used the revised Anderson and Krathwohl's taxonomy of the cognitive learning domain to create tasks requiring different levels of mental effort (i.e., cognitive complexity) [9, 97]. This taxonomy has six levels of increasing complexity: remember, understand, apply, analyse, evaluate, and create.

Many researchers have used task complexity in research to study search behaviour [11, 16, 46, 108, 218]. For example, Aula et al. investigated the behaviours in participants' search interactions when they were engaged in tasks in which the answer was hard to find [16]. They showed that when participants had difficulty finding information, participants formulated more diverse queries, used Boolean and advanced operators more, and spent more time on the SERP. Kelly et al. showed that participants engaged in more interactions such as more queries, clicks, and time on task as task complexity increased [108]. These studies illustrate the importance and influence a task can have on the interaction behaviour.

2.2.1.2 Task Complexity and Discourse

Tasks, task difficulty or complexity, and discourse (i.e., the communication of a series of linked utterances) have been studied extensively in areas such as linguistics and pedagogy [59, 79, 151]. For example, Gilabert et al. investigated the impact of increasing task complexity on interactivity in learner's communication behaviour [79]. They showed that different task types affected communication behaviours, with more complex tasks generating more interactions. Other research has suggested that both lexical behaviour and the use of confirmation checks increase as tasks become more complex [151].

It has been proposed that the increase in a task's cognitive demands generates more communication breakdowns and therefore increases the number of interactions to repair these breakdowns [152]. The researchers suggested that these breakdowns occur because of demands placed on the cognitive resources, which are needed to solve the task itself. As a result, there are fewer cognitive reserves available for maintaining the task discourse

interactions. Despite the disrupted communication, these breakdowns may also bring a positive side-effect [79, 143]. It has been suggested that the extra interactions to solve the communication malfunctions may lead to further negotiations about the meaning of a message [143]. These negotiation functions can contribute to improved accuracy of the information exchange [143].

Research in computer-mediated communication has also shown that task complexity influences discourse behaviour [59]. That is, more complex tasks required more discourse, particularly, more meta-communication (i.e., the conversation about the communication). Furthermore, in that study, task complexity impacted on adjacency pairs (i.e., discourse routines or bigram interactions) which indicates that discourse interactions were essential for the conversation and that these discourse routines (i.e., bigrams) could be exploited to predict interaction pairs [59]. All the above research suggests that task complexity impacts the communication behaviour on interaction, discourse, and meta-communication activities in decision-making and learners' conversations. Future research into discourse routines may help predict interaction behaviour in SCS and change interaction techniques according to task complexity. These are areas addressed in this thesis.

2.3 Information Seeking Processes and Models

Models are an abstraction of reality and are often used before the development of a formal theory [48]. Models are regularly displayed in diagrams or flowcharts with the aim of making them easier to understand and enabling researchers to focus on specific problems [48]. Information seeking also uses models to explain or abstract what is observed in the search process, making it easier to recognise if hypotheses are consistent with real-life observations [148, 212]. As Wilson describes, most models (i.e., mostly diagrams) in information seeking are explanations describing the information seeking actions, their motivation and outcomes, or their relationship with other states [215].

Information seeking is well studied in IIR and often adopts a search model process or cycle which includes the user's recognition and definition of their information need, the examination of results, and the reiteration of the process until the user's information need is satisfied [127]. Many researchers have studied and formed models of this process (e.g., Belkin [22], Ellis [72], Kuhlthau [116], Marchionini [126], Saracevic [158], Wilson [215]). These models were often derived or based on observations of how people worked through their search process alone, in specific environments, or how they interacted with intermediaries (i.e., reference librarians) [91, 212].

One of the general models described by Marchionini and White [127] defines the information seeking process as consisting of:

- Recognising a need for information,
- Accepting the challenge to take action to fulfil the need,
- Formulating the problem,
- Expressing the information need in a search system,
- Examining the results,
- Reformulating the problem and its expression, and
- Using the results.

The above stated actions are often said to form the core information seeking actions [91]. It is generally accepted that search engines support the user in their expression of the information need, examination of the results, and to some extent the reformulation of their problem [91].

Other models are more focussed on the psychological processes during the search process. For example, [Belkin's](#) Anomalous States of Knowledge (ASK) hypothesis explains the user's information need from a cognitive viewpoint [22]. Thus, ASK states that users experience a gap or anomaly in what they know and what they would like to know. To fill that gap, users need to obtain information until the anomaly is resolved. According to the ASK hypotheses, only once a user has identified a gap, can they start formulating their information need. Other researchers, such as [Taylor](#) observed and proposed a similar concept [179]. [Taylor](#) divided the expression of an information need into four stages that the user works through to formulate a query which can be submitted to a search engine [179]. These four stages of expressing an information need are:

1. **Visceral:** The need for information is formed.
2. **Conscious:** A mental description of the information need emerges.
3. **Formalised:** A formulation of the question is formed.
4. **Compromised:** A formulation of the question is formed in a way it can be presented to a search engine.

Another kind of model is [Saracevic's](#) stratified model [158]. The elements in the stratified model are related to the user and system, each with different levels or *strata*, discou

through an interface. Thus the stratified model includes the interactions between the user and system as dialogue interactions, with each participant bringing their own layers of specifications to that dialogue. For example, users bring their levels of cognitive, affective, and situational influences and the system brings its hardware, data processing, and structures. Their interaction is the exchange between each's strata. The critical point made by the stratified model is that the strata are not independent of each other and that the weakest point in the user–system relationship can impede achieving the best outcome for the search process.

2.3.1 Modelling Information Seeking Through Dialogue

Even though little research has been devoted to the new search paradigm of SCS, early work in the 1970s in man-machine IR through dialogues was introduced by Oddy [139]. A reference retrieval program called THOMAS was developed which aimed to help users select documents without explicitly formulating queries. Instead, THOMAS helped the user narrow down their search scope by asking questions and presenting suggestions. The program only displayed the relevant documents at the end of the question–answering process which can be interpreted as the conversation. Croft and Thompson designed the Intelligent Intermediary for Information Retrieval (I³R) in the 1980s where the system is modelled on an expert intermediary [61]. In contrast to a search system which allows a user to search with a single retrieval strategy (a query), the I³R system supported the user with domain knowledge acquisition, explanation, browsing, retrieval, and evaluation. The system could also confirm or request more information from the user in some kind of dialogue for unspecified information needs.

Other researchers also proposed ways to incorporate searching for information through dialogue but with the use of Dialogue Acts (DA) [168, 173]. DA are a schema which represents the generic meaning of an utterance. For example, Sitter and Stein developed the CONversational Roles (COR) model [168] based on DA as a general model for information seeking dialogue and combining it with a dialogue plan (a list of pre-defined intended dialogue actions) [7]. The plan is then used to guide users through stages of information seeking. The model is shown in Figure 2.1. In this model, the actors are noted as A (information seeker) and B (information provider). The circles and squares symbolise the states as part of the dialogue. Arrows represent the progress between the states. For example, in step ① the seeker makes the first move with the possible outcomes outlined in example ②. This atomic move is annotated with DA as *request(A,B)*.

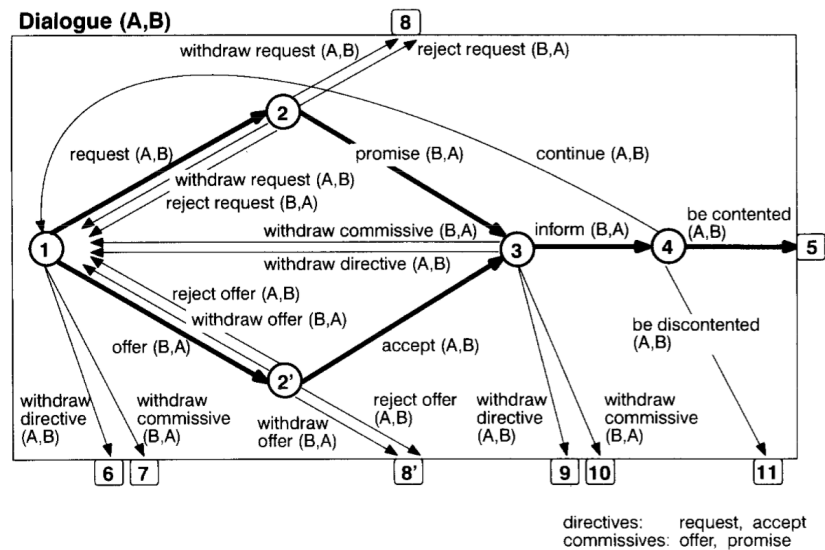


FIGURE 2.1: COR model by Sitter and Stein [168].

Other features of the COR model include the flexibility in *mixed-initiative*, meaning that at any given time one of the actors can decide what happens next or ask questions. Mixed-initiative dialogues allow for a more natural interaction but are more complicated for the system to handle [132]. The model also allows for meta-communication by permitting the conversation to go through one of the loops at any point in time. Nevertheless, only one move in an utterance is possible according to the COR model, making the model unaccommodating for the flexibility of voice input and output.

A more recent DA-based model for information seeking dialogues is the Query, Request, Feedback, Answer (QRFA) loops by Vakulenko, Revoredo, Di Ciccio, and de Rijke [205]. These are four User-Agent feedback loops trying to explain the conversational flows based on real data. Similarly to the COR, the QRFA aims to provide the structure of a single dialogue contribution or move from the actors in the conversation. Nevertheless, these aforementioned schemas are based on DA and only provide broad categories of the action taken in that utterance. These DA-based models can be applied in any common dialogue and fail to reveal further possible interactions between a user and system in a SCS process. Furthermore, these simple actions are uncommon in more complex information seeking situations [25].

Other studies have focused on discourse aspects of conversations without using DA. For example, Belkin et al. proposed a coding schema to annotate communications between librarians and users to better understand the design of expert systems [24]. Their schema showed that one could extract a range of contextual information from dialogues, including the description, states, modes of problems at hand, user models, search strategies, and search interactions. Later, Belkin et al. introduced a concept of scripts that described functions of dialogues and applied them to the design of IIR systems [25].

The authors argued that, depending on the kind of information need, different interactions may be appropriate to provide an (ideal) abstraction of the problem and enable an understanding of the question, from which responses (scripts) could be created.

Thus the COR [168], QRFA [205] models and scripts [25] enable the prediction of which kind of interaction will be necessary following from a previous move. These predictions are a form of discourse routines and can become “predictable” defaults adaptable to maximise efficiency by demanding minimal encoding of the system. Hence, if we could predict and simplify the input from the user, we may be able to provide appropriate responses generated by the system.

A more relevant conceptual framework of User-Agent actions was recently created by Azzopardi et al. [19]. This framework combined the action and interaction space discussed in Radlinski and Craswell [146] and Trippas et al. [199]. The conceptual framework, therefore, is not restricted to the DAs but provides a broad overview of the potential actions taken by either actor as shown in Table 2.2. Nevertheless, Azzopardi et al.’s conceptual framework still needs to be empirically validated [19].

TABLE 2.2: An overview of the actions and interactions hypothesised in Azzopardi et al. [19].

		User	Agent		
Query Formulation [199]		Reveal	Inquire	User Revealment [146]	
		Disclose			
		Non-disclose	Extract		
		Revise	Elicit		
		Refine	Clarify		
	Expand				
Set Retrieval [146]	Results Explorations [199]	Inquire	Reveal	System Revealment [146]	Memory [146]
		List	List		
		Summarise	Summarise		
		Compare	Compare		
		Subset	Subset		
	Similar	Similar			
		Navigate	Traverse		
		Repeat	Repeat		
		Back	Back		
		More	More		
			
	Note	Note			
Mixed Initiative [146]		Interrupt	Suggest		
			Recommend		
			Hypothesise		
		Interrogate	Explain		
	Understand	Report			
	Explain	Reason			

2.4 Fundamental Search Actions Through Audio

The principle way in which people interact with browser-based search systems is by expressing information requests or queries and investigating ranked results lists. Indeed, the main interaction mode between user and system is through *primary* (or *atomic*) *actions* which have been itemised as *search queries*, *selection recommendations* (query and document suggestions), and *item selection* by White [212]. Many of these interactions can be logged and these logs assist the training of algorithms such as query suggestions. In this section, we review the actions of spoken queries (i.e., search queries), and results presentation and answer organisation (i.e., selection recommendations) through audio.

2.4.1 Spoken Queries

Traditionally, search systems have been receiving browser-based written search queries from users which represent the user's information need. These written queries can be expressed through Boolean statements or operators, but overall are mostly short statements of the user's intent. The system then uses these queries to retrieve information units or documents and presents these to the user. While the process of submitting written queries is relatively easy⁶, spoken queries first need to be processed by Automatic Speech Recognition (ASR) to become text representations.

Research in voice queries has often compared text and voice queries based on log analyses or lab-based experiments [13, 60, 86]. This research has shown inconsistencies in results [87]. For example, Schalkwyk et al. reported that voice queries were shorter than typed queries (2.5 versus 2.9 on average, respectively). However, other studies found voice queries to be longer than the average text queries of 3.2 words [60, 86, 225]. Furthermore, Guy also reported that voice queries have many other unique characteristics such as: they are closer to natural language, the topics are different, and user behaviour (time of use and clicks) differs [87].

2.4.2 Results Presentation and Answer Organisation Through Audio

Web search systems most commonly display search results in a vertical list which summarises the top-ten retrieved documents. This list is often referred to as the SERP. One item on the SERP consists of a document title, a short summary (i.e., snippet), URL, and often other meta-data such as date or author. Such representation of a document

⁶Note: This is in contrast to the expression of one's information need which can be very challenging.

is also referred to as the document surrogate and aims to help the user understand the meaning of the underlying document [127].

Studies have shown the importance of how to present the document surrogate and its usability. For example, it has been suggested by [Clarke et al.](#) that all query terms should appear in the surrogate to reflect their relationship with the underlying document; that when query terms are present in a title, they do not need to appear again in the summary; and that the URLs should be displayed in a less complicated manner while showing their relationship to the query [54].

Other researchers have examined the snippet length to understand the trade-off users are willing to accept between the length of the snippet versus the snippet's informativeness. [Cutrell and Guan](#) investigated the effect of different snippet lengths (short [1 text line], medium [2-3 lines], and long snippets [6-7 lines]) [63]. They found that for information queries, the performance improved if the length of the snippet increased. However, the performance degraded for navigational queries. Later work from [Kaisser et al.](#) also suggests that different types of queries benefit from an optimised summary length [102]. More recently [Maxwell et al.](#), indicated that users preferred longer, more informative summaries as they were perceived to be more informative, even if they did not contribute more to helping users correctly identify relevant documents [129].

2.5 Speech User Interfaces

Human speech is the most widely used form of communication, as well as the most complex one. Even though, human speech is considered a natural way to interact, speaking to a computer is still mostly seen as “unnatural” [114, 203]. However, with the recent developments of spoken interactive systems such as Apple's HomePod or Google Home, speaking to a computer is becoming more widely accepted. The use of speech systems in particular situations, such as when one's eyes or hands are busy [56], allows for information to be accessed without requiring a keyboard or typing [132, 224]. In addition, these speech systems can be used by people who may otherwise be unable to access information via text, such as visually impaired people or people with dyslexia [156, 203].

It is important to address user needs, including the users' context, to improve data access through intelligent information systems [84]. For example, when users are presented with search results to their query in a visual representation, the search query terms are highlighted [91]. [Ajmera et al.](#) argue that when search results are presented by speech, audio feedback could be used to display whether a specific query term shows up in the query results [2]. Other studies have indicated that a notifying sound could be used

instead of speech feedback [204, 224]. Winterboer et al. [216] tried to implement a similar approach whereby a beep was used as a discourse marker to help users compare options. Other concerns which have been identified concerning speech user interfaces are:

- The user talks before the system is ready.
- The user reads meanings in pauses while the system is still working [117, 224].
- The user might find it easier to produce speech output than consume speech input [71].
- The user might not know what to say [224].
- The information must be presented sequentially [67].
- The fact that speech output is easier to forget than written output [117].
- The trade-off between presenting enough information to the user (confidence for a good overview of search results) and keeping utterances short and understandable might be unsatisfying for users [149].

Overall, we recognise that speech user interfaces have many challenges which need to be overcome to facilitate a good user experience. All these challenges are inherently present in SCS interactions and impact on the behaviours and limitations of the audio-only channel.

2.5.1 Spoken Dialogue Systems

A SDS is an instance of a speech user interface. Such systems provide a platform for people to interact with computer applications such as databases with the use of spoken natural language. SDS exchanges information on a turn-by-turn basis providing an interface between the user and the computer [78]. Extensive research has been conducted into how to best present information and interact over audio [78, 132]. For example, researchers have investigated the cognitive resources users need to interact with SDS and have suggested that instead of just reading out results, SDS should help the user make decisions by providing suggestions [206] or providing an overview of (ir)relevant options [66]. It has been suggested that this may make the user feel in control of having heard all possible options.

In recent years, interest in SCS has grown, as speech technology [219] and machine learning for spoken systems [220] have developed. A range of SDS are available, from

question answering to semi-conversational systems [132]. Research has been devoted to task-oriented SDS which has defined search boundaries, such as travel planning or route planning, and can be developed with slot filling approaches [209].

Task-oriented dialogue systems are created on a particular closed domain. However, non-task-oriented dialogue systems or open-domain conversations such as search for SCS systems may not benefit from a rigid plan-based dialogue approach and introduce many new challenges [92]. These challenges include how to deal with the variety of user utterances and how answers or replies could be simplified or abstracted to generate appropriate system responses [178].

2.5.2 Dialogue Analysis

Research interest in SCS has increased the recording of spoken search interactions [185, 205]. Such records are a valuable source of data to understand how users interact in this unique search paradigm and which tactics are used for driving effective search performance. Thus, this data is useful to understand the characteristics of a search conversation to build SCS systems which can act as a dialogue participant [78]. The spoken data recordings themselves are of limited value and these recordings need to be appropriately transcribed and “annotated” [120]. Thus, exposing the structure of the conversations by annotating the actions taken is one of the first steps towards analysing these spoken interactions [227].

Previously, much research has been devoted to creating annotation schemas and classifying taxonomies for dialogues and SDS [7, 41, 164]. These annotation schemas are often developed for speech but are also applicable to written conversations such as online discussion forums. Annotating these dialogues has been based on the understanding that classifying utterances provides insight into the dialogue behaviour [147]. For example, annotated conversations can help to identify answers in texts and unanswered questions which need to be addressed, as well as characterise user intents or model which actor plays a particular role in a conversation [112, 144].

Many different annotation schemas have been proposed which cover the general speech interactions. Some schemas emphasised information seeking, such as the Dynamic Interpretation Theory (DIT) by Bunt [40]. The DIT was based on the empirical investigation of spoken human–human information dialogues. Bunt suggested that these information dialogues have two motivational sources, namely to proceed in the task and to exchange communicative functions to drive the conversation [40]. He noticed that an information dialogue consisted of the expected greetings, apologies, and acknowledgements but also included information-exchange utterances such as questions, answers, checks, and

confirmations. Later, Bunt developed an annotation schema called DIT++ for these information dialogues [41]. Nevertheless, DIT++ lacks the detailed distinctions made when a user interacts with a search system while satisfying their information need, for example the techniques used to represent documents or information units.

2.6 Conclusion

As demonstrated in this chapter, research on conversations is not new in IIR. However, there is a resurgence of interest in SCS, especially in abstracting and defining conversational interactions, which we refer to as the *conversational search revolution*. We reviewed previous studies in spoken (semi-)conversational models highlighting the lack of models which combine the unique aspects of SCS. For example, the previous models do not cover multi-turn, open domain, natural language exchanges in which a system can take the initiative. Such taking of initiative by the system implies that the system actively listens and keep track of the interaction history. These abstractions and definitions enable researchers to gain understanding of characteristics, descriptions, and structures of the interaction itself, facilitating the specifications of the SCS system design.

2.7 Chapter Summary

In this chapter, we reviewed prior studies related to conversational search and how this conversational search differs from spoken or web-based search. We began this chapter by explaining where conversational systems are located historically, and the vision people had for these systems. We then reviewed previous research in IIR, highlighting the importance of studying task complexity and discourse interactions. With respect to information seeking processes and models, we outlined previous work in models which are relevant to audio-only SCS and illustrated that we are now in the *conversational search revolution* era. In addition, we examined the atomic actions which take place in search (i.e., queries and results presentation) with respect to speech input and output. Concerning speech user interfaces, we outlined some differences with visual user interfaces and presented previous research in defining the interaction space.

Given this overview of prior research related to SCS and the undetermined possible spoken conversational atomic actions, we present our first contribution to exploring spoken conversational interactions through log analysis of an existing but limited audio-only interaction application in the following chapter.

Part II

User Preferences in Results Presentation and Access over an Audio-Only Communication Channel

Chapter 3

Accessing Media Via an Audio-only Communication Channel

Studies of interaction log analysis are a common tool to investigate behavioural data and can contribute to insights of the interaction patterns of users with a system [167, 212]. We present the results of a log analysis from RealSAM¹, an audio-only interaction application. RealSAM is an accessible media assistant in which users can navigate and interact with media content through natural language. The assistant is designed for people with a vision impairment or other disability that prevents a person accessing printed material, and developed by Real Thing². The RealSAM log analysis is part of a Linkage project between RMIT University and Real Thing Entertainment Pty Ltd.

The exploratory analysis was conducted to provide an initial insight into the communication and interaction behaviours between users and this audio-only application. We focus on understanding how users utilise the application. The study reveals the challenges of analysing and designing these audio-only interactive systems, with implications for the design of future voice-enabled tools.

RealSAM allows users to interact through multi-turn audio-only interactions in their natural environment. Even though these interactions may be specific and limited to this particular application, we believe it also provides a starting-point for further analysis of SCS.

¹<http://www.realsam.com.au>

²<http://www.realthing.com.au/>

The insights gained from working with the logs influenced our experimental setup and analysis in Part III while complementing the discussions presented in Part IV. Thus, we combine the results of a real-life application from this chapter and lab-study from later chapters to provide a more holistic discussion of SCS.

This chapter is structured as follows. Section 3.1 provides an introduction to the RealSAM application and its target audience including an overview of the interaction methods and the content which can be accessed. We then describe the available dataset. Section 3.2 presents the results of the log analysis including the general and session descriptives of the interaction dataset. We show how RealSAM is used over time including the interaction frequencies based on pre- and self-defined interaction categories. We describe sessions which consist of a single interaction and introduce search interaction behaviours. We conclude the results section with displaying the TTS output settings of RealSAM users. Section 3.3 discusses the results and limitations of the study. Section 3.4 sums up the lessons learned during the analysis process which influenced our observational study's data-capture process presented in Part III. Finally, we conclude this chapter with a summary in Section 3.5.

3.1 Introduction

RealSAM is an application with which users can interact and search for audio material, such as podcasts, news articles, and audiobooks, exclusively via an audio-only interaction channel. The application is tailored to provide accessible media for people who are visually impaired. We use the RealSAM logs to understand the interaction behaviour between users and the application.

Displaying search results for people with a visual impairment is problematic. Systems such as Siri allow users with a visual impairment to pose queries, but they will not receive answers to their query via audio unless it is a factoid question. For non-factoid or ambiguous questions, this user group relies on additional assistive software (e.g., screen reader, VoiceOver³, or TalkBack⁴) to translate the written SERP into speech. Thus, a user with a visual impairment who uses Siri to search must switch to using assistive software to read out the search results. The volume of information read out also presents the user with cognitive challenges which may lead to unsatisfactory search interactions.

³<http://www.apple.com/au/accessibility/osx/voiceover>

⁴<https://support.google.com/accessibility/android/answer/6007100?hl=en>

Thus, much work remains to be done to allow equal information accessibility [156]. As the first step, we need to understand how users behave in an audio-only interaction setting which we do by investigating the interaction logs.

3.1.1 RealSAM Application

RealSAM consists of a Samsung Galaxy Pocket with a single-app Android ROM installed on it (see Figure 3.1). This device has a main button on the bottom front of the device which is the *talk button*. When users press this button, they can either start their spoken interaction or interrupt (i.e., *barge-in*) the device. The volume buttons on the device work, however, the other buttons and touch functionality of the screen are disabled for accessibility reasons. Users can also turn on a *hands-free mode* which allows them to interact with the device without having to press the talk button. However, in this mode, RealSAM will only start listening again after it has finished speaking and thus users cannot interrupt.



FIGURE 3.1: RealSAM device.

RealSAM provides the following five categories of content:

- **Podcasts:** Listen to podcasts from sources such as the Australian Broadcasting Corporation (ABC) or the British Broadcasting Corporation (BBC).
- **Newspapers:** RealSAM currently indexes news from ABC News, The Conversation, The New Daily, and a wide range of papers provided by the Vision Australia Library, including The Age, The Sydney Morning Herald, and The Australian.
- **Books:** RealSAM provides access to the books offered by The Gutenberg Project⁵ and Bookshare⁶.

⁵<https://www.gutenberg.org>

⁶<https://www.bookshare.org>

- **Service:** RealSAM allows users to check the current time, weather conditions, and geographical location.
- **Device:** RealSAM provides commands to configure the device, check the battery level, or listen to announcements from RealThing.

RealSAM uses sound cues (i.e., ear-cons or discourse markers) to guide the user through the system. For example, a *falling tone* and a *tick tock sound* means that RealSAM is considering the user's request and will respond soon. When a user submits a command, the device RealSAM presents the first five results to the user with an option to hear more results. Thus, one "result page" consists of five results. An example interaction is shown below.

USER: Which newspapers do you have?

REALSAM: I have the following newspapers:

1. ABC News
2. Adelaide Advertiser
3. The Age
4. The Australian
5. Australian Financial Review.

Please select one or say continue.

USER: Number 3

REALSAM: OK, selecting The Age. The first page of 29 unread headlines from the News Section:

1. Faulty fire system puts lives at risk
2. Mum's the word in Melbourne
3. Greens go for...

USER: Read me the Finance section from the Australian.

[barge-in]

Interactions with RealSAM are classified based on system-defined rules which are triggered by pre-mapped voice inputs. For example, RealSAM starts reading news headlines when a user inputs "*read me the news headlines*". As such, this interaction is classified by the system as "** headlines **". This is an illustration of the inherent linguistic and functional limitations of this restricted system.

3.1.2 Dataset

The log set includes interactions between 17 February 2014 and 17 May 2016.⁷ Input interactions can be seen as a voice command to the system. This voice command is then translated into a text command using ASR and from this point onwards it is treated as a text command. The output text is translated with TTS for the user to listen to. The audio output is in contrast to many multi-modal systems where the input is by voice but returns results using the standard mobile or desktop interface (i.e., the screen).

Each interaction or voice input has a timestamp (beginning of interaction), anonymised user ID, output interaction from the system, voice type and speed, and the system rule triggered by the input received. However, no information is recorded as to whether the user barged-in to the application and there are no end timestamps.

3.2 RealSAM Log Analysis

We first present the general descriptive statistics about the logged RealSAM interactions and examine the pre-identified RealSAM *Interaction Categories*. We then continue to group these Interaction Categories in *Super Interaction Categories* allowing us to investigate how people use RealSAM through communication, one-interaction sessions, and search sessions. The final part of this section discusses the user settings of the TTS output.

3.2.1 General and Session Descriptives

The RealSAM interaction log consists of 411,201 interactions from 236 unique users. An interaction comprises of an action from the user and a reaction from the system.

Interactions are grouped in sessions where a session lasts until there are at least 15 minutes of inactivity [96, 183]. The interaction log contains 46,859 sessions.

On average, users spent 19.74 minutes per session. The average sessions per user was 199 (median is 23). There were 8.77 interactions per session. A total of 24,507 sessions (52.29%) consisted of only one interaction.

When we examine the RealSAM session patterns over a 24-hour time frame, we observe that more sessions take place in the mornings throughout the 7-day week. However, when comparing weekdays and weekend days we notice a trend that users interact more

⁷The interaction logs cannot be made publicly available.

frequently with RealSAM during weekday morning hours than weekend mornings as seen in Figure 3.2. After 2pm on weekdays the number of sessions declines while on the weekends the number increases.

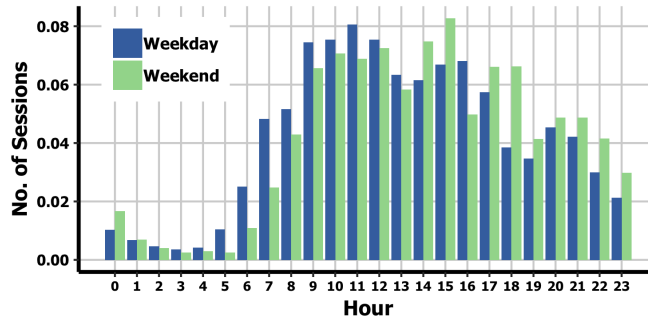


FIGURE 3.2: Normalised session frequency in 24 hours on weekday and weekend days.

3.2.2 How People use RealSAM

We removed all the stopwords including unrecognised voice input and create a frequency list of the most highly used terms.⁸ We found the most frequent term from the users was “next”, corresponding to 21.72% of the total input terms as seen in Table 3.1.

TABLE 3.1: Top-five frequent input terms.

Input Term	Count (%)
Next	143,399 (21.72%)
Number	38,569 (5.84%)
Read	22,980 (3.48%)
Headlines	18,681 (2.83%)
Back	16,653 (2.52%)

A total of 43,918 distinct pre-mapped rules were recorded in the log. We sorted these rules in 87 interaction categories including the categories “Null”, defined by the application, and “Other”, which we were not able to classify. The Null and Other categories accounted for 12.8% and 5.8% respectively of the total logs.

We sorted the remaining 85 interaction categories by investigating the voice input transcripts. For example, if the pre-mapped voice input recorded “* headlines *” we examined all rows within the log containing this particular input to conclude that this rule is indeed related to asking for news headlines. We then classified this pre-mapped input accordingly.

Thus, a total of 85 interaction categories were created with the most frequently used categories presented in Table 3.2. The table shows that several interactions are similar

⁸We used the SMART stopword list.

and could be categorised in a *super category*. For example, the category *next article* and *next response* are both navigational interactions indicating reading out the next response and therefore belong to the newly defined super category *Interaction Management*. The classifying processes were conducted iteratively by myself and reviewed by supervisors.

TABLE 3.2: Most frequently used RealSAM interaction categories.

Interaction Category	Count (%)
Next article	93,309 (27.88%)
Select response	52,365 (15.65%)
Next response	24,302 (7.26%)
News headlines	16,893 (5.05%)
ASR error recovery	16,819 (5.03%)

We grouped the 85 interaction categories through examination into super categories. These super categories create a further abstraction while reducing the number of categories for a more meaningful analysis. The interactions categories are divided into the next five super categories:

1. **Search (S)**: a user searches for a specific document,
2. **Browsing (B)**: a user wants to hear the news headlines,
3. **Interaction Management (IM)**: how a user interacts with the device, such as “next”, “stop”, or “resume spoken document”,
4. **Device and Service (D/S)**: interactions related to operating RealSAM, such as changing the voice or checking the battery and weather⁹, and
5. **Error Handling (EH)**: the device attempts to recover from errors.

Figure 3.3 shows that Interaction Management is the most commonly used. The second most commonly used is Error Handling followed by Device and Service. The high Interaction Management would be expected given that this category includes the commands to use the device such as resuming a spoken document, navigating to the next section, or repeating an article.

3.2.2.1 One-Interaction Sessions

As mentioned, 52.29% of the sessions consisted of one interaction. The 15 most frequent interaction categories cover 79.55% of the one-interaction sessions as presented in Table 3.3. The Search super category did not contain any one-interaction sessions.

⁹Weather information is stored on the server and classified as a service.

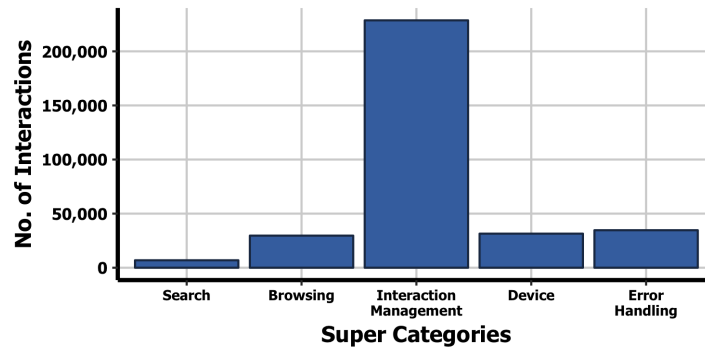


FIGURE 3.3: Interaction frequency of super categories.

TABLE 3.3: Most frequent one-interaction session categories.

Interaction Category	Super Category	Interaction Category Count (%)
Access source	IM	3,873 (17.43%)
Check the battery level	D/S	1,805 (8.13%)
Next article	IM	1,790 (8.06%)
No match found	EH	1,768 (7.96%)
Select response	IM	1,588 (7.15%)
Check the weather	D/S	1,426 (6.42%)
ASR error recovery	EH	1,120 (5.04%)
News headlines	B	780 (3.51%)
User guide	D/S	748 (3.37%)
List books	B	681 (3.07%)
Next response	IM	500 (2.25%)
Time	D/S	496 (2.23%)
Part of command missing	EH	478 (2.15%)
Response to “hello” input	D/S	332 (1.49%)
Go back	IM	287 (1.29%)

NOTE: Browsing (B), Device and Service (D/S), Error Handling (EH), Interaction Management (IM), Search (S)

3.2.2.2 Search Sessions

The Search super category consisted of 3,399 (7.25%) sessions where users posed one or more queries. The total number of queries in the Search super category was 6,888, consisting of 2,238 news article searches (32.49%), 2,106 podcast searches (30.57%), 629 book searches (9.13%), and 1,915 (27.80%) unclassified searches. These unclassified searches were due to users posing an unspecified query which the system could not classify in any of the specified interaction categories.

The average query length for the voice queries was 3.29 words ($SD=1.49$, $max=24$) which were obtained after lowercase conversion, tokenisation, and stopword removal and 4.87 words ($SD=1.88$, $max=31$) without the removal. Query characteristics presented in Table 3.4 show that 56% of the queries were unique.

TABLE 3.4: Query characteristics.

	Count
Total number of queries	6,888
Unique queries	3,872
Most frequent queries:	
Read articles about rugby	406
Read articles about wallabies	214
Play me the health report	69
Play me the science show	63

Table 3.5 shows the most frequent terms in search queries. Popular terms suggest that search was used as a mechanism to access specific sources (e.g., ABC, report, show) or to find content related to a given topic (e.g., rugby, wallabies).

TABLE 3.5: Most frequent query terms.

Query Term	Count	Query Term	Count
Rugby	622	ABC	214
Wallabies	288	Australia	199
Report	265	Science	179
Show	236	Health	172
Vision	232	Margaret	139

3.2.3 Text-to-Speech Output

This section investigates the voice and speed of the TTS output per interaction. A female Australian voice and 1.0x speech reading rate were the default settings but six different voices and other speeds are available. Interactions were performed 51.78% of the time in these default settings, where 60.12% had the default speed, and 71.82% used the default female Australian voice. Thus 39.54% of the interactions were performed either with a slower (18.26%) or faster (21.28%) voice speed (see Figure 3.4).

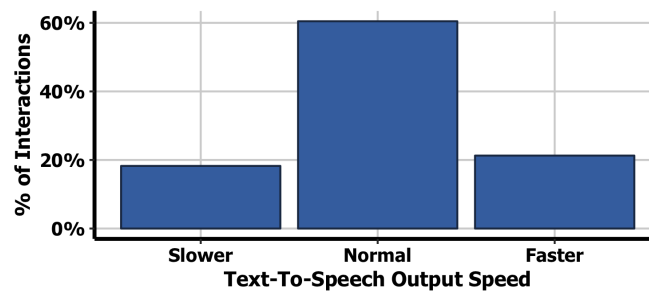


FIGURE 3.4: Speed of the output in the interactions.

3.3 Discussion

We now discuss the results of the log analysis, the use of navigational interactions, one-interaction sessions, search sessions, and speech output configurations. We conclude this section by discussing the limitations of this exploratory study.

The results show that navigational “next” interactions such as “*next article*” or “*next response*” were frequently used commands. We propose to add an “infinite-reading” mode to RealSAM allowing users to listen to document titles in a more efficient manner. This mode resembles a search engine’s infinite scroll mode which automatically loads and displays next search results when the user reaches the end of the page. Thus, the infinite-reading mode would continue reading the document titles until a user interrupts the system.

Two thirds (66.37%) of the one-interaction sessions shown in Table 3.3 can be interpreted as *good abandonment*. This is where a user accesses the device with a clear goal, retrieves the information, and then leaves the device [94]. System defined interaction categories, such as *Access source*, where a user inputs “*read me ABC news*”, are also considered good abandonment, as are classifications such as checking the battery level or the weather, and accessing the news headlines. In contrast, 19.05% of one-interaction sessions can be seen as *bad abandonment*, which is where a user leaves without being able to achieve their goal [94]. Bad abandonment classifications often happened when an error occurred such as *no match found*, *ASR errors*, or *part of the command is missing*. The remaining 14.58% corresponded to noise in the logs (errors splitting the sessions or null interactions).

With regard to the Search super category, the average spoken RealSAM query length (4.87 words) is similar to that reported in a recent study of spoken queries from a commercial search engine (4.2 words), but is longer than the length of typed queries (3.2 words) [87]. Guy [87] also reported in this study that one-word queries were rarer in voice (12%). In our dataset, one-word queries were uncommon and only accounted for 1% of the queries. Other researchers have reported that voice queries are on average one word longer than typed mobile queries [225] while Schaller et al. [160] suggested that it may be easier to create long queries with a voice interface than with a keyboard. Although users cannot type queries into RealSAM, and we cannot make a direct comparison between typed or spoken queries in RealSAM, the longer average voice query and the lack of one-word queries may indicate that users find it more natural to create longer queries.

With the third and fourth most frequent queries “*play me the health report*” and “*play me the science report*” the user is presented with a search result list. This list consists of the podcasts containing the corresponding query terms (i.e., “*health report*” or “*science*

report” anywhere within the document). However, RealSAM only reads out the titles for these podcasts, and as these may not contain the query term, the podcasts’ relevance may be unclear to the user. Therefore it may be helpful for the users to hear their query words in the context of the found document. For podcasts, this may mean that users listen to a snippet extracted from the podcast audio in order to understand the context of their query word [171].

Almost half of the interactions were conducted in the original speed and with a female Australian voice, while 48.23% of the interactions were in a different speed or voice. In order to give users more freedom in their interactions with the content, we have evaluated the effect of audio transformations (i.e., prosodic modifications) and our initial results suggest that some of the proposed prosodic modifications lead to better comprehension and identification of the answers in a snippet at the expense of slightly degraded naturalness of the audio signal [51]. Future research could investigate whether skimming or time-compression techniques, such as pause-based skimming, would be useful [14, 15].

3.3.1 Limitations

The quality of the logging process as well as the system’s linguistic and functional limitations hindered the analysis of the interaction logs. For example, RealSAM was updated several times during the data capture process and therefore had different predefined rules in place. Simultaneously, each input from the user was logged through text, but no audio file was present to check whether the ASR had correctly recognised the input from the user. ASR input errors may have resulted in 12.8% of the logs with a Null input from the system; however, we were not able to check this. Furthermore, the RealSAM logs did not indicate if a user had barged-in during the output of the text. For example, it was not possible to establish whether the user listened to all results before making a decision of which results they would like to select. Lastly, we were unable to utilise the timestamps series fully due to the different speeds in voices and the lack of end timestamps.

3.4 Conclusion

The aim of this log analysis was to explore interaction and communication behaviours between users and RealSAM, an audio-only application for accessing media. The strength of this analysis is that we were able to investigate people’s in-context interactions with the application. The log analysis provided insight into users’ behaviours and media accessed, and how users satisfied their information needs. The discussed findings suggest

that a truly conversational system needs further research and development to establish how people want to interact with content over voice without pre-conceived constraints placed on them by the system.

3.5 Chapter Summary

In this chapter, we presented a log analysis of RealSAM, an audio-only application to access books, news articles, and podcasts for people with a visual impairment. We examined how users utilise the application including session descriptives, one-interaction sessions, search sessions, and the personalisation of the TTS output.

The implications of this chapter are for both researchers working and creating similar logs and system developers who are logging audio-only interactions. Recommendations from this chapter include:

1. Log the start and end time for each utterance.
2. Log each interaction of the user and system separately.
3. Log where and when the user interrupted the system output or indicate whether the user listened to the full output.
4. Where possible, retain the audio to check ASR errors or add ASR term confidence values in the output transcription [12].

Our analyses suggest that audio-only interactions systems are still in the early stages of their development, as reflected in the need for improvement in navigational commands, query intent recognition, and skimming techniques over audio. From this chapter, we conclude that audio-only interactions are not straightforward to log and need to be designed carefully.

Chapter 4

Results Presentation for Audio-only Communication

In this chapter, we study search results summary length over an audio-only communication channel.¹ We focus on understanding user preferences for results presentation of an audio-only communication channel with a novel experimental design setup using crowdsourcing. Previous studies in browser-based search results presentation have shown the importance of a document surrogate and its usability [54, 63, 129]. Presenting search results over an audio-only communication channel, however, involves many challenges for users due to the serial nature of speech [15, 224]. Limited studies have been conducted investigating results presentation over an audio-only communication channel. To study search results presentation, we developed a novel experimental design allowing us to obtain insight into users' preferences in the information exploration stage over an audio-only channel. Thus, the aims of this study are twofold (*i*) we want to understand the presentation of results over an audio-only channel, and (*ii*) we want to create a new and re-usable crowdsourcing framework to test search results presentation.

We investigate the impact of search results summary length in audio-only web search and compare our results to a text baseline. The study was designed to collect quantitative data for results presentation preference through CrowdFlower.² A novel aspect of this study is the inclusion of multiple steps in the search task which was designed to reflect multiple turns in the search interaction. To the best of our knowledge, at

¹This chapter has been published as J. R. Trippas, D. Spina, M. Sanderson, and L. Cavedon. Towards understanding the impact of length in web search result summaries over a speech-only communication channel. In *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*, pages 991–994, 2015.

²CrowdFlower has since rebranded to Figure Eight (<https://www.figure-eight.com/>). In this thesis we will keep the reference to CrowdFlower.

the time of the experiment, no previous studies had investigated interactive results presentation for audio-only systems in IR.³ Based on our crowdsourcing experiments, we found that users preferred longer, more informative summaries for text presentation. However, this trend was not observed for audio summaries. Instead, the results indicated that user preferences depended on the query style; for example, shortened audio summaries were preferred for single-facet queries. However, for multi-facet queries (i.e., ambiguous queries), user preferences were not as clear, suggesting that more sophisticated techniques (i.e., conversations) are required to handle such queries.

The broader outcome included the transferability of our crowdsourcing setup to other results presentation studies (for example Chuklin et al. [51] and Spina et al. [171]). We therefore suggest that our experimental setup is robust.

This chapter answers the research questions of the crowdsourcing experiment as well as informing our task choices for our observational study (see Part III). Specifically, we found that different types of queries benefited from a different kind of summary. For example, short audio summaries may be appropriate for single-facet queries with one query intent, while more advanced techniques and conversational approaches may be suitable for multi-facet queries. We included different task complexities in our observational study. Furthermore, by executing this crowdsourcing experiment, we gained insight into future research directions which could be conducted in this framework and these are discussed in Part IV.

This study confirms that translating a text summary into audio may not be sufficient and more sophisticated techniques and conversational procedures may be required to create summaries suitable for an audio-only setting. Moreover, techniques which allow users to interact directly with the document's content rather than with a surrogate may be suitable for an audio-only channel. Additionally, this study supports that further research is required into interactions and search result presentation techniques which alleviate the cognitive load placed on users in an audio-only communication channel.

The chapter is organised as follows. In Section 4.1 we introduce the importance of studying the summary length and state our aims and purposes of our crowdsourcing experiment. Section 4.2, presents the methodology of our experiment, including the crowdsourcing method, experimental design, and participants. We then present the results in Section 4.3, followed by the discussion in Section 4.4, and conclusion in Section 4.5. We end this chapter with the summary in Section 4.6.

³Ethics approval for the experiments was obtained from RMIT University (reference: BSEHAPP 10-14). See Appendix A.1.

4.1 Introduction

Few studies have investigated techniques for effective presentation of web search results via an audio-only communication channel [51, 68]. This chapter seeks to address this. In particular, we examine how to present search results over an audio-only communication channel while not overwhelming the users with information [203], nor leaving users uncertain as to whether what they heard covered the information space [206].

The length of a spoken search result summary plays a crucial role in the success or failure of presenting search results over audio. Successful presentation of search results occurs when the users' information need is satisfied. A short summary might not yield enough information to judge whether the retrieved document is relevant or not; in contrast, a more descriptive summary might take too long to be played and thereby diminish the user experience [19]. Thus a trade-off is necessary between a short summary and a longer, more descriptive summary.

4.1.1 Aims and Purpose

This crowdsourcing study investigated these trade-offs via a crowdsource-based interactive experimental design. The study aimed to develop a baseline of the result summary length users prefer in audio. In particular, this study answers the following research questions:

- What is the impact of search result summary length in a spoken retrieval scenario?
- Do users prefer a longer or shorter summary?

4.2 Methodology: Results Presentation

We conducted a within-subjects crowdsourcing experiment with people from English speaking countries to investigate different summary lengths. Our participants had to indicate the relevance of query summaries, which were of different presentation lengths, and indicate their preferred query summary. We analysed the results with statistical tests to understand the significant differences for these presentation lengths. In this section, we start by describing the crowdsourcing method and experimental design. The remainder of this section explains the participant selection criteria.

4.2.1 Crowdsourcing Method

Our experiments used a crowdsourcing platform to present queries and search results to users. Result summaries of various length were presented in text or audio form. Summary length was either a *full* Google-length summary or a *truncated* version extracted from the original summary. Users were asked to select a result that best addressed the query. The CrowdFlower crowdsourcing tool was employed [100, 101].

4.2.2 Experimental Design

We first describe the task participants undertook, followed by the queries, search engine results summaries, post-task and exit questionnaires, and the use of text as a baseline for audio. The crowdsourcing setup for presenting result lists to participants and collecting judgements is also described.

4.2.2.1 Tasks

Participants were presented with a task which consisted of three queries, corresponding lists of result summaries (i.e., full length summary, truncated summary, and a control question), a post-task questionnaire, and an exit questionnaire. Users were asked to read the three query descriptions and read/listen to the summaries, before stating their preferred result description in the questionnaires.

4.2.2.2 Queries

We designed the tasks to reflect everyday search tasks on the Web. We used query topics from the Text REtrieval Conference (TREC) 2013 Web Track [58] which are based on commercial search engine logs. Since this was a preliminary study, a subset of twenty queries from the TREC 2013 Web Track dataset was used. An assessment of the queries indicated two categories, *single-facet queries* (queries with clear intent) and *multi-facet queries* (typically broader in intent and represented with subtopics). We decided to investigate whether these categories impacted on result summary preference.

The study included seven single-facet and thirteen multi-facet queries. Table 4.1 shows two examples for each type of query, the query itself, and the task description. Only informational subtopics were selected for this study because they have a primary interpretation which is reflected in the description field and often have a large amount of relevant documents in contrast to navigational queries [53].

TABLE 4.1: Examples of queries and query descriptions for single-facet and multi-facet queries.

Type	Query	Description
Single-facet	eggs shelf life	What is the shelf life of a chicken egg—that is, how old can it be and still be safe to eat?
	what was the name of elvis presley's home	What was the name of Elvis Presley's home?
Multi-facet	old town scottsdale	Find restaurants in Old Town Scottsdale, AZ.
	occupational therapist	What is an occupational therapist?

4.2.2.3 Search Engine Results Summaries

Each query was sent to the Google search engine and the text summaries were extracted for the top-five search results, which formed a summary set.⁴ The summaries were converted into a spoken synthetic voice (audio) with the system voice *Alex* from OSX 10.9. The following instructions were presented with each summary set: *These are summaries of the top results. Select the summary that leads to the information you are looking for.* A list-rank number was added to the front of each summary to allow easy identification of the users' selection. Table 4.2 shows a sample summary before and after the conversion for the task.

Truncated versions of the original Google-generated summaries were created manually. Here, a contiguous subset of nine words was selected from each full summary. Nine were found to be a little less than half the length of a standard Google-generated summary. For this initial work, manual summaries were created to avoid bias introduced by poor automatic truncation, which may negatively impact user perception. Human judgement was assumed to be the best way to preserve the meaning of the summary. Thus truncated summaries contained mostly the same information though they were shorter than the original full summary.

The presentation of the twenty search queries was randomised with the use of a Latin square design. Each user saw three queries per task. The order in which the users were presented with the result description (original summary versus truncated summary) was rotated. These steps were implemented to avoid learning effects such as usage order and participants becoming accustomed to a synthetic voice [4, 106].

A problem reported with crowdsourcing is that users try to receive payment without completing the task properly [45]. To overcome this problem, we populated every task with a *Gold Question* to help with data integrity and to detect if the participant was paying attention to the task [39]. The Gold Questions in this study used queries with

⁴Only the top-five were presented to keep the audio task manageable.

TABLE 4.2: Examples of full and truncated summaries for occupational therapist query.

	Full summary	Truncated summary
1	In its simplest terms, occupational therapists and occupational therapy assistants help people across the lifespan participate in the things they want and need to ...	help people across the lifespan participate in the things
2	Occupational therapists treat injured, ill, or disabled patients through the therapeutic use of everyday activities. They help these patients develop, recover, and ...	Occupational therapists treat injured, ill, or disabled patients through
3	Occupational therapy (OT) is the use of assessment and treatment to develop, recover, or maintain the daily living and work skills of people with a physical, ...	Occupational therapy (OT) is the use of assessment and
4	U.S. News’s occupational therapist job overview with comprehensive information on necessary job training, expected salary and job satisfaction, plus tips on job ...	occupational therapist job overview with comprehensive information on necessary
5	Occupational therapy can help improve kids’ cognitive, physical, and motor skills and enhance their self-esteem and sense of accomplishment.	help improve kids’ cognitive, physical, and motor skills and

clear pre-determined answers. Participants were presented with three query descriptions and corresponding summaries. However, one of these summaries was populated with unrelated summary results. A participant that was unable to identify that the summaries were not related to the query had their judgements discarded.

4.2.2.4 Post-task and Exit Questionnaires

Post-task questionnaires are frequently implemented to assess the system–task interaction and gather user feedback on their experiences with using a particular system to complete a specific task [106]. Since no validated questionnaire has been published for studying user reaction to audio-only summaries, we used questionnaires adapted from previous studies, in particular from SDS [66, 216].

Participants completed the post-task questionnaire three times for each of the queries given. The post-task questionnaire, as seen in see Table 4.3, consisted of five questions on a five-point Likert scale (1–5); one question on query judgement with multiple choice answers (6); one question on how the participant listened to the audio with tick boxes (7); and a text box for further comments (8). An example task with a post-task questionnaire is presented in Figure 4.1.

Kelly suggests conducting a questionnaire at the end of the completed task to capture comparisons for within-subjects studies [106]. Thus, participants were also presented

TABLE 4.3: Post-task questionnaire questions.

Post-task Questionnaire
1. The search results I heard are informative.
2. The search results give me a good overview of the available options.
3. The search results give me enough information to select the most relevant result.
4. The search results are presented in a way that is easy to understand.
5. I am confident I can recall the search results that I heard.
6. Which search result would you select to hear further information for?
7. Which statement describes how you listened to the audio?
8. Further comments.

with an exit questionnaire. By using a dynamic panel, the exit questionnaire was available only to participants who were successful in answering a Gold Question. The exit questionnaire was used to measure participants' preferences for information exploration using different result description configurations. The exit questionnaire is presented in Table 4.4 and was analysed based on the post-task questionnaire responses.

TABLE 4.4: Exit questionnaire questions.

Exit Questionnaire
1. Which audio would you recommend to a friend?
2. Which audio did you find easier to get to the most relevant results?
3. Which audio do you think gave you the best result?
4. Which audio do you think was more efficient to use?
5. Why would you recommend the first/second search results to a friend?
6. Why did you find it easier with the first/second search results to get the most relevant result?

4.2.2.5 Using Text as Baseline for Audio


Tasks were paired for analysis, whereby one task's summaries were audio and the others were text. The text output was used to create a baseline measure of the system, facilitating analysis of the difference in preference between audio and text [106], and enabling us to compare audio against the text baseline. The text baseline and audio summary were identical other than particular wording which indicates the output method, for example, *heard* became *read*.

4.2.3 Participants

CrowdFlower allows *contributors* (e.g., researchers who submit tasks to CrowdFlower) to place constraints on the users or crowd workers assigned to a task. The following constraints were put in place for the present study:

A) The search results you will hear answer the query:

What are the health benefits associated with eating dark chocolate?



[Link to MP3](#)

Did you hear a voice speak in the audio?

Yes, I did.
 No, I didn't.

1. The search results I heard are informative.

Strongly disagree	1	2	3	4	5	Strongly agree
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	

2. The search results give me a good overview of the available options.

Strongly disagree	1	2	3	4	5	Strongly agree
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	

3. The search results give me enough information to select the most relevant result.

Strongly disagree	1	2	3	4	5	Strongly agree
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	

4. The search results are presented in a way that is easy to understand.

Strongly disagree	1	2	3	4	5	Strongly agree
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	

5. I am confident I can recall the search results that I heard.

Strongly disagree	1	2	3	4	5	Strongly agree
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	

6. Which search result would you select to hear further information for:

What are the health benefits associated with eating dark chocolate?

Option 1
 Option 2
 Option 3
 Option 4
 Option 5
 None

7. Which statement describes how you listened to the audio? Select all applicable statements. Multiple answers are possible.

I listened to all the search results in the audio.
 I paused the audio and pressed play to hear further options.
 I played the audio more than once.
 I stopped the audio partway through and didn't continue listening.

8. Further comments.

FIGURE 4.1: Example CrowdFlower task.

- Only users with an IP address from Australia, Ireland, New Zealand, the UK, and the USA were allowed to participate to maximise the likelihood that users were native English speakers or had a high level of English proficiency.
- Users were able to participate only once in a particular task to maximise the worker pool.
- Users who took less than sixty seconds to complete the whole task were discarded on the basis that it would take longer than sixty seconds to listen to/read the summaries. This may have mitigated the impact of response bias resulting from the lack of counter phrasing in survey questions 1–5.

Although users were not permitted to participate more than once in a task which had the same set of queries, they were allowed to participate in tasks with different queries. A minimum of 36 participants were recruited for each given task [115]. It was found that 11.8% of all participants did not answer the Gold Questions successfully, and their submissions were discarded. These participants were also not allowed to participate in later tasks.

4.3 Results

In this section, we discuss the data gathered through our crowdsourcing setup. We describe the query judgement distribution based on the post-task questionnaire in Section 4.3.1. We then present the exit questionnaire results in which users compared the two results styles in Sections 4.3.2 and 4.3.3. We show results of the length preference for text and audio summaries (Section 4.3.3).

4.3.1 Query Judgement Distribution

Participants were asked in the post-task questionnaire which summary made them want to know more about the underlying document.⁵ These judgements were analysed with the two-sample Kolmogorov-Smirnov test (*KS test*) to determine whether two given samples follow the same distribution [128]. If no difference is expected, these distributions should be similar. We extracted the “click” distribution from this post-task questionnaire (i.e., the document users wanted to know more about) to compare the “click” concentration with the other results as a proxy for analysing click-through behaviour [122]. We compared the query judgement “click” distributions where the length of the summary was manipulated. The tasks were conducted in pairs: audio and text [106].

The results show that participants made very similar query judgements despite the different presentation styles (audio versus text). That is, 35 out of 40 query judgement distributions followed the same distribution. The KS test showed that for full-length summaries two out of 20 queries had different distributions for audio versus text-based summaries. When investigating these two queries, the KS test revealed that one out of seven single-facet query judgements was statistically significantly different ($p < .05$) when comparing full-length audio to the full-length text baseline. The KS test showed that one out of the 13 multi-facet query judgements was statistically significantly different ($p < .05$) when comparing full-length audio to the full-length text baseline.

⁵This is equivalent to asking which result they would *click* in a traditional SERP.

The KS test showed that for truncated summaries three out of 20 queries had different distributions for audio versus text-based summaries. When investigating these three queries, the KS test revealed that two out of seven single-facet query judgements was statistically significantly different ($p < .05$) when comparing truncated audio to the truncated text baseline. The KS test showed that one out of the 13 multi-facet query judgements was statistically significantly different ($p < .05$) when comparing truncated audio to the truncated text baseline.

4.3.2 Preferred Length of Text Summaries

The exit questionnaire was analysed using the χ^2 goodness-of-fit test to compare the distribution of scores across two levels. The χ^2 goodness-of-fit test was used to assess whether changing the result summary presentation length affected user preference [106].

Table 4.5 indicates that participants tended to prefer full summaries when presented as text. For instance, 57% of participants would recommend complete text summaries to a friend and 57% indicated that full summaries gave better results. The χ^2 goodness-of-fit tests were statistically significant ($p < .01$) for three exit questions concerning the use of the original summary for presenting text results, indicating that this information exploration style was preferred.

TABLE 4.5: Exit questionnaire results for preferences in the search engine result summaries.

Exit Question	Text Summary		Audio Summary	
	Full	Truncated	Full	Truncated
Recommend to a friend	572 [▲] (57%)	434 (43%)	529 (51%)	512 (49%)
Easier to find relevant result	548 [▲] (54%)	458 (46%)	514 (49%)	527 (51%)
Gave better result	576 [▲] (57%)	430 (43%)	539 (52%)	502 (48%)
More efficient to use	529 (53%)	477 (47%)	499 (48%)	542 (52%)

[▲] $p < .01$

4.3.3 Preferred Length of Audio Summaries

In contrast to the summaries presented via text, summaries presented via audio do not indicate a clear preference between full and truncated (preferences differ at most by only 2%). The χ^2 goodness-of-fit tests were not statistically significant ($p > .05$) for any of the exit questions about audio results presentation. No statistically significant difference was found for multi-facet queries. However, for single-facet queries using audio, there was a statistically significant ($p < .05$) preference for truncated summaries.

Participants reported that overall it was easier to recall truncated audio summaries (54.4%) than full audio summaries (49.9%). Moreover, fewer participants stated that they had to listen to the audio more than once (16.8%) for truncated summaries than for full-length audio summaries (23.7%). Only three participants reported that they stopped the audio for truncated summaries, possibly indicating that both the information presented and the length of the information were short enough to avoid cognitive overload.

4.4 Discussion

The aim of this study was to investigate whether summaries of shorter length would be preferred for audio presentation as they could avoid cognitively overloading users [224]. The results showed that summaries which are optimised for a visual space may not translate into audio without consequences [117].

In general, the same “click” distribution was found in the KS test for query judgements between the text baseline and audio. This indicates that participants made very similar query judgements regardless of whether the presentation style was audio or text (both for full-length and truncated summaries). The exit questionnaire responses suggested that for text summaries (single- and multi-facet) full-length summaries were preferred. However, for audio, no length preference was found.

Nonetheless, truncated summaries were preferred in audio for single-facet queries. Thus for more straightforward, less ambiguous queries, shorter audio summaries were both effective and preferred. Furthermore, for multi-facet queries, participants may have benefited from a more informative audio response even at the cost of listening time.

Participants reported that it was easier to recall truncated audio summaries and they were less likely to listen to these audio summaries more than once. This suggests that the information presented and the length of the information did not cognitively overload the user.

The single-facet query judgement distribution for both audio and text followed the distribution reported in past work, where query results ranked first and second received most user attention [99]. However, this expected “click” distribution was not reflected in the multi-facet query judgements; instead, audio summaries ranked first and last obtained the most attention. This is also of interest: the serial nature of audio seems to lead to a bias towards most-recently-heard results, a behaviour not found in visual presentation.

Participants left comments in questionnaires. For summaries of multi-facet queries, they indicated that the summaries were missing critical information. This suggests that the way of presenting summaries may differ depending on query intent: short audio summaries may be appropriate for explicit intent queries (single-facet), whereas broader intent queries (multi-facet) may need more complex techniques (e.g., interactive/conversational approaches).

4.5 Conclusions

The chapter described an investigation into results presentation for audio-only based search. This study aimed to address the following research questions:

- What is the impact of search result summary length in a spoken retrieval scenario?
- Do users prefer a full or truncated summary?

Differences were observed when result summary lengths were presented in the spoken retrieval scenario. In general, there was no preference for fuller descriptive summaries or truncated summaries. However, results revealed that different kinds of queries (single-facet versus multi-facet) benefited from an optimised summary (full versus truncated) depending on the type of query. This suggests that SCS systems will need to identify the users' needs and their context, and adapt the result presentation style accordingly. The SCS system will also need to adapt to the users' query style inside a search session. That is, users may pose single- and multi-facet queries within one session, particularly as they refine their search.

4.6 Chapter Summary

In this chapter, we presented an experiment on user preferences in results presentation over an audio-only communication channel. This study can be viewed as the first step towards informing the design of SCS. We answered the research questions related to the investigation and showed that our experimental design of the study provides new insights into how crowdsourcing can be used interactively.

A limitation of the study is the use of TREC queries to generate the snippets. These queries are highly relevant in a written domain; nevertheless, it is possible that tasks based on spoken information needs would be more suitable. Nonetheless, this initial experiment has provided insight into the different extensions possible to this crowdsourcing

study in results presentation which we present in Part IV of this thesis. Simultaneously, our crowdsourcing methodology has been re-used in further investigations for results presentation by Chuklin et al. [51] and Spina et al. [171].

As a result of this work, different kinds of queries need to be investigated to understand the optimal results presentation which we attempt to address in our observational study in Part III.

In this chapter, we presented a novel experimental setup to investigate user preferences in results presentation over an audio-only communication channel. Our contribution of this part is twofold, we provide *(i)* further evidence that one cannot translate text into audio without consequences for user preference in an IIR setting, and *(ii)* a novel crowdsourcing framework for user preference evaluation in results presentation.

From this study, we conclude that additional research is needed to understand precisely which factors impact the results presentation preference and their usability. We expect that other aspects, such as interaction frequencies, in addition to results presentation and queries, are also different for SCS than in a traditional search engine setting.

In Part III, we overcome the limitations from initial audio-only interaction systems as identified in Chapter 3 and the results presentation in Chapter 4 by studying interactivity through an observational study.

Part III

Towards a New Model of Spoken Conversational Search

Introduction to Part III

In this introduction to Part III we present the approach of our qualitative research. We provide the overall aims and purposes of our observational study including the methods, analysis, and approach.

The aim of this qualitative part in the thesis is to explore SCS as a new search paradigm and seeks to understand the exhibited behaviours demonstrated in an ideal scenario for SCS. Thus, we aim to gain a deeper understanding and overview of the interaction behaviours of a group of participants through qualitative analysis. We first create a rich and detailed dataset through a Natural Dialogue Study (NDS) [221], which we refer to as our *observational study*, to explore SCS interaction behaviours and to seek patterns within this data through an inductive method. In particular, we use our observational study dataset to better understand the communication behaviours at first-hand instead of relying on questionnaires or self-report. This qualitative approach often takes longer to complete than quantitative research because no pre-defined process is present [37]. Nevertheless, the strengths of our qualitative analysis are that it provides an in-depth and detailed analysis to explain complex interactions during the information seeking process.

This observational study, analysis, and results contribute to the broader SCS research in several ways:

- We create the first SCS dataset (*SCSdata*) and we outline the method of capturing and creating this dataset including details of the transcription process in Chapter 5;
- We summarise the observed information seeking conversational and behavioural differences between browser-based and SCS interactions in Chapter 6;
- We identify, classify, and validate the interaction space for SCS which forms the basis of our annotation schema, *SCoSAS*, in Chapter 7; and
- We conclude Part III by analysing the interactivity in the *SCoSAS* utterances on task complexity, interactivity, and discourse in Chapter 8.

Below, we set out the aims and purposes of our qualitative work while providing a methodological overview of NDS, the exploratory observational analysis, thematic analysis, and validation of our investigations. Finally, we conclude by specifying the overall approach and setup used for this qualitative component of the thesis.

Aims and Purpose

Many steps are involved in the analysis of spoken interactions between human participants in order to understand tactics and strategies while performing a task such as collaborative search. These steps include identification and classification of operations designed to interpret search results or to progress the interaction to a successful outcome.

This section presents the overall setup of our observational study with the aim of creating a dataset which is precise with repeatable protocols to provide a high-quality dataset. To our knowledge, no publicly available guidelines are available for SCS data capture and analysis, and we therefore adapted methods from Social Sciences [36] and SDS research [133].

Natural Dialogue Study

Our first step is to explore how people interact or speak in the task they are trying to accomplish [118]. In the case of a SCS system, one could investigate the reference interview techniques or record elicitation processes librarians undertake with information seekers [24, 69]. However, a more direct approach is to record a situation where people are acting as close as possible to the task of interest [118]. A natural setting will encourage participants to converse more intuitively and thus provide insights into the language or vocabulary people use, their turn-taking behaviours, and the information flow [40, 221].

NDS supports an understanding of the accepted conversational patterns in human dialogue. More natural and usable conversational systems can be created by studying human dialogue [221]. Furthermore, NDS provides insight into the grammar usage while conversing in a particular task and gives examples of feedback or prompts. In other words, NDS helps to explore the behavioural patterns and provides insights to improve the design of the system while creating a conceptual understanding of human dialogue behaviour [40].

NDS is not a Wizard of Oz (WOZ) technique. In a WOZ setting, a human acts as a system while the user thinks they are interacting with a live system [76, 83]. Furthermore,

a WOZ experiment can only be conducted if certain pre-conditions are met such as how a system will respond in a particular setting. The WOZ methodology can be used once preconditions, such as knowing how a system should respond, are met since the system is simulated by a human. Thus, WOZ is suitable for hypotheses testing in contrast to the NDS hypotheses forming approach [76].

Exploratory Observational Analysis

A number of disciplines such as communication studies [57], epidemiology [176], and psychology [27] use observational analysis to investigate what people do or say rather than what they say they do. Observational analysis provides a rich understanding which leads to an in-depth explanation of the meaning and the context of phenomena [150]. However, conducting an observational analysis is time consuming and the communication behaviour may be impacted by the lab-based setting [8].

We analyse the NDS by observing and examining the search and interaction behaviours related to the audio-only interaction channel. Our exploratory observational analysis is needed since SCS research is still in its preliminary stage. Moreover, due to limited real-life multi-turn audio-only interaction systems, it is challenging to collect this specific interaction data to infer behaviours. Thus, our exploratory observational analysis of the NDS is a flexible method to provide insight into this new paradigm [150].

Thematic Analysis

Thematic analysis is widely-used for analysing qualitative data [37]. It involves identifying, analysing, and reporting patterns (themes) within the qualitative data [36, 37]. The purpose is to help researchers to organise and interpret data in a meaningful way through an inductive process with possible outputs of themes, categories, or concepts [28]. Furthermore, thematic analysis allows for analysing qualitative data in an accessible and theoretically flexible manner [36]. We adopted the six-step process as outlined by Braun and Clarke [37]:

Step 1: Familiarising self with data;

Step 2: Generating initial codes (codes are concepts or labels which describe important elements of the data and can be seen as the most basic segment [34]);

Step 3: Searching for themes (themes represent a pattern or overarching construct in the data which are typically derived from the codes while increasing the level of abstraction; themes mostly have more explanatory power than codes [75]);

Step 4: Iteratively reviewing each theme with reference to initial codes and the other themes. This ensures themes reflect unique elements of the data;

Step 5: Defining and naming themes; and

Step 6: Producing the report.

I conducted steps 1–3. Steps 4–6 were conducted with the supervisory team.

Thematic analysis in SCS can be used for creating new information seeking models or identifying issues in particular search stages. For example, mapping identified search stages formed by the six-step process can provide insight into a precise information seeking model by instantiating the different steps. Simultaneously, one stage of the information seeking process (e.g., examining results) could be investigated for particular purposes (e.g., sense-making). Thus, thematic analysis allows for systematic investigations of a non-functional system. To our knowledge, this is the first application of thematic analysis in the examination of SCS.

Validating Thematic Analysis

Qualitative research uses different validity criteria than quantitative research [37]. There are several criteria which can be calculated to guard the quality of the analysis. For example, two researchers can code the dataset independently in which case the inter-rater reliability in the form of Cohen’s Kappa can be calculated [119]. Further, the generalisability and transferability of the results can be tested by applying the results to a different dataset and thus strengthening its broader relevance.

Overall Approach and Setup

The development of spoken language datasets is a work-intensive and time-consuming process. Nevertheless, these datasets are invaluable for conversational modelling, as a resource for system development, or defining of vocabulary coverage [78]. The development and evaluation of SDS is a well-studied problem and has shown that iterative analysis and assessment is needed.

To enhance our understanding of SCS, we adopt NDS as a well-established technique used in SDS to develop a spoken language dataset and utilise qualitative analysis to identify meaningful patterns in our dataset [36, 78]. The purpose of our experimental setup is to specify the interaction possibilities in SCS. By outlining these different interactions, we provide the first step towards uncovering the details of the SCS process [78].

Chapter 5

Methods

In this chapter, we present the methodology for our observational study, the data collection setup, the transcription methodology, and the methodology for the data analysis and annotation schema.¹ To our knowledge, we created the first SCS dataset that captures the patterns in complex search tasks, the *SCSdata*, and provide the resources to recreate similar datasets. We generated the *SCSdata* especially to investigate the interaction behaviour between the two actors who are involved in an information seeking dialogue. The work helps us answer the following:

- How are information-dense documents communicated in an audio-only setting?
- What are the components or actions of an information-seeking process via audio?
- What is the impact of task complexity on the interactions and interactivity in SCS?

The *SCSdata* has been used in research published by us [198, 199] and has also recently been used in a study within the broader IR community [205].

Additionally, we describe the qualitative methodology based on thematic analysis for the creation of our SCS annotation schema, *SCoSAS*. Further analysis in this thesis is built on the *SCSdata* which provides a rich understanding of how users communicate over an audio-only communication channel.

¹This chapter consists of the following publications J. R. Trippas, D. Spina, L. Cavedon, and M. Sanderson. Crowdsourcing user preferences and query judgements for speech-only search. In *SIGIR 1st International Workshop on Conversational Approaches to Information Retrieval (CAIR'17)*, 2017. 3 pages and J. R. Trippas, D. Spina, L. Cavedon, and M. Sanderson. A conversational search transcription protocol and analysis. In *SIGIR 1st International Workshop on Conversational Approaches to Information Retrieval (CAIR'17)*, 2017. 5 pages.

This chapter is structured as follows. Section 5.1 provides the general approach of the observational study, including the framework to collecting our data, the development of the annotation schema, the validation of the schema, and the development of design recommendations. Section 5.2 provides key definitions used in this chapter. We then provide the observational study design (Section 5.3). Section 5.4 covers the recruitment and sampling approach for our observational study. Details for the data collection setup, including the task design, experiment procedure, questionnaires and scale, semi-structured interview, and apparatus are presented in Section 5.5. Participants demographics are described in Section 5.6. The transcription methodology together with the principles, protocols, and quality assurance measures are presented in Section 5.7. In Section 5.8, we provide details of the data analysis and annotation schema creation including the methodology to code the transcriptions and validate these codes. This chapter concludes with a summary in Section 5.9.

5.1 Approach

We conducted a laboratory study to collect utterances and search interactions to develop the SCSdata. This dataset captures the utterances of two participants or actors communicating to fulfil an information need. In particular, the purpose of the proposed dataset is to understand how users communicate in an audio-only search setting where no screens are available and focuses on the issues one could encounter when using such a search system. Thus, observing how people search in this setting may provide initial insight into the interactions taken [199].

In this chapter, we describe the qualitative approach of the SCoSAS development which is an annotation schema created for SCS. The schema is analysed and then validated with inter-rater reliability. Further validation is done against an alternative dataset to the SCSdata. Our analysis provides insight into the interaction space, design recommendations, and a first SCS schematic model on which new hypotheses for further research into SCS can be created. These new SCS models can lead to further performance measures and evaluation of different features or interactions which can help with the development of advanced SCS systems.

Iterative processes for design and testing have been used for many decades to develop natural dialogue systems [65]. For example, Gorin et al. [82] first collected and analysed human-to-human dialogues for a call-routing task in order to design a system based on this data which they then analysed via a WOZ setting [35]. Our approach is similar, we design the NDS which we refer to as the observational study and conduct this study with participants. We process the collected data, and we create the SCoSAS, which is

an annotation schema for this data. We analyse the annotation schema, validate it, and derive descriptive statistics from the schema. Then, we extract design recommendations and provide a schematic overview of the SCoSAS based on the previous steps as seen in Figure 5.1.

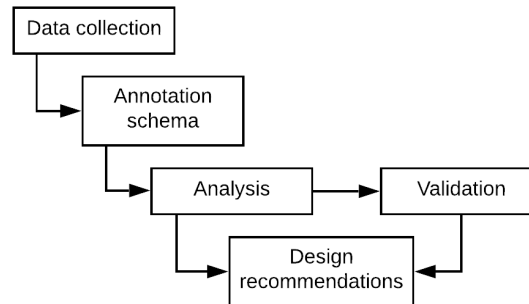


FIGURE 5.1: Schematic overview of methodology.

5.2 Definitions

We now provide definitions used in the observational study.

Turn: Where a participant talks for an amount of time without being disrupted. If a speaker interrupts another speaker, then the first speaker's turn is finished. The second speaker now takes the initiative.

Move: A move is an atomic interaction in a dialogue with a communication goal. Thus, a dialogue consists of an array of moves. In a traditional visual-textual interface, a mouse click or key press are single moves in the dialogue. Every action(s) a system needs to take is linked to an atomic move from the participant. One turn can consist of multiple moves.

Seeker: The Seeker is an observational study participant or actor who receives an information need (*backstory*) but does not have access to the search engine to fulfil that information need. The Seeker has to communicate with the Intermediary to receive information from the search engine which is verbalised by the Intermediary.

Intermediary: An Intermediary is an observational study participant or actor who does not have access to the information need. However, the Intermediary has access to a search engine and has full control over it. The Intermediary has to co-operate with the Seeker to resolve the Seeker's information need.

Backstory: A short information need statement which motivates and contextualises a search need.

5.3 Study Design

Our observational study consisted of one session with two participants, where one participant acted as the *Seeker* and the other participant as the *Intermediary* as illustrated in Figure 5.2.

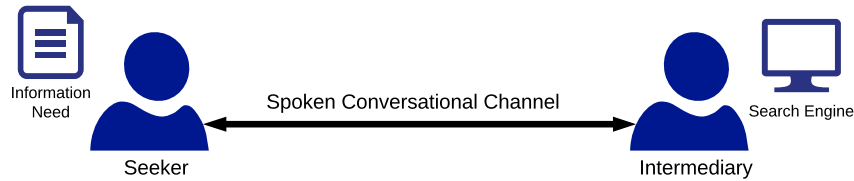


FIGURE 5.2: Experimental setup.

The Seeker received a short information need as *backstory*. The Seeker had to read the backstory and verbalise the information need without reading out the backstory to the Intermediary. Instead, the Seeker had to personally formulate their information need problem to convey it to the Intermediary. The Intermediary had access to a search engine through a desktop computer. In other words, the Seeker acted as the searcher and the Intermediary simulated the audio-only interface. Participants could not access each others' tasks or search engine, were not able to see each others' facial expressions, and could only verbally communicate. All backstories were randomised and the participant roles were randomly assigned.

The participants had to collaborate to satisfy the information need. Both participants completed a pre-test questionnaire at the beginning of the study. The participants answered a short questionnaire (pre- and post-task questionnaire) before and after each scenario, at the end of the study (exit questionnaire), and a semi-structured interview was conducted to conclude the study. The overview is presented in Figure 5.3 where the blue line represents the Seeker, and the yellow line represents the Intermediary. Participants could leave at any time, and there were no adverse consequences from participating. All interactions were audio and video recorded and transcribed for analysis.²

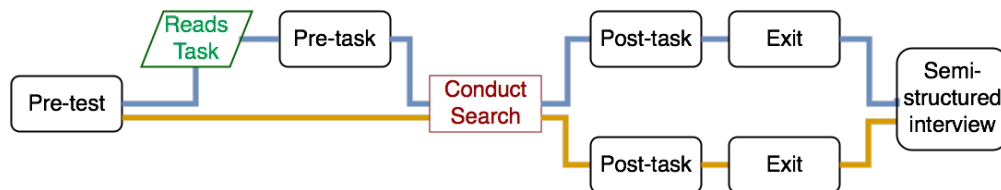


FIGURE 5.3: Visual overview of experiment procedure.

²Transcripts are publicly released and can be accessed on http://bit.ly/SCSdata_thesis. Labelling and a codebook information are on the webpage. More information can be found in Appendix C–D.

The participants did not receive an example of how a search task could be solved to avoid biasing the results.

5.4 Recruitment and Sampling

The study consisted of 15 observational study sessions of participant pairs (i.e., 30 participants in total) completing pre-defined information seeking tasks.³ These studies took place in a computer lab at RMIT University in Melbourne, in June 2016. Two participant pairs were used as pilots and are not included in the analysis. We distributed a call for participation through the RMIT University Behavioural Business Lab mailing list.⁴ Participants with a high self-reported level of English were contacted for participation. Convenience sampling of participants was used for this study.

5.5 Data Collection Setup

5.5.1 Task Design

Search tasks are an important element in IIR studies. It has been shown that different information seeking tasks can have different search behaviours or characteristics [44, 97, 125, 218]. For example, Arguello et al. indicated that aggregated search in more complex tasks received greater user interaction such as longer queries, a greater number of queries, more SERP clicks, and more visited pages [11].

To examine the behavioural differences among our observational study participants, we used backstories created by Bailey et al. [20]. These are based on three levels of cognitive complexity suggested by Wu et al. [218] adopted from the Taxonomy of Learning updated and redefined from Bloom's taxonomy of educational objectives [9]. This taxonomy provides six dimensions to reflect the cognitive process and knowledge. These cognitive process dimensions include: remember, understand, apply, analyse, evaluate, and create, and their definitions are presented in Table 5.1. Each level increases the degree of cognitive effort required. Although the taxonomy was established for educational purposes, it has also been used extensively to present cognitive complexity in search tasks [11, 20, 46, 108, 218].

We describe an evaluation of nine search tasks based on the cognitive complexity framework of the Taxonomy of Learning [9]. The following three cognitive dimensions were

³The setup was reviewed and approved by RMIT University's Ethics Board (ASEHAPP 08-16). See Appendix A.2.

⁴<https://orsee.bf.rmit.edu.au>

TABLE 5.1: Anderson and Krathwohl’s Taxonomy of Learning objectives (cognitive process dimension) [9].

Dimension	Definition
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 differentiating, organising, and attributing.
Evaluate	Making judgements based on criteria and standards through checking and critiquing.
Create	Putting elements together to form a coherent or functional whole; re-organising elements into a new pattern or structure through generating, planning, or producing.

used: *Remember*, *Understand*, and *Analyse*. We chose these different cognitive complexities to observe different techniques and search behaviours used in an audio-only search setting.

Table 5.2 presents the nine queries and backstories with relation to their cognitive dimension used in this study [134].

5.5.2 Procedure

Each session took 90 minutes per participating pair. Participants were rewarded with 20 AUD for their time. The procedure of the user study was as follows:

1. Welcome the participants to the lab and give them a brief introduction of what will happen in this session. Ask the participants to read the information statement. (The Participant Information Statement was a file on the webpage where the participants signed up for the study. They were also sent the Participant Information Statement by email with the confirmation of the time and date of their appointment.)
2. Ask the participants to sign the consent form.⁵

⁵See Appendix A.3 for Participant Information Statement and consent form.

TABLE 5.2: Example search tasks taken from Bailey et al. [20].

Dimension	Query and Example Backstory
Remember	What river runs through Rome, Italy? Many great cities have rivers running through them, as rivers facilitated trade and commerce as well as supplying fresh water to drink. You remember that Paris has the Seine, London has the Thames, but what does Rome have?
	What language do they speak in New Caledonia? You and your partner are thinking of places to go on holiday. New Caledonia is an option, but you realize you don't know what language is spoken there and you decide to find out.
	Where does cinnamon come from? The other day you were eating some spiced biscuits from Europe, when it occurred to you that cinnamon probably isn't native to that part of the world. You would like to know where it comes from.
Understand	Recycle, automobile tires You need to buy new tires for your car, and the local dealer has offered to take the old ones for recycling. You didn't know tires could be recycled and you wonder what new uses they are being put to.
	Outsource job India A recent report on the radio quoted a politician as saying that one of the causes of rising unemployment in the U.S. was the outsourcing of jobs to India. This has made you interested in finding out what jobs that used to be in the U.S. have been outsourced to India.
	Marine vegetation You recently heard a commercial about the health benefits of eating algae, seaweed and kelp. This made you interested in finding out about the positive uses of marine vegetation, both as a source of food, and as a potentially useful drug.
Analyse	Turkey Iraq water Looking at a map, you realize that there are several rivers that commence in Turkey and then flow over the border into Iraq. You wonder if Turkish river control projects, including dams and irrigation schemes, have affected Iraqi water resources.
	Airport security Every time you go through the security screening at an airport, you wonder whether it is making any difference. Find out how effective the many new measures (beyond just standard screening) at airports actually are, both for scrutinizing of passengers and their checked and carry-on baggage.
	Per capita alcohol consumption You recently attended a big party and woke up with a hangover, and have decided to learn more about the average consumption of alcohol. You are particularly interested in any information that reports per capita consumption, and want to compare across groups, for example at the country, state, or province level.

3. Provide details about the different roles as Seeker and Intermediary including the protocol to stop and start the search between the tasks.
4. Ask one of the participants to act as the Intermediary and ask them to move to the computer which is set up for the study.
5. Ask both participants to complete the Pre-test Questionnaire (See section 5.5.3.1).
6. Ask the Seeker to read the task for themselves (See section 5.5.1) and request the Seeker to complete the Pre-task Questionnaire (See section 5.5.3.2). Then prompt

the Seeker to start the search by saying “Start search” when ready. Reiterate to the Seeker that they can stop the search when they believe enough information has been collected to satisfy the information need by saying “Stop search”. A time limit of 10 minutes was also put in place, and the researcher stopped the search when this limit was reached to continue with the next task.

7. Ask the participants to complete separate Post-task Questionnaires (See Section 5.5.3.3).
8. Ask the participants to repeat steps 6–8 with two different search tasks.
9. Ask the participants to complete separate Exit Questionnaires (See Section 5.5.3.4) to capture overall feedback on their experience of the search and study.
10. Ask the participants to join the researcher for a semi-structured interview.

A total of 15 pairs (13 pairs excluding the pilot participants) completed three tasks each.

5.5.3 Questionnaires

After each task, both participants completed questionnaires adapted to their role. These differences are reflected in the example questionnaires in the following sections and are presented as a high-level overview in Figure 5.4. In general, the questionnaires’ objective was to gather information about the participants’ familiarity to online searching and the participants’ search success during the study.

All items are evaluated with a five-point scale, where 1=Not at all, 2=Slightly, 3=Moderately, 4=Very, and 5=Extremely, unless otherwise stated. Questions or statements in the questionnaire annotated with an asterisk (*) indicate that they are adapted from Kelly et al. [108]. All questionnaires and semi-structured interview questions are in Appendix B.

The questionnaires were administered on paper forms. The paper forms were because the Seeker had no access to a computer.

5.5.3.1 Pre-test Questionnaire

Both participants completed a pre-test questionnaire before they started the full experiment. This pre-test questionnaire gathered demographic data such as age, gender, the highest level of education, employment, and computer and search engine usage.

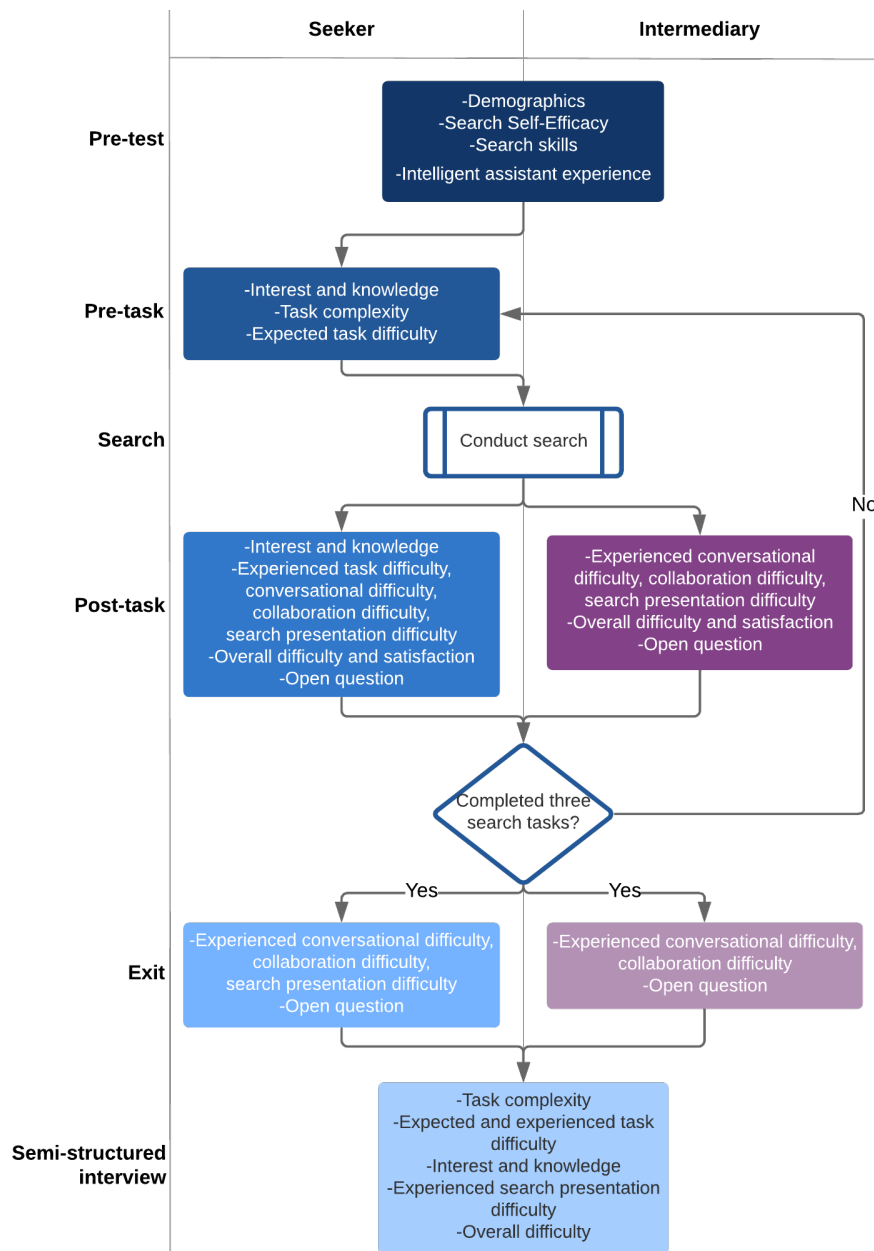


FIGURE 5.4: Overview of questionnaires and their measures.

Participants completed the Search Self-Efficacy scale [38] and rate their overall search skills.

Participants were asked if they had experience with intelligent personal assistants such as Google Now, Siri, Amazon Echo, or Cortana.

5.5.3.2 Pre-task Questionnaire

The Seeker completed a pre-task questionnaire after reading the backstory and before initiating each search task. This questionnaire measured interest and knowledge in the

backstory, task complexity, and expected task difficulty. The questions for the Seeker are given in Table B.1.

5.5.3.3 Post-task Questionnaire

Both the Seeker and Intermediary completed a different post-task questionnaire tailored to their role at the end of each search task. The post-task questionnaire assessed the system–task interaction.

The questions for the Seeker are provided in Table B.2. The questions for the Intermediary are provided in Table B.3.

5.5.3.4 Exit Questionnaire

Participants completed the exit questionnaire after they finished all the search tasks and before the semi-structured interview was conducted. Both participants received different questionnaires which were tailored to their role.

The questions for the Seeker are provided in Table B.4. The questions for the Intermediary are provided in Table B.5.

5.5.4 Semi-structured Interview

At the completion of all search tasks and questionnaires, the participants were invited to participate in a semi-structured interview. The interview investigated the following topics from the questionnaires in more details: task complexity, expected and experienced task difficulty, interest and knowledge, experienced conversational difficulty, experienced collaboration difficulty, experienced search presentation difficulty, and overall difficulty (for the semi-structured interview questions, see Appendix B.6).

5.5.5 Apparatus

Intermediaries completed the search tasks on a 21.5inch screen iMac with 8GB ram. Intermediaries received a mouse and keyboard to interact with the search results. Chrome version 51 was used as a browser, together with Silverback⁶ version 2.7 to record the screen, audio, and face of the participants.

⁶<https://silverbackapp.com/>

Google was used as the default search engine although participants were allowed to switch to different search engines. The browser history and cache were reset after each data collection session.

5.6 Participants

The observational study involved 26 participants (13 participant pairs) recruited through a mailing list⁷. Fifteen participants were female and 11 were male with a mean age of 30 years ($SD=11$, median=26, range 18–54).

Twenty-two participants reported being a native English speaker, and four participants said they had a high level of English proficiency. The highest level of degree held was a Master's degree. Eighteen participants reported that they were awarded a Bachelor's degree or higher and eight participants said their highest level of degree awarded was High School graduation. The majority of participants were students (73%), 19% was employed, and 7% were unemployed. The most common fields of education were Science and Engineering (both 19% respectively) and Law (11%). Participants reported that they had been using a computer for more than ten years (85%) and 15% reported using a computer for 5–10 years. All participants said that they used search engines daily with the majority of participants reporting that they used a search engine more than eight times per day (54%) as seen in Figure 5.5.

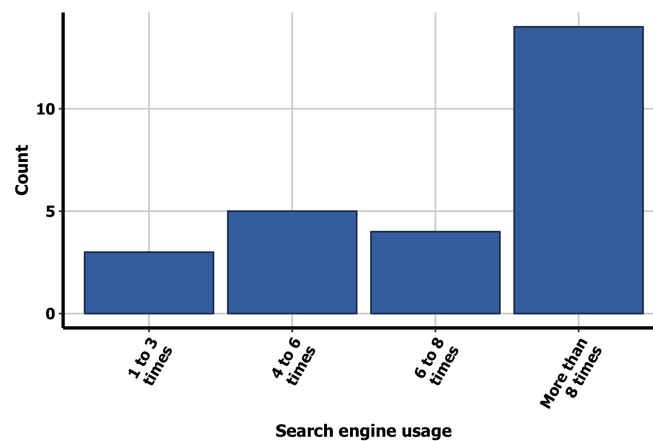


FIGURE 5.5: Participants' search engine usage per day ($N = 26$).

Participants rated their search skills on a 5-point scale, where 1=novice and 5=expert. Participants' mean search skills were 3.9 ($SD=0.5$), with a minimum score of 3 and a maximum of 5.

⁷The mailing list is created and maintained by the Behavioural Business Lab at RMIT University, <https://orsee.bf.rmit.edu.au/>

Participants' search self-efficacy was measured with the Search Self-Efficacy scale [38], which contains 14 items describing different search activities. Participants indicated their confidence in completing each activity using a 10-point scale, where 1=totally unconfident and 10=totally confident. Participants' average Search Self-Efficacy was 7.3 ($SD=1.51$ and Cronbach's $\alpha=0.93$).

Participants reported their usage of intelligent personal assistants, such as Google Now, Apple's Siri, Amazon Alexa or Microsoft Cortana. Four participants reported never having used an intelligent assistant and eight had used one a couple of times but no longer did so as seen in Figure 5.6. The majority (54%) of the participants said they used an assistant, consisting of five participants using one at least once a month and nine participants using one at least weekly.

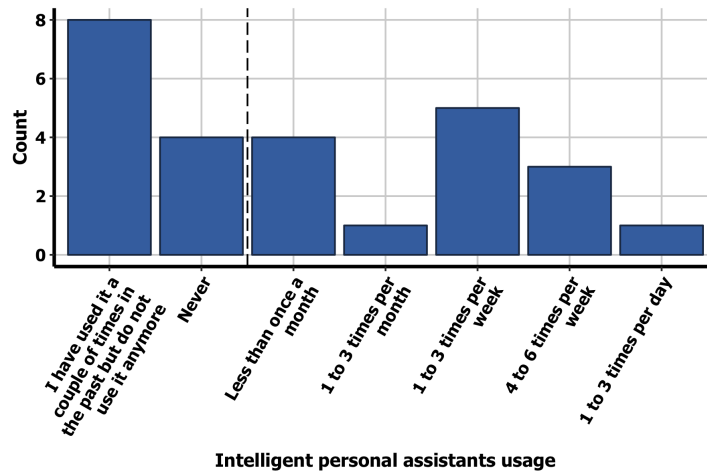


FIGURE 5.6: Frequency usage of intelligent personal assistants ($N = 26$).

5.7 Transcription Methodology

We captured a total of 6.5 hours of information seeking conversations in our laboratory study. To create a dataset which is reusable by other researchers we transcribed the audio recordings. However, limited information for transcribing information seeking conversations is available. Thus, we created a protocol to transcribe SCS which included the quality assurance processes and information on the importance of choosing transcription tools.

We first present general transcription principles followed by more detailed examples of how these principles are translated for our transcription protocol. We then suggest tools for transcription and introduce quality assurance processes.

Other fields have established guidelines for both transcriptions and analysis (e.g., Social Sciences [37] or ASR [184]); however, to our knowledge, there are no publicly available guidelines for SCS. Given the importance of consistent research techniques to establish a body of comparable work, we propose a protocol for information seeking conversations which includes data preparation, quality assurance, and analysis to assist future researchers in the field.

5.7.1 Transcription Principles

We followed the principles presented below allowing for high-quality transcriptions. In the transcription process, we wrote what was said; thus we did not include non-linguistic observations such as facial expressions, body language, or intonations. Our transcription is therefore verbatim and often referred to as orthographic transcription.

For our transcriptions, we aimed to capture how people expressed themselves in a search situation and therefore we transcribed all recorded utterances with the following transcription principles from McLellan, MacQueen, and Neidig [131]:

1. *Preserve the morphological naturalness of transcription.* Keep word forms, the form of commentaries, and the use of punctuation as close as possible to speech presentation and consistent with what is typically acceptable in written text.
2. *Preserve the naturalness of the transcript structure.* Keep text structured by speech markers (i.e., like printed versions of plays or movie scripts).
3. *The transcript should be an exact reproduction.* Generate a verbatim account. Do not prematurely reduce text.
4. *The transcription rules should be universal.* Make transcripts suitable for both human/researcher and computer use.
5. *The transcription rules should be complete.* Transcribers should require only these rules to prepare transcripts. Everyday language competence rather than specific knowledge (e.g., linguistic theories) should be required.
6. *The transcription rules should be independent.* Transcription standards should be independent of transcribers as well as understandable and applicable by researchers or third parties.
7. *The transcription rules should be intellectually elegant.* Keep rules limited in number, simple, and easy to learn.

We used ELAN (EUDICO [European Distributed Corpora Project] Linguistic Annotator)⁸ [121] because ELAN accommodates both the use of the above principles and our precise transcription protocol which is explained in the next section. These principles and rules allowed us to create high-quality transcripts with an iterative manner which was systematic and consistent.

5.7.2 Transcription Protocol

Transcription protocols have two main goals: minimising the probability that the transcripts produced are inconsistent and reducing the likelihood that the data analysis will be weakened or delayed [131]. We developed the following transcription protocol adapted from Braun and Clarke [37] and McLellan et al. [131] based on the transcription principles described previously.

- Turns were identified and every first word of each new turn was capitalised.
- Audio recordings were transcribed verbatim (i.e., recorded word for word, exactly as said), and non-complete words or sentences were transcribed to the best of the transcriber’s ability. Nonverbal or background sounds were not included (e.g., laughter, sighs, or coughs).
- If participants mispronounced words, these words were transcribed as the individual said them. The transcript was not “cleaned up” by removing slang, grammatical errors, or misuse of words.
- While “aha”, “hmm” or “uhm” were included, linguistic- or phonetic-type transcripts were not produced.
- Abbreviations were written as said, such as “TV” for “television”.
- Numbers were all spelled out (e.g., “90” is written as “ninety”).
- Spelled out words were capitalised (e.g., participant spells the country “New Caledonia” which is transcribed as “NEW CALEDONIA”).
- URLs were written as pronounced (e.g., “drive dot com dot AU”).
- Place names and brand names were written with an initial capital.
- Portions of the audiotape that were inaudible or difficult to decipher was transcribed as *[inaudible segment]*.

⁸<http://tla.mpi.nl/tools/tla-tools/elan/>

- Pauses in the speech were indicated with an ellipsis. A brief pause was defined as a two- to five- second break in speech. Pauses longer than five seconds were transcribed as *[long pause]*.
- A style guide with vocabulary was kept throughout the project.

We did not focus on overlapping speech since this was not in the scope of our analysis [131].

5.7.3 Transcription Quality Assurances

We developed a rigorous process for reading and reviewing the text. In particular, we checked the audio a minimum of three times against the transcript before the transcript was submitted.⁹ This technique is also referred to as the three-pass-per-tape policy [131]. All transcripts were audited for accuracy by a professional editor.

5.8 Data Analysis and Annotation Schema Creation

In this section, we provide all the necessary steps to create an annotation schema for SCS by using thematic analysis. The purpose of our annotation schema, the SCoSAS, is to understand all components and interaction paths of this new paradigm while providing researchers with a labelled dataset for further research such as machine learning [221]. We not only describe the steps required to code transcriptions, but also provide information on the analysis of the codes or annotation schema, and discuss the validation of the schema.

5.8.1 Coding Transcriptions With Thematic Analysis to Develop SCoSAS

We coded (i.e., labelled) our transcriptions using thematic analysis as described previously in Part III Introduction. The labels of the SCSdata form the annotation schema, SCoSAS. The Seekers and Intermediaries did not have access to each others' search task or search engine interface, could not see each other, and could only communicate verbally. This setup can be seen in Figure 5.7 ((a) Seeker, (b) Intermediary).

The Seeker, Intermediary, and the Intermediary's screen were filmed during the session. The recordings were synchronised and merged for transcription. Recordings were transcribed and coded in order of their historical occurrence. The codes were created from the video and transcriptions in ELAN. We adopted the following steps:

⁹I created the transcriptions.

The screenshot displays the ELAN 4.9.4 interface. At the top, there's a menu bar (File, Edit, Annotation, Tier, Search, View, Options, Window, Help) and a toolbar. The main window is divided into several panes:

- Top Left:** A browser window showing search results for "effectiveness of new security measures at airports".
- Top Right:** A video player showing a person at a desk. Two red circles labeled (a) and (b) are overlaid on the video.
- Right Side:** A table with columns: Nr, Annotation, Begin Time, End Time, Duration. It lists three annotations:

Nr	Annotation	Begin Time	End Time	Duration
1	Where does cinnamon come...	00:00:02.710	00:00:11.550	00:00:08.840
2	Airport Security	00:06:34.290	00:06:47.590	00:00:13.300
3	Outsource job India	00:17:24.040	00:17:36.080	00:00:12.040
- Bottom:** A timeline view showing the transcription of the video. It includes a waveform at the top and a list of participants and their utterances below. Annotations (1), (2), and (3) are marked on the timeline.

FIGURE 5.7: Sample screenshot of ELAN transcription and analysis tool (anonymised). Annotations indicate (a) Seeker, (b) Intermediary, (1) Controlled vocabulary Seeker, (2) Transcription, (3) Query.

Step 1: Identifying when each participant spoke, i.e., identifying turns.

Step 2: Transcribing the turn.

Step 3: Designing and assigning codes to each turn with ELAN. Observational notes were added. The full dataset was coded with each utterance receiving equal attention. We classified concepts from the recordings and devised a coding scheme according to the similarities across different actors. The codes were designed to identify the single action occurring in that turn, describing features of the data and defining the function of the turn. Thus, turns were annotated with the actions taking place. Consequently, meaningful labels were developed from the original annotations. *Controlled Vocabulary* was added to a *dictionary* which was created during coding. This dictionary was then developed into a full *codebook*.

Step 4: Combining codes into themes for further analysis.

Step 5: Checking quality assurance. Transcriptions and codes were exported from ELAN to a text file. Spelling and codes were checked.

Step 6: Importing files into R and aggregating codes to check whether codes within a theme conceptually belonged to that theme.

Note: Steps 3–6 were conducted iteratively. This process reduced the initial 100 codes to 84 through the identification of overlapping codes. To preserve the nuanced action described in the codes for future information seeking research, distinctions between closely defined codes were retained. For example, the codes “Information request” or “Information request within document” were retained to identify in which section of the interaction particular information was requested.

5.8.2 Analysis of Coding

The ELAN transcriptions with codes and observational notes were transferred to R for further analysis. The transcriptions were modified to lower case, and punctuation and extra spacing were removed. We deliberately did not eliminate any errors, false starts, or confirmations since these occur in real case voice search scenarios.

In the context of mixed-initiative information retrieval dialogues, the terms *control* and *initiative* are used interchangeably. However, we used the approach of *taking the initiative equals taking the turn*, as described by Hagen [88]. Adopting this approach means that one turn can consist of multiple moves or communication goals [187]. We coded the complete dataset identifying aspects of relevance to our research aim. Thus, codes were applied to each turn taken by either the Seeker or Intermediary and these codes were collated and given *themes*. Themes may consist of *sub-themes* which capture specific concepts of that theme as illustrated in Figure 5.8.

5.8.3 Validation of Annotation Schema SCoSAS

To reduce the possibility of missing important data points, we validate our coding schema in two ways. We computed (1) inter-rater reliability and code overlap, and (2) overlap and coverage based on the coding of a different dataset, The Microsoft Information-Seeking Conversation data (MISC)¹⁰, with our predefined codes [185].

A second independent annotator, who is familiar with information seeking and information retrieval research, recoded all utterances in the SCSdata to obtain the inter-rater reliability with Cohen’s Kappa and code overlap [119]. The second annotator used the codebook for closed coding (i.e., the categories were already determined).

¹⁰The MISC data was accessed at <http://aka.ms/MISCv1>.

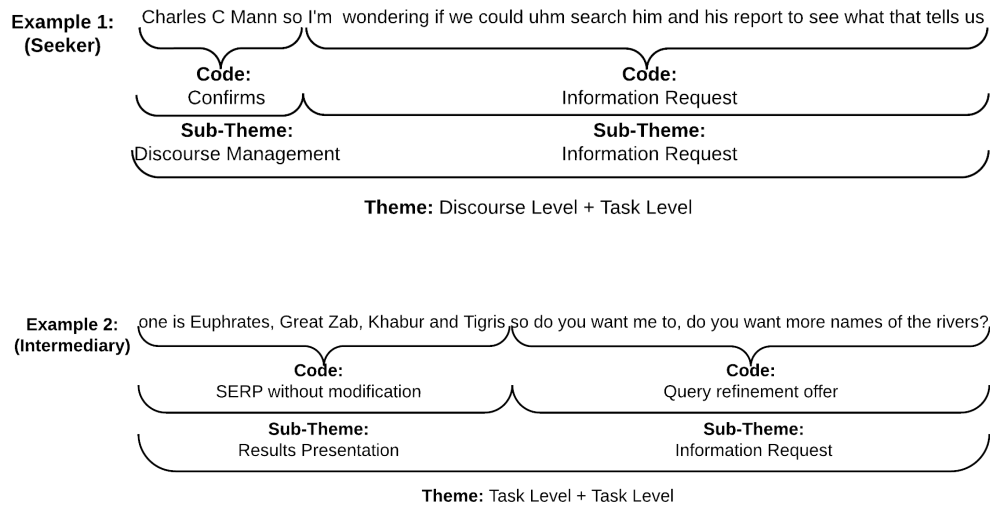


FIGURE 5.8: Example of coding utterances for Seeker and Intermediary.

Identifying useful actions for SCS which have not been covered in the SCoSAS provides an understanding of the scope of our coding schema. Therefore we apply the SCoSAS to a second and similar dataset, the MISC to calculate the overlap and coverage [185]. We took a random sample from the MISC and coded the utterances according to our dataset. Nevertheless, it may not be possible to achieve complete coverage with our annotations given the complexity and unexplored interactivity of a SCS information seeking dialogue [174]. In addition, achieving full coverage is difficult and often not possible to achieve [175]. Hence, declarations which were not covered in SCSdata received new codes according to the steps mentioned earlier in Part III Introduction.

5.9 Chapter Summary

This chapter aimed to outline the methodology for creating the dataset SCSdata. Furthermore, the chapter discussed the analysis of the dataset to create the first annotation schema for SCS, the SCoSAS, including validating the dataset.

The detailed methodological setup contributes to (i) the experimental setup to re-create and develop more SCS datasets, (ii) a novel annotation schema creation methodology through the coding of transcriptions for SCS, and (iii) a method to evaluate the annotated SCSdata.

Chapter 6

Observing Spoken Conversational Search Interaction Behaviour

In the previous chapter we described the creation of the SCSdata. In this chapter, we use the data and we present the results from the observational study by inspecting the conversational interactions.¹ By examining the conversations from our study, we identify commonly used interactions which apply to SCS. Thus, we observe the characteristics of spoken exchanges in an information seeking environment. We describe and analyse these interactions and provide examples where possible.² The results are further discussed in Part IV.

The key finding of this chapter is that interactions can be divided into search communications (e.g., how people express their information needs or how found results are communicated) or non-search communications (e.g., utterances to repair the conversation). We illustrate with our observations and examples that complexity and interactivity are intrinsic components of SCS. We highlight that sophisticated systems will be needed to overcome the difficulties these inherent components pose. Furthermore, our results suggest that we may need to review the existing information seeking models to include this increased complexity and interactivity of SCS. This chapter provides a basis for illustrating that necessary re-examination.

The observational results are divided into two sections outlining high-level observation interactions. First, Section 6.1 describes observations which are related to search interactions. We frame these observations in the different stages of an information seeking process as specified by Sahib et al. [157], enabling us to introduce our observations

¹This chapter consists of the following publication J. R. Trippas, D. Spina, L. Cavedon, H. Joho, and M. Sanderson. Informing the design of spoken conversational search. In *Proceedings of Conference on Information Interaction and Retrieval (CHIIR)*, pages 32–41, 2018.

²Examples have been edited for readability and are marked accordingly.

in a structured manner. Second, Section 6.2 presents non-search interactions, which are not explicitly constrained to search but cover the integral features of SCS, such as communication and cognitive user models.

6.1 Search Interactions

We present observations at four stages of the information seeking process: *Query Formulation*, *Search Result Exploration*, *Query Reformulation*, and *Search Results Management* as defined by Sahib et al. [157]. These stages are analogous to Marchionini’s *Express*, *Examine*, *Reformulate*, and *Extract Information* stages in Information Seeking Process (ISP) [126]. Furthermore, Sahib et al.’s stages of information seeking provide broad phases for the collected observations while still offering a structure which is embedded in search interactions [157].

For each of the four information seeking process stages, we describe the observations, present an analysis of the observations, and provide examples. In the query formulation stage, we discuss the naturalness of information requests. We then continue with the search results exploration stage where we illustrate the difficulty Seekers had to distinguish whether information came from a SERP or a document. We also outline how relevance feedback can be captured in this audio-only environment, and how Intermediaries had to assist the Seekers with more visual information, such as the awareness of novel or previously seen information including the interpretation of graphical content. We progress to strategies addressing the query reformulation stage and examine how Seekers presented repetitive search tasks and provided information requests within documents. Finally, we investigate the search results management stage and how the extracted information was stored.

6.1.1 Query Formulation

For the observational study, we provided the Seekers with a backstory for each information seeking task, allowing them to verbalise their own “information request”³. In this section, we provide examples of these information requests and how they are formed when they are articulated instead of typed.

³We use the notion of information request because these expressions were often not precise queries but more an explanation of what the Seekers were looking for.

Naturalness of Information Request

Seekers varied the way they verbalised their information request: from uttering a query-like expression to describing a detailed and carefully crafted information request. The examples in Table 6.1 illustrate the wide range of information requests posed.

TABLE 6.1: Example information request utterances.

	Example utterance	Characteristic
1	<i>“Turkish river control”</i> (P7)	Query-like
2	<i>“So uhm what jobs that used to be in the US are no longer have been outsourced to India?”</i> (P13)	Natural language type query
3	<i>“So I’m trying to find the count part in uhm a biscuits that you are get from Europe uhm it contains cinnamon and I want to know where the cinnamon is coming from are there is this uhm is this coming from Europe uhm so how to uhm search for uhm cinnamon Europe biscuits”</i> (P23)	Query babbling [138]
4	<i>“Maybe start off with uhm type in the origins of cinnamon”</i> (P5)	Instructions plus query-like
5	<i>“Can you please search car tyre recycling [long pause] and then in the results I am looking for examples of what uhm recycled car tyres are used for”</i> (P15)	Step-wise information request revealment (Instructions plus query-like and additional information on what to look for in the results)
	<i>“Have Turkish river control projects affected Iraqi water resources [long pause] so we’re looking for if dams or irrigation schemes have affected uhm any of the Iraqi people”</i> (P17)	Step-wise information request revealment (Natural language type query plus additional information on what to look for in the results)
6	<i>“Uses for old car then the query or, passenger vehicle tyres TYRES (Seeker spells tyres) or in caps tires TIRES (Seeker spells tires) ... and I wanna uhm do a date range so the data is from the most recent twelve months, so uses for old car caps or passenger vehicle or tyres TYRES (Seeker spells tyres) caps or tires TIRES (Seeker spells tires) and data in the last twelve months that’s the query”</i> (P3)	Detailed and carefully crafted information request (teleporting [180]) plus utilising extra features such as date range from the system

Many different expressions of information requests were observed that did not conform to the conventional length of browser-based search query (of 3.2 words) [87]. Instead, these information requests included natural language requests, instructions, or additional information to the original information request. Other observations include Seekers wanting to spell keywords in their queries or use advanced search mechanisms such as Boolean syntax. Note that in an audio-only setting, allowing spelling may be a primary feature, given that typing or copying/pasting keywords is not readily available.

It could be argued that some of these information requests are observed because Seekers are not restricted to a typical web-based search box and do not have to translate their thoughts into queries. More precisely, by not conforming to the average query format used in a browser-based search setting, we now see an increase in the range of ways in which information requests can be expressed. Besides this increase, we also notice

that the scope of this range is so diverse, from query-like to query-babbling information requests, that we experience an increase in information request complexities.

6.1.2 Search Results Exploration

In this section, we investigate the interactions between the Seeker and the Intermediary after the initial information request (i.e., any interactions after the first turn).

First, we investigate the concept of the boundaries between the SERP and the documents. We then cover how both Seeker and Intermediary are actively involved in relevance judgements, and outline what happens when previously encountered results are seen. Finally, we investigate how graphical information can be useful in an audio-only setting.

SERP and Document Boundaries

In traditional IR, the SERP and the documents linked to the SERP can be thought of as different entities. Nevertheless, in our study, where a human simulated the SCS system, these differences were not present for several Seekers during their search. There were instances where a Seeker asked an Intermediary to access a particular document from a SERP assuming the Intermediary was located on the SERP. However, the Intermediary was already positioned within the document without the Seeker realising this. An example of the Intermediary not communicating that they navigated from the SERP to a document, resulting in ambiguity for the Seeker as to whether a navigational interaction has taken place, is seen below:

P6 -INTERMEDIARY: I have an article on marine natural products and their potential applications as anti-infective agents

[Scanning document without modification]

P5 -SEEKER: Yeah maybe have a look into that [...]

[Access source]

Seekers believed that information items were accessible (i.e., clickable) even though they were not (i.e., non-hyperlink click [212]). It suggests that Seekers misunderstood cues of which information was accessible.

P3 -SEEKER: Uhm... could I open the recyclers recycle uhm in a new tab [...]

[Access link within document]

P4 -INTERMEDIARY: It doesn't seem to have... a... tab for that
[Feedback on what is happening]

In the above example, the lack of visual feedback played a crucial role in the navigational interactions. Providing such information in audio would improve usability for the Seeker by informing them when an item is hyperlinked or not (i.e., clickable).

We also observed Intermediaries providing summaries covering aspects of multiple documents without the Intermediary indicating this to the Seeker, and thus not giving information about the boundaries between different documents. This may suggest that incorporating *multi-document summarisation* [21] may be beneficial in transmitting information in an audio-only search setting.

The idea of a SERP (the tool⁴) and the document (the goal) is not distinctively presented in an audio-only communication setting. The lack of location-aware information throughout the search experience increases the need for further clarification communication.

Explicit Relevance Feedback

In our spoken search environment, we observed Seekers providing explicit relevance feedback without being prompted. For example, a Seeker provided positive feedback by saying: *“Yeah I think yeah that actually sounds pretty good that could potentially be relevant is there anything else or is that it?”* (P5). We also observed utterances which may be interpreted as negative relevance feedback: *“OK alright that's probably not relevant then so yeah we wanna just find something actually where does the spice cannanon cinnamon come from”* (P5).

Novel versus Previously Seen Information

A change in link colour is typically used to indicate whether a particular link on a SERP has been previously clicked.⁵ These changes provide feedback to users as to whether they have visited the underlying document. We observed several groups indicating that the same search results were displayed: for example, Intermediaries would state *“I keep on getting the same [ed. search result]”* (P6) or *“we're back to that [ed. search result] again”* (P2).

⁴We note that search engines now provide cards on the SERP and these have often become the goal.

⁵Cache and browser history was reset after each observational experiment.

Interpretation of Graphical Information

For the majority of browser-based search engine users, graphical information, such as images, tables, charts, or videos is accessible, but this is more challenging in an audio-only setting. In our observational study, Intermediaries interpreted graphical information to convey the presented information, and most of the interpretations were made of images and graphs in a document.

The next example shows a detailed conversation about a graph and table including the Seeker querying inside this information:

P4 -INTERMEDIARY: OK so it looks like it's covering World Health Organisation data from 2010 uhm and the report was published in 2014 uhm it has calculations used by people aged fifteen years and older uhm... [...]

P3 -SEEKER: Does the data uhm illustrate per capita consumption... by country?

P4 -INTERMEDIARY: Uhm... I believe that that would be... the first column... OK this is the list of countries by alcohol consumption measured in equivalent litres of pure ethanol consumed per capita per year

P3 -SEEKER: Fantastic uhm please read out the top ten

P4 -INTERMEDIARY: Uhm Belarus, Moldova, Lithuania, Russia, Romania, Ukraine, Andorra, Hungary, Czech Republic

P3 -SEEKER: Where is Australia in the list?

We also observed another interpretation of images whereby the Intermediary navigated to the image tab on the SERP to quickly gather insight into an object which she then described to the Seeker. Thus, this way of accessing images provides an overview of the information space (i.e., a set of knowledge or information units and their relationship) [135].

6.1.3 Query Reformulation

In the query reformulation stage of the information seeking process, we examine how Seekers express conditional information request statements (i.e., if a search result is true to a particular condition then use that result for further searches) and information requests within documents (i.e., the “find” (*Control+F*) function in a browser).

Automated Repetitive Search or Conditional Information Request

To save time and effort, people try to find ways to automate repetitive tasks into batches instead of performing each task individually. We observed instances of this notion of “automation” or adding conditional statements during the conversational search setup. For example, one participant pair wanted to find more information about the health benefits of eating seaweed. The Seeker (P23) had different types of seaweed in mind that she wanted to look up and in collaboration over multiple turns the pair created a short query loop. This can be summarised as illustrated in Algorithm 1.

Algorithm 1: Automated Repetitive Search (Seaweed)

Result: Which are the health benefits of different seaweeds

```

1 foreach Seaweed do find health benefits;
2 else
3   | Seaweed not relevant to search
4 end

```

Another participant pair created a conditional search task with multiple conditions. This time the Seeker (P25) wanted to investigate rivers in Turkey and Iraq before searching for dams among those rivers. For each river that had a dam, the Seeker wished to know the construction date and water volume. The example is given in Algorithm 2 to illustrate this kind of behaviour.

Algorithm 2: Automated Repetitive Search (Rivers Turkey)

Result: Did Turkish river control projects affect Iraqi water resources

```

1 foreach River in Turkey and Iraq do
2   | if They have a dam in Turkey then
3     |   | if Building date of dam and volume is stated then
4       |   |   | Compare river's volume in Iraq before and after building of the dam
5       |   |   end
6     |   end
7 end

```

It appears that these Seekers had already planned their search path before starting the search or had formed a model of the Intermediaries' capabilities. These two examples could be seen as one way of “taking control” over the search interactions by planning ahead and commanding this particular search flow. In other words, the Seeker has set out a clear path of how they want to search without handing over decision making responsibilities to the Intermediary.

Information Requests Within a Document

We observed Seekers providing an explicit information request only once they navigated to a particular SERP/document. Here, Seekers requested information about the document that was being inspected by referencing to the given backstory or pieces of information within the document. Furthermore, in some cases Seekers requested information within the navigated SERP/document concerning the given backstory, thus, revealing their information need in step-wise fashion.

P7 -SEEKER: Health benefits of marine vegetation

[Initial information request]

...

P8 -INTERMEDIARY: It just says a lot of comparing and uhm like there are some articles that start to talk about like uhm sort of plants and stuff

P7 -SEEKER: Uhm do some articles mention the use of marine vegetation as a drug like in medicine

[Information request within SERP]

In other cases, Intermediaries presented some information from the given document and Seekers wanted to know more about a specific entity provided in that document and thus explicitly queried in the document.

P5 -SEEKER: Yeah maybe click on it and see what it says so we can get a bit more information

...

P6 -INTERMEDIARY: It mainly describes about marine uhm marine cellular organisms of the sea

P5 -SEEKER: Does it say anything about it them being food?

[Information request within document]

6.1.4 Search Results Management

This last stage of the information seeking process, as defined by [Sahib et al.](#), is concerned with the search results management after the information extraction [156] and in particular which techniques users use to store the found information (e.g., note-taking, bookmarking, or favouriting).

We had not asked our participants to take notes throughout the search interactions, and we explicitly said we would not quiz them on their found information. However, five out of 13 Seekers took notes throughout the search process on the paper document with the search tasks printed.

6.2 Non-Search Interactions

This section introduces observations made which are not explicitly related to search interactions but cover broader considerations. We demonstrate how SCS introduces new aspects to the well-known one-action search paradigm observed in a conventional browser-based search setting. We continue by outlining differences among user and system models and memory, and highlight that these can be created over multiple turns in one search session or over multiple sessions. We then analyse how a SCS system can become actively involved in a search session beyond “taking initiative” and investigate the impact of the audio-only communication channel. To conclude this section, we link these four non-search interaction observations.

6.2.1 One Utterance Consists of Multiple Moves

Complexity appeared to be added in a search process by allowing users to verbally convey their query. In a browser-based search, a mouse click or key press are single *moves*. Each action a system needs to take is linked to an atomic move from the user. It could be said that we have a one-action search paradigm (action–response) in a browser-based search setting: if a user provides input (query), the system will respond (results). Search interactions in such a context can be seen as a linear process.⁶

However, we observed that this one-action search paradigm does not hold in our observational setting where information is conveyed via audio through spoken interactions with another person. Instead, we saw Seekers describing multiple moves in one utterance. An example of multiple moves in one utterance is shown in Figure 6.1. The codes below each utterance describe the actions of the turn. In this example, the Seeker first defines a navigation action and with the second part of the utterance asks the Intermediary for feedback. Intuitively, this full utterance now consists of two actions or codes.

We also observed utterances with more than two moves; however, this was rather unusual (0.47% of total dataset). These two or more moves in a single utterance increase the complexity of Seekers’ and Intermediaries’ interactions.

⁶Note, this one-action search paradigm could be manipulated by Seekers, for example by opening several tabs from the SERP.

Example: *Maybe you can get out of it then [long pause] so what's [was] the search term...*
(Seeker)

FIGURE 6.1: Example of multiple moves in one utterance.

6.2.2 User and System Models and Memory

“The overall approach is based on the idea of cognitive models or images that the components of the system have of one another and of themselves” Belkin [23, p. 111]

We observed participants building mental models (i.e., a representation or mechanism for explaining one’s understanding of an application or system) of their partner during the experiment:

Seeker Built Model of Intermediary

Some examples include Seekers creating concepts or representations of which actions Intermediaries could perform. In one instance, the Intermediary offered a function to the Seeker by asking if they would like to open a link in a new tab. The Seeker now knows that this is an option of the “system” and later in that session, the Seeker requested several links to be opened in different tabs. Later in that same session the Seeker examined the extent of the function by asking *“Could I open the recyclers recycle uhm in a new tab... if it allows that”* (P3) and thus challenged their built Intermediary model.

Intermediary Built Model of Seeker

Other instances were recognised where Intermediaries started creating an understanding of what Seekers preferred to hear as output. From the Intermediaries, we noticed two distinct differences in their utterances. Firstly, Intermediaries assumed *how the information should be presented* to the Seeker. For example, through the interaction between the participants, one of the Intermediaries was able to form a model of how the Seeker preferred to pose information requests (this particular Seeker represented her information requests distinctively with Boolean syntax). As such, the interactions allowed the Intermediary to establish a model of how the Seeker would form or structure her information and was able to mimic and present this information request formulation to satisfy her need.

Secondly, the Intermediary formed a cognitive model about *which information should be presented* to the Seeker. In this instance, the Intermediary reported the names of objects. When the Seeker posed another information request, the Intermediary checked whether the Seeker wanted object names again, even though it was not specified in the Seeker's information request. As such, a SCS should make the distinction between how the information should be presented (**form**) and which information should be presented (**content**).

Creating Memory over Multiple Turns/Sessions

In this example, the Seeker asked for “numbers” (i.e., numerical information) for a particular backstory. For the next backstory, the Intermediary directly asked whether the Seeker would like to navigate to the statistics section. This demonstrates an example of creating memory over multiple turns. In another example, a participant pair had learned from a previous backstory that they could use Google Scholar which the Seeker preferred. In the next search task, the Intermediary explicitly mentioned that scholarly articles were available for their information need. The interactions demonstrate that memory may be created over multiple sessions as well as multiple turns [146].

6.2.3 Decision Offloading and Taking Control

We observed Intermediaries applying many different techniques to deal with the challenge of transferring information through an audio-only communication channel. Examples include reading out search results sequentially, summarising a SERP, or requesting feedback as to whether more information had to be transferred.

We also noticed that Intermediaries became more involved in assisting to express the Seeker's information need, adopting a leading approach. In the following example, the Intermediary refines or *rewrites* the Seeker's utterance into a specific query.

P23 -SEEKER: ... cinnamon is from Europe, so I was trying to look uhm is it from Europe or from other places

[Information request]

P24 -INTERMEDIARY: I look up cinnamon suppliers... in Europe

[Query refinement offer]

We observed Intermediaries actively trying to satisfy the Seeker's information need by making decisions and thus taking more control over the search process. More specifically,

these assisting and leading Intermediary utterances suggest that Intermediaries have a significant role in deciding which information is transferred. The Intermediaries are making selections as to what information is appropriate to share at a given moment. This decision making process may also suggest that Intermediaries have to calculate the *cost-benefit*, which may influence Seeker-satisfaction, associated with each strategy to decide which one would be more likely to benefit the Seeker.

These observations corroborate that, given the high cost of delivering information via a linear channel such as speech, it is not optimal to present all found information. Instead, the system needs to decide which information it should offer at each interaction by continuously estimating the cost-benefit to the user.

In contrast to having the Intermediaries decide what information to transfer, we also observed Seekers explicitly requesting the Intermediary to make decisions for them, e.g.: “*Uhm do you think that’s enough to get an idea of where it actually came from or do you think we should keep going?*” (P5). It could be suggested that this particular decision-offloading example is an artefact of the Seeker being aware that there is an Intermediary (i.e., a human). However, this would warrant further exploration in a WOZ setting.

6.2.4 Effective Information Transfer

Sometimes actors misheard each other (i.e., the information transfer was not successful or was disrupted) and had to *repair* their conversation [165]. To repair, actors requested a repetition of a previous utterance: “*Sorry what, can you repeat that sentence*” (P3) or “*can you repeat that please*” (P20). Actors were also observed hesitantly repeating back what the other had said. In other situations, actors misapprehended a message and were later corrected by their partner.

6.2.5 Linking Non-search Related Observations

In the above sections, we provided observations which suggest that the audio-only interaction channel will greatly impact the interactions between the users and the future SCS system. Interacting verbally increases the flexibility of what users can provide as input, which was illustrated with the observation that one utterance could consist of multiple moves. However, this flexibility also increases the complexity of the belief regarding what a system or user can do (cognitive user model) as there are no conventional or pre-set interaction paths. Even though the responsibility for decision making could be shared between actors or shifted from one to another, all this is only possible when the information transfer is successful and effective.

6.3 Chapter Summary

This chapter explored SCS, an emerging interactive search paradigm wherein all interactions are performed through audio. We conducted an observational study to learn about this new phenomenon in a structured way. We not only presented the observations concerning search but also included non-search interactions (i.e., utterances to help repair the conversation). We suggest that non-search interactions have become an integral part of the search process and should be included in future information seeking models. Each presented observation was described, analysed, and where possible examples were given to illustrate the workings of that particular observation. We concluded that this new paradigm is much more complex and interactive than the search scenarios/paradigms covered by existing models.

The primary inference of this chapter is that even though many audio-only interactions are similar on the surface to a conventional text search interaction, each interaction comes with additional complexity due to the uncertainty of the correct information transfer. We also suggest that future information seeking models may have to include the multi-level classification of search and non-search interactions. In particular, we suggest systematically investigating utterances from the observational study to better understand these nuances which form an intricate part of this audio-only search. We aim to address this in the next chapter.

Chapter 7

Identifying, Classifying, and Validating the Interaction Space for Spoken Conversational Search

In the previous chapter, we identified that a more rigorous investigation is needed to better understand the possible user–system actions in SCS. In this chapter, we analyse the SCS data through identifying and classifying the transcribed utterances of our empirical laboratory study.¹ Thus we investigate the role of these information seeking communications [197]. We present the themes and sub-themes which are based on the constructed codes (or labels) derived from the thematic analysis and present them in the SCS annotation schema, *SCoSAS*. These themes provide the characteristics of information seeking dialogues in a conversational setting, the actor’s role, and the actor’s relationship with the conversation. We also validate the *SCoSAS* to verify its consistency, correctness, and usefulness.

Even though annotation schemas for SDS often include some notion of information seeking/providing functions, currently these functions are very high-level and are created for general purpose functions only [42]. Therefore, we created this annotation schema due to the lack of previously defined classification designs for information seeking.

The overview of the methods used in this chapter has been explained in the introduction of Part III. In summary, the methodology relevant to the experiment reviewed in this chapter is thematic analysis which is described in Section 5.8.1. Our key outcome is

¹This chapter consists of the following publications J. R. Trippas, D. Spina, P. Thomas, H. Joho, M. Sanderson, and L. Cavedon. Towards a model for spoken conversational search. *Information Processing & Management*, 2019. (Submitted) and J. R. Trippas and P. Thomas. Data sets for spoken conversational search. In *Proceedings of the CHIIR 2019 Workshop on Barriers to Interactive IR Resources Re-use (BIIRRR 2019)*. *CEUR-WS*, pages 14–18, 2019.

a classification formation that establishes the SCoSAS and consists of three themes: *Task Level*, *Discourse Level*, and *Other Level*. Additionally, we validated the annotation schema through calculations of inter-rater reliability and code overlap with the SCSdata and an external annotator. Finally, we calculated the code overlap and coverage between the SCSdata and MISC for validating our schema.

The significance of this chapter is twofold: we *(i)* develop a classification schema and *(ii)* test and validate this schema. More importantly, we show that our schema is generalisable and replicable for a SCS setting.

This chapter is a key component in the process of creating meaningful annotated SCS datasets by extracting and abstracting communication behaviours. Thus, we demonstrate how to construct accessible datasets by distilling the complexity of spoken information seeking conversations. In particular, this chapter allows for the organisation of our collected data by way of classification.

To summarise, this chapter is divided into the following sections: in Section 7.1, we set out the aims of the information seeking conversation analysis. In Section 7.2, we describe all the identified themes and sub-themes, and we then validate the coding consistency within the SCSdata in Section 7.3. We further validate our coding schema in Section 7.4 by transferring our annotation schema to a different dataset. This chapter concludes with the summary in Section 7.5.

7.1 Aims

Our observational study was conducted to investigate the possible interaction space of a SCS system [197]. This study was designed to understand how users communicate in an audio-only search setting and focusses on the characteristics of this interaction paradigm. Thus, observing how people search in this setting provides initial insight into the possible scope of these future systems.

This analysis chapter aims to outline and explain the coding process for thematic analysis including the production of our coding schema, SCoSAS, and validity checks (generalisability and replicability of coding schema). Our detailed descriptions strengthen our analysis and aim to address the lack of documentation transparency on how annotation schemas are developed.

7.2 Utterance Classification: Themes for SCS

In this section, we first set out the specifics of the code separation to analyse these codes independently of each other. We then describe each theme and their corresponding sub-themes to finalise the deduction of the SCoSAS from those identified classifications (i.e., themes and sub-themes).

To understand which actions are taken, we split all codes where more than one code was attached to an utterance — thus creating atomic actions per utterance for a more natural grouping of these actions into themes and sub-themes. We present the three themes and their corresponding sub-themes as follows. The first theme, *Task Level*, is related to search interactions and the topical investigation. The second theme, *Discourse Level*, is associated with communicative functions between the Intermediary and Seeker for smooth collaboration. The third theme, *Other*, consists of utterances that belong to neither the Task nor the Discourse levels. Example utterances are provided for each sub-theme throughout the chapter. Tables of all the themes, corresponding sub-themes, participants (or actors), and codes are included in Appendix D.

The themes, sub-themes, and codes are all based on the SCSdata. All examples given in this section are from our dataset. We validate our annotation schema to a different dataset in Section 7.4 to illustrate the validity of the SCoSAS.

7.2.1 Theme 1: The Task Level

The Task Level theme covers search actions such as query expressions and search results presentation utterances. In other words, this theme is related to the performed task which, in our case, is a search task. The theme includes four sub-themes: *(i) Information Request* which includes utterances related to (re)forming information needs by both Seeker and Intermediary, *(ii) Results Presentation* which includes of search result transfer utterances from the Intermediary, *(iii) Search Assistance* includes Seekers asking for or Intermediaries providing help with the search task, and *(iv) Search Progression* includes Seeker’s feedback on the progress of their search task.

Information Request

The Information Request sub-theme covers utterances which are associated with the topical information requests and is used by both Seeker and Intermediary. Thus, all utterances which are related to forming, suggesting, refining, confirming, repeating,

spelling, or embellishing information requests are captured in this sub-theme. The following example is of two information requests.

P13 -SEEKER: So which state in Australia consumes the most alcohol per person?

[Information request]

P14 -INTERMEDIARY: Again 2016 or the most recent information?

[Information request]

Information requests from Seekers could be expressed at any time, and Seekers often asked for information from a document itself, provided clarification related to their search intent, or requested more meta-information about a document or SERP. Conversely, Intermediaries were more likely to provide support in (re)forming the information request, for example by providing information request refinements, suggesting query expansions, or eliciting extra information.

In other words, this sub-theme is linked to query formulation and reformulation stages as covered by Sahib et al. [157] and discussed in Chapter 6.

A distinction can be made within the Information Request sub-theme for Seekers where they interact or manipulate results by requesting further information in the following two ways:

1. Requesting information *about* a document or SERP (which could be interpreted as a meta-information request),
2. Requesting information *within* a document or SERP.

Results Presentation

The Results Presentation sub-theme is where Intermediaries read out, interpret, or provide an overview of a SERP or document to the Seeker. These sub-theme utterances convey the results from the search engine or documents. Moreover, this sub-theme is only used by Intermediaries and the majority of their actions are linked to this sub-theme.

In the next example, the Intermediary reads out the results exactly as they were displayed in a document:

P6 -INTERMEDIARY: The history of valuable cinnamon. The first mention of cinnamon is in Chinese documents dating from 2800 BC. The

ancient Egyptians logged cinnamon as a spice used in the embalming process...

[Results presentation]

Other categories of utterances where Intermediaries conveyed the documents or search engine results but modified them (i.e., interpreting the results so that they would be most beneficial for the user) are also categorised in this theme. Intermediaries modified SERPs or documents in the following ways:

- Synthesised (synthesis is a combination, usually a shortened version, of several texts made into one) or provided an overview
- Interpreted
- Paraphrased (the ideas of another person in your own words)
- Summarised (a shortened version of the text)
- Clarified
- Compared

Search Assistance

This sub-theme captures interactions where the Intermediary assisted the search process by providing explicit search suggestions and advised searching for more information or moving to the next topic. Other examples include relevance judgements as seen below:

P2 -INTERMEDIARY: So uhm here it talks about call centres outsourcing uhm... then it talks about human resources outsourcing uhm... there is a lot on health benefits conversation uhm [long pause] I don't see how some of these are relevant...

[Search assistance]

In contrast to directly providing assistance, Intermediaries sometimes asked for support to create a better understanding of how to help the Seekers in their search process. For example, Intermediaries asked about the usefulness of a result, requested spelling, or indicated that switching to a different search engine would be helpful (e.g., Google Scholar or library search).

Additionally, this sub-theme captured the Seeker explicitly asking for assistance during their search session, for example, by asking for recommendations or judgements on whether they covered enough of the information space as seen in the next example:

P5 -SEEKER: Uhm do you think that's enough to get an idea of where it actually came from or do you think we should keep going?
[Search assistance]

Search Progression

This sub-theme is only used by the Seeker to provide feedback on the search progression to the Intermediary. Examples include specific performance feedback on their search, rejecting search results, or informing the Intermediary whether they found enough information for a topic:

P15 -SEEKER: OK that's probably enough information
[Search Progression]

7.2.2 Theme 2: The Discourse Level

The Discourse Level theme includes utterances with a communicative function from both Seeker and Intermediary. This theme covers traits related to the audio channel. The Discourse Level theme consists of four sub-themes: *Discourse Management* which allows the conversation to take place between the actors; *Grounding*, also referred to as “common ground”, which captures the dialogue interactions for the creation of mutual knowledge, beliefs, and assumptions between the two actors [52, 189]; *Navigation* which covers the communications of moving around web pages, documents, and browser tabs; and *Visibility of System Status* which allows actors to provide insight on what is happening throughout the interactions.

Discourse Management

This sub-theme includes conversational coherence and cohesion between the actors [163]. In other words, the utterances in this sub-theme are part of the communication between the actors to check whether the message has been understood (i.e., meta-discourse). In our dataset, these discourse building utterances are independent of the participant role. For example, both Seeker and Intermediary confirmed, checked, asked for repetitions, or repeated utterances as illustrated in the snippet below:

P1 -SEEKER: So uhm can you go and change the search question to effectiveness of uhm... passenger and baggage screenings at airport

P2 -INTERMEDIARY: Passenger and
[Discourse management]
 P1 -SEEKER: Baggage
[Discourse management]

Often an information request was repeated, echoing back the previous speaker's exact words, to confirm a command. These discourse actions are crucial to have a meaningful conversation, for example indicating that one actor has understood the other actor. Thus, repetition of an utterance such as an information request, without adding extra information, is part of the discourse management to keep the conversation going.

Grounding

Grounding in communication as described by [Clark and Brennan](#) is “sharing and synchronising mutual beliefs and assumptions” and is fundamental for communication between actors [52]. We observed utterances belonging to this particular sub-theme which was used by Seekers to coordinate the shared information or common ground [52]. This function allowed the Seeker to share their meaning of aspects of beliefs and values framework. In other words, the two actors' mental model of each others' beliefs may benefit from adapting continuously to coordinate the build of a mutual understanding. Seekers summarised or paraphrased the information given to them and created a larger picture of the search results as a way of synchronisation. Through this dynamic updating process of the Seekers' mental model, Seekers provided insight into what they understood from the information provided. By receiving this feedback Intermediaries then had an opportunity to know whether the provided information was correctly conveyed. In this example, we see the Seeker coordinating their beliefs of alcohol consumption quantities:

P14 -INTERMEDIARY: [...] yeah 20 to 29 is the most high risk drinking people in Australia for alcohol related harm... I don't know what that means about consumption
 P13 -SEEKER: Yeah so they consume a lot
[Grounding]

Grounding differs from Search Progression and Discourse Management. While Grounding involves sharing the beliefs and values of the information, Search Progression is concerned with the feedback on the search task progress and Discourse Management is related to effective information transfer.

The Grounding sub-theme was only seen in Seekers' utterances. This is because Intermediaries, by having the information to hand, summarised results presented and they did not need to confirm or share their beliefs or meaning of the content. As such, their utterances are captured by the theme Results Presentation.

Navigation

Navigational utterances are part of the discourse between the actors to progress the task allowing them to manoeuvre around the online information space. Seekers navigated the search results by instructing the Intermediaries. In our case, Seekers asked to access specific sources, navigated between documents, singled out particular documents, and read more from a document or the next document. In other words, Seekers provided information about how and where they wanted to navigate to as seen below:

P9 -SEEKER: Uhm maybe uhm can you go into the result [...] that mentions
how uhm outsourcing damages the industry
[Navigation]

Visibility of System Status

“The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.” Nielsen [136, p. 1]

Seekers asked the Intermediaries to provide information on what was occurring throughout the interactions. Intermediaries provided feedback on what was happening for example if they had seen certain items before, or by orienting where they were positioned. The example below illustrates the Intermediary indicating that their process is still pending followed by the Seeker requesting an update.

P25 -SEEKER: Oh TIBER sorry Tiber yeah
[Discourse management]

P26 -INTERMEDIARY: Yeah uhm just searching just one second
[Visibility of system status]

P25 -SEEKER: Any luck?
[Visibility of system status]

7.2.3 Theme 3: Other

Five utterances from the Seeker were not classified in any of the above (sub-)themes. Two of these utterances were disfluencies from the Seeker, one utterance was where the Seeker provided information about the search engine, one utterance was asking if the Seeker was allowed to embellish a query, and the last unclassified utterance involved the Seeker offering to spell a word. These five categories were not classified after much deliberation and given the theme “Other” instead (see Appendix D).

7.2.4 SCoSAS Subtraction

An overview of the themes and sub-themes used by each actor in the SCSdata is presented in Table 7.1. The development of the utterance classifications in themes, sub-themes, and codes form the basis of the SCoSAS.

TABLE 7.1: Themes and sub-themes used by different actors.

Theme	Sub-theme	Seeker	Intermediary
Task Level	Information Request	✓	✓
Task Level	Results Presentation		✓
Task Level	Search Assistance	✓	✓
Task Level	Search Progression	✓	
Discourse Level	Discourse Management	✓	✓
Discourse Level	Grounding	✓	
Discourse Level	Navigation	✓	
Discourse Level	Visibility of System Status	✓	✓
Other		✓	

7.3 Inter-rater Reliability and Code Overlap

As part of the validation and quality protection of the SCoSAS, we calculate the inter-rater reliability and code overlap in this section. These measures quantify the agreement and consensus between different coders. The inter-rater reliability illustrates the consistency between the coders.

Assessor 1 (myself) analysed the data using thematic analysis and created codes and a codebook which acted as the annotation schema. A second independent researcher (Assessor 2) used the codebook for closed coding of all utterances in the SCSdata. The inter-rater reliability on code level was moderate (Cohen’s $\kappa=0.59$) [119].

Both coded datasets were then converted to sub-theme level² where the inter-rater reliability between Assessor 1 and 2 on sub-theme level was calculated. The inter-rater reliability on sub-theme level was substantial (Cohen’s $\kappa=0.71$).

The overlap of codes used between the two independent assessors was high with 90% of the predefined codes being used by both assessors. Assessor 1 applied 84 different codes consisting of 41 codes for the Seeker and 43 for the Intermediary. Assessor 2 used 76 codes, 38 codes for the Seeker and 38 for the Intermediary as seen in Table 7.2. In other words, Assessor 2 used a smaller range of pre-defined codes to label all utterances. Substantial agreement was met for inter-rater reliability on sub-theme level.

TABLE 7.2: Independent Assessors’ code overlap.

	Assessor 1	Assessor 2
Total number of utterances	1044	1044
Total number of codes used	84	76
Total number of codes for Seeker	41	38
Total number of codes for Intermediary	43	38
Unused codes	0	8 (10%)

The remaining 10% of codes which were used by Assessor 1 but not by Assessor 2 represented eight codes which were used 13 times in the dataset, as seen in Table 7.3. The differences between the codes of Assessor 1 and 2 were investigated and considered to be minor discrepancies. Assessor 2 was also consulted about the coding schema and code definitions were refined for clarity.

TABLE 7.3: Codes used by Assessor 1 and not by Assessor 2.

	Code	Actor	Number of times used
1	Definition lookup or person	Seeker	1
2	Asks to repeat nth search result	Seeker	1
3	Automated repetitive search	Seeker	3
4	Wayfinding	Intermediary	3
5	Interpretation of photos	Intermediary	1
6	Image overview on SERP	Intermediary	2
7	Within-document search result entity lookup request	Intermediary	1
8	SERP overview without modification	Intermediary	1

7.4 Validation of SCoSAS

In this section, we validate the labelling schema SCoSAS by applying it to a similar spoken conversational dataset, The Microsoft Information-Seeking Conversation data (MISC) [185].³ First we outline the MISC dataset which was created by Thomas et al.

²The coding of the utterances was completed on utterance level.

³The MISC data was accessed at <http://aka.ms/MISCv1>.

to support conversational retrieval interfaces [185]. Then, we provide the validation process of the SCoSAS with MISC. We describe the similarities and differences between the SCSdata and MISC. We outline how the two datasets were comparable through transposing SCoSAS labels and we calculate the overlap and coverage of the SCoSAS with the MISC. We conclude this section by presenting differences in the labelling and discussing the results of the SCoSAS validation.

7.4.1 MISC Dataset

The MISC is a set of recordings of spoken conversation between human “Seekers” and “Intermediaries” [185]. It was designed to support research on questions such as: Do human Intermediaries show behaviours which correlate with Seeker satisfaction?; Do Seekers show behaviours which we could use as a baseline for online metrics, appropriate to conversational agents?; What role does politeness or other conversational norms play?; What tactics do we see in information seeking conversation, and do particular structures help or impede progress or satisfaction? The MISC has been used in published work on conversational style [186] and on multimodal collaboration [130].

MISC Study: The MISC data includes audio and video signals; transcripts; prosodic and linguistic signals; entry questions on demographics and personality; and post-task surveys on emotion, engagement, and effort. Screen recordings are also available, as is data on affective and physiological signals.

The overall setup for both the SCSdata and MISC recordings is similar. In the MISC, tasks were also assigned to a “Seeker” who was responsible for gathering information and writing down a final answer. An “Intermediary” substituted the future SCS system. The Intermediary had unrestricted access to the web, including search engines. Instead of having the two participants in one room as in the SCSdata, the MISC participants were connected over an audio link, and both video and audio of both participants were recorded.

Available MISC Data: The available MISC dataset includes different raw and derived data. For example, the raw audio is included as well as the transcripts.

The MISC includes five information seeking tasks, one of which was used as practice. These tasks were selected to reflect a range of complexity and task difficulty. The MISC also includes tasks which elicited positive and negative emotional responses. As in the SCSdata, participants solved the tasks by using the open web.

7.4.2 Validation of SCoSAS with MISC

The SCSdata is very similar to the MISC and includes recordings of information seeking conversations between two people as Seeker and Intermediary. The MISC contains audio and video recordings with ASR transcriptions of these recordings.

To understand whether we covered the majority of possible actions for this new SCS interaction paradigm, we applied our coding schema to the MISC. We coded the MISC dataset according to our predefined labels to investigate which actions were or were not covered by our annotation schema. Thus, by using our predefined SCoSAS codes, we validate the coverage (i.e., is there an action applicable for every situation?) and overlap (i.e., is there a situation where more than one action could be relevant?). We used these measures to review whether the saturation of themes reached for the SCoSAS, was an appropriate validity measure in qualitative analysis [10].

Every utterance in the MISC dataset $u \in D$ is part of an information seeking conversation C which has been transcribed using ASR. Thus, an information seeking conversation is represented as a sequence of utterances $C = \{u_1, u_2, \dots\}$ which receives a tag t from the pre-defined set of tags $t \in T$. The tags were generated from the SCSdata. We report on the coverage of the utterances u to the pre-defined set of tags T .

We performed the following steps described in Sections 7.4.3.4–7.4.7:

1. Labelling a subset of utterances from MISC
2. Creating comparable datasets
3. Reporting on the overlap and coverage of SCoSAS on MISC
4. Describing the non-overlapping codes from MISC
5. Discussing the results of the SCoSAS validation

The above steps enabled us to validate our existing annotation schema, the SCoSAS.

7.4.2.1 MISC Data Statistics and Subset

The MISC dataset was created from 22 participant pairs with each pair completing five information seeking tasks. The participants were randomly assigned a role as Seeker or Intermediary and had ten minutes to complete each information seeking task. The participant pairs spent on average 8 minutes 20 seconds on each task. Participants exchanged on average 857 words per task.

We selected a random set of four participant pairs and labelled four tasks per participant pair. The following pairs were selected as the MISC subset:

- Pair A (Participants 1–2)
- Pair D (Participants 7–8)
- Pair J (Participants 19–20)
- Pair N (Participants 27–28)

In the four pairs, we have a total of 701 turns with an average of 175.25 turns per pair and an average of 43.81 turns per task. However, 4.99% of the total turns which we labelled in the MISC dataset were inserted due to ASR errors and were not present in the audio. This is further explained in Section 7.4.3.3. These turns were ignored which means that a total of 666 turns were labelled on code-level with an average of 166.5 turns per pair and an average of 41.62 turns per task.

The SCSdata consists of 1044 turns with an average of 80.30 turns per pair and 26.76 turns per task.

7.4.3 Differences Between the SCSdata and MISC Datasets

The setup and instructions between the SCSdata and MISC dataset were marginally different. We provide an overview of the differences in this section.

7.4.3.1 Search Tasks

The MISC search tasks were selected by [Thomas et al.](#) for their varied level of difficulty and complexity, as illustrated in Table 7.4 [185]. These tasks were also designed to elicit positive and negative emotions.

TABLE 7.4: MISC search tasks.

Task	Difficulty	Complexity	Emotion
1	Low	Low	Positive
2	Low	High	Negative
3	High	Low	(NA)
4	High	High	Positive

7.4.3.2 Setup of SCSdata and MISC

MISC Seekers were given a search task and asked to write down an answer for each provided search task. They did not have access to any information source but received an information need which they were allowed to read out in order to share it with the Intermediary. The Intermediary had access to a computer with a search engine. The

Seeker and Intermediary were located in different rooms, and all communication was done through an audio connection.

In contrast, SCSdata Seekers were not allowed to read out the search task and had to share the information seeking task with the Intermediaries by paraphrasing the task. The SCSdata Seeker did not have to write an answer for each task.

7.4.3.3 Transcription Differences

The MISC data was transcribed with ASR while the SCSdata was manually transcribed and subjected to the three-pass-per-tape policy [131]. ASR incorrectly transcribed utterances, including inserting “thanks” when speakers did not say this in the audio. Such utterances were therefore ignored for this analysis.

The following snippet of a conversation is an illustration of the ASR transcribing utterances which were not present in the audio, making speakers appear more polite.

P20 -INTERMEDIARY: [...] She wanted them to donate to charity

P19 -SEEKER: Thanks

[Utterance not present in audio]

P20 -INTERMEDIARY: To provide clean water // and she um

P19 -SEEKER: Thank you

[Utterance not present in audio]

We encountered segments where the researcher interfered due to a technical issue (Participant Pair D) and sections where the ASR created many unnecessary turns between the actors because it detected that someone was talking, but there was no evidence of this in the audio.

7.4.3.4 Utterance Labelling

To ensure good labelling performance, we coded the SCSdata on the video and audio recordings and on the transcriptions, including the Intermediary’s screen capture. We also coded the MISC subset with the predefined codes on audio recordings and transcripts. However, we were unable to label Results Presentation utterances at code-level and coded them at sub-theme level. This was due to the unreleased screen capture videos at the time of analysis. Thus labelling Results Presentation utterances at code-level meant that subtleties such as whether an Intermediary was reading from a SERP or a document could not be distinguished.

Furthermore, new possible labels were generated if none of the existing labels were suitable. We discuss these additional labels in Section 7.4.6.

7.4.4 Creating Comparable Datasets

The code level investigation provides insight into the number of actions shared between the SCSdata and MISC. The SCoSAS, as defined on the SCSdata, consists of 84 unique codes where Seekers used 41 different codes and Intermediaries used 43. These 43 Intermediary codes from the SCSdata were reduced to 25 codes to create a comparable labelled dataset with MISC. Thus, codes in the SCSdata such as “Scanning document without modification” were coded at code-level but later mapped and transposed to the Results Presentation sub-theme to create comparable datasets. Transferring the Results Presentation codes to the sub-theme level reduced the 84 unique codes of the SCSdata to 66 codes as seen in Table 7.5.

TABLE 7.5: SCSdata and MISC dataset descriptives.

	SCSdata	MISC subset
Total number of utterances	1044	666
Total number of unique codes*	66*	49*
Unique codes Seeker	41	31
Unique codes Intermediary	25	18

*NOTE: Due to insufficient details, utterances which were related to presenting results were aggregated to the Results Presentation sub-theme level. The SCSdata’s unique number of codes without aggregation of the Results Presentation is 84.

7.4.5 Code Overlap and Coverage Between SCSdata and MISC Data

In this section, we investigate the code overlap and coverage by examining the SCoSAS between the SCSdata and MISC which provides an understanding of the scope of possible actions.

Overlap: $SCSdata \cap MISC = \{x : x \in SCSdata \text{ and } x \in MISC\}$. The overlap between SCSdata and MISC datasets is 35 codes (71%).

Coverage: The coverage or union between datasets SCSdata and MISC shows how the sets relate to each other where $SCSdata \cup MISC = \{x : x \in SCSdata \text{ or } x \in MISC\}$. In total, the transposed SCoSAS consists of 66 different codes⁴ (41 Seeker codes and 25 Intermediary codes). The MISC dataset consists of 49 different codes (31 Seeker codes and 18 Intermediary codes). The union of the two datasets’ codes creates a set of 80

⁴Note: Results Presentation codes have been transferred to sub-theme level for comparison with the MISC dataset.

different codes, with the SCoSAS covering 82.5% of these possible codes and 94% of the MISC utterances could be coded with the SCoSAS.

The set of 14 supplementary codes coded in MISC but not in SCSdata is presented in Table 7.6.

TABLE 7.6: Set difference between MISC and SCS datasets.

	Code	Actor	Nr used
1	Chitchat	Seeker	1
2	Communication about the task	Seeker	2
3	Decision offloading	Seeker	1
4	Feedback on writing down the answer for the given task	Seeker	3
5	Negotiation	Seeker	7
6	Rejects spelling offer	Seeker	1
7	Requests spelling	Seeker	1
8	Uncertainty expression of what to search	Seeker	2
9	Chitchat	Intermediary	5
10	Enough information?	Intermediary	9
11	Negotiation	Intermediary	6
12	Offers to spell	Intermediary	1
13	Spells	Intermediary	5
14	States “too many results to sum up”	Intermediary	1
Total number of instances of code used by MISC and not by SCS			45 (6%)

7.4.6 Descriptions of Code Set Differences

In this section, we investigate the 14 different codes found in the MISC but not in the SCSdata.

Chitchat or Negotiation

We encountered new types of utterances in the MISC where the actors were negotiating or chitchatting. The negotiation utterances were used to bridge differences and reach agreements [229]. Examples include instances where actors share their own experiences about particular topics or subjects. However, this is not to be confused with the already defined Grounding sub-theme which covers utterances from the Seeker expressing their beliefs and values of information provided by the Intermediary.

Chitchat and negotiation utterances have greater overlap between speakers, meaning that more than one actor at a time is speaking [161]. For example, the following utterances overlapped while the Seekers and Intermediary negotiated their shared understanding of non-traditional medicine:

P1 -SEEKER: I think herb sounds more like // not
[Negotiation]

P2 -INTERMEDIARY: More like medicine
[Negotiation]

P1 -SEEKER: I think it sounds more like naturopathic but that fits it
[Negotiation]

Participants seemed forthcoming in sharing their own opinions and experiences. The following example is from an Intermediary who shares her own travel experiences which are related to the task:

P8 -INTERMEDIARY: That's what I love to do actually when I traveled all the public transportation and all sorts of continents
[Chitchat]

Communication About the Task

SCSdata participants were instructed not to read out their provided search task to the Intermediary but instead were asked to rephrase and formulate their information request. In contrast, MISC participants were allowed to read out their search task. This resulted in Seekers also talking informally about the search task itself and how they understood or interpreted the task. For example,

P1 -SEEKER: Yeah the task is a bit // um very generalised so um

Agency and Decision Offloading or Taking Control

Due to reading out the search task in the MISC, both the Seeker and the Intermediary shared a similar objective of their search need. This shared search task created a balanced level of collaboration between the two actors which allowed the Intermediary to instantiate agency more frequently. By contrast, Intermediaries in the SCSdata acted more as the interface between the Seeker and the found information.

The notion of agency returned throughout our subset of the MISC in utterances resulting in the following codes “Enough information?” (Intermediary), “Too many results to sum up” (Intermediary), and “Decision offloading” (Seeker). For example, the Intermediaries suggested that a search task has been finished “*excellent, so we are finished...*” (P8), or they stated that they were not going to sum up all the results because there

were too many. Simultaneously, the Seekers also handed over the decision making to Intermediaries by uttering “*it’s up to you [ed. if we look at the other site or not]*” (P20). Hence, we cluster these codes into ‘agency and decision offloading or taking control’.

Feedback on Writing Down the Answer for the Given Task

As part of the MISC data collection setup, Seekers were asked to write their answers for the information seeking tasks. The MISC Seekers communicated how they were progressing with the writing task as presented in the example below.

P1 -SEEKER: Okay so I’ll just have to put this in another category [ed. on the answer sheet]

Spelling

We encountered instances of spelling actions in the MISC which had not been encountered in the SCSdata (i.e., offers from the Intermediary to spell out words). These spelling actions may have been because Seekers were required to write down the information they found. Therefore they needed to know the spelling more frequently.

Uncertainty Expression of What to Search

In this utterance, the Seeker is expressing their confusion regarding what the information need asks them to fulfil. This Seeker is expressing their uncertainty which possibly could be seen as asking the Intermediary for help to critically investigate the search task which had been read out.

P19 -SEEKER: I am not sure what you’re supposed search

7.4.7 Discussion of SCoSAS Validation

The majority of the codes (71%) which were coded in the MISC overlapped with the SCoSAS. The remaining codes (29%) were instantiated 6% of the time throughout the full MISC subset. In other words, the most significant utterances in the MISC subset are covered by our SCoSAS coding scheme. After investigating the different codes from the MISC which did not appear in the SCSdata, we believe that some of these newly encountered codes could be candidate expansion codes to the SCoSAS, such as the

array of possible spelling requests, suggestions, or rejections. We also believe that some of these codes were not encountered in the SCSdata due to the difference in study setups, such as the instantiation of the communication about the task. Nevertheless, 95% of all the utterances in the MISC were covered by our coding schema developed on the SCSdata. No new themes were identified suggesting that saturation, which is often used as a justification for sample size in qualitative work, was reached [70]. Furthermore, our sample of 13 participant pairs for the SCSdata provided all themes with most codes.

7.5 Chapter Summary

This chapter analysed the SCSdata's information seeking conversations by interpreting and classifying the utterances with thematic analysis. The analysis resulted in three themes, eight sub-themes, and 84 codes. The internal coding consistency was then validated by calculating the inter-rater reliability and code overlap of a second annotator. The remainder of the chapter was devoted to illustrating the generalisability and replicability of our SCoSAS by applying our schema to a similar dataset, the MISC.

This transparent annotation process contributes by strengthening the analysis and the methodological foundations of annotation schema development.

Our analysis is validated through several methods. Firstly, the theme and sub-theme process was monitored by multiple researchers. Secondly, an external assessor recoded the full SCSdata with our codebook, and lastly, we coded a different but similar dataset with our set codes. Despite this validation, we acknowledge the limitation of only having one researcher develop the leading labels.

The implications of this analysis are many. Firstly, this analysis can support the feature extraction of particular utterance-types, or assist with the engineering and evaluation of conversational retrieval. The analysis can also be used for language modelling of information seeking conversations and the development of results presentation strategies.

Our contributions in this chapter are the following: *(i)* we establish the interaction space for SCS which resulted in themes, sub-themes, and codes to extract the SCoSAS, *(ii)* we provide a transparent and well-documented analysis of the utterances to define that interaction space which strengthens the findings, *(iii)* we illustrate that our coding schema is generalisable and replicable through validation calculations with our own SCSdata and a second SCS dataset, the MISC. In the next chapter we use the SCSdata and SCoSAS to demonstrate the applicability of the dataset and annotation schema.

Chapter 8

Task Complexity and Interactivity for Spoken Conversational Search

We utilise the SCoSAS-labelled SCSdata which we developed in Chapter 7 to perform further analysis to investigate the effect of task complexity on interactivity (search and discourse).¹ Search tasks are an essential component of interactive search studies [26]. These tasks are often used to evaluate a system or to observe people’s behaviour with a system. In many cases, search tasks are manipulated as part of the research design to study different interaction behaviours [213].

In Chapter 7, we analysed the interaction behaviour of thirteen participant pairs who executed search tasks with different levels of cognitive complexity based on the Taxonomy of Learning [9]. This chapter aims to understand whether different interaction behaviours are used depending on the cognitive complexity of the task. Our results show that users require greater interactivity to satisfy the information need in more cognitively complex tasks. On more complex tasks, participants spent more time on the task, posed a higher number of information requests, and engaged more in meta-discourse interactions. These results contribute to the formulation of the SCS complexity, the information seeking behaviours, and the relationship among the characteristics of the audio-only communication channel.

¹This chapter consists of the following publication J. R. Trippas, D. Spina, L. Cavedon, and M. Sander-son. How do people interact in conversational speech-only search tasks: A preliminary analysis. In *Proceedings of Conference on Information Interaction and Retrieval (CHIIR)*, pages 325–328, 2017.

Our results emphasise the complexity that an audio-only interaction channel imposes in a search process. We observe this greater complexity through the increase in meta-discourse interactions in more complex tasks. We suggest that more meta-discourse utterances such as confirmations are used to overcome disruptions in commutation flow in more complex tasks. In addition, further research into interaction behaviour may help us understand how users recover from errors in an audio-only setting, how they navigate when no visual boundaries are present, and whether particular interaction chains (i.e., conversational routines) are a predictor for success or failure of a task [16, 212].

The chapter is structured as follows: Section 8.1 introduces the importance of studying the cognitive complexity and interactivity in SCS, and we state our research questions. Section 8.2 presents the methodology of our experiment and the interaction behaviour measures used for our analysis. We then present the results in Section 8.3, followed by the discussion in Section 8.4, and conclusion in Section 8.5. We end this chapter with the summary in Section 8.6.

8.1 Introduction

Many studies have investigated cognitive complexity of search tasks in browser-based search [11, 16, 108, 125]. Other studies in linguistics or pedagogy have investigated how discourse is affected by task complexity [59, 79, 151]. However, little is known about the impact of task complexity in SCS. This chapter seeks to address this. In particular, we investigate how task complexity in SCS affects the interactivity or interaction behaviours in both search and non-search interactions. We chose different cognitive complexities to observe various techniques and interaction behaviours used in an audio-only search setting.

We use the SCSdata as described in Chapter 5 and the annotated dataset as specified in Chapter 7. We use categorisations and classifications (themes and sub-themes) as defined by the SCoSAS to study the interaction behaviours between the participants.

To reiterate, our lab-based study (Chapter 5) had participant pairs where one participant acted as a Seeker and the other as an Intermediary. The participants performed searches to satisfy three different information needs that we provided. We filmed the interactions between the two participants and our analysis is performed on the transcriptions of the search interactions. In this chapter, we use non-parametric tests to investigate differences in interaction behaviours such as time on task, total number of interactions, or total number of information requests.

8.1.1 Research Questions

We aim to investigate the interaction behaviours of our study participants in the SCS-data. We focus on search tasks with different levels of cognitive complexity. We study the effect of these tasks on the participants' interactions with each other while trying to satisfy their information need. Previous work has suggested that complex tasks take longer to complete (time on task), require more queries, and more search results/documents are inspected [108, 218]. It has also been suggested that task complexity affects discourse functions [59, 153]. We now examine task complexity in the context of SCS with our research questions as shown in Table 8.1.

TABLE 8.1: Research questions and hypotheses.

	Research Question	Hypothesis
1	How does task complexity affect search interactions in SCS?	People will interact more with search-related interactions when conducting complex tasks.
2	How does task complexity affect the use of discourse interactions in SCS?	People will use discourse more when conducting complex tasks.

8.2 Methods

We conducted an observational within-subjects study with 13 participant pairs. We provided Seekers with a short information need as backstory which was based on the cognitive complexity adopted from the Taxonomy of Learning [9] (see Section 5.5.1). Seekers verbalised their information need to the Intermediaries who had access to a search engine. Intermediaries then helped the Seekers satisfy their information need by communicating found information as explained in Section 5.3. Our observational study was conducted to understand the possible interactions in SCS which we identified in Chapter 7. We now investigate these identified themes and sub-themes in regards to their interactivity and frequency usage. We analyse the two main themes, the five most used bigrams, and the three most used sub-themes with statistical tests to understand the significant differences for different task complexities.

Next we outline the measured interaction behaviours.

8.2.1 Interaction Behaviours

We derive and measure the following interaction behaviours on different levels: *General interaction behaviours* by time on task and turns per task which are directly observed in the SCSdata; *Theme interaction behaviours* by the two identified themes (Task and

Discourse Level which cover 89.36% of the full dataset) and bigram interactions between themes; *Sub-theme interaction behaviours* with the three-most used sub-themes (Information Request, Results Presentation, and Discourse Management which cover 79.88% of the dataset) as identified in Chapter 7; and in *Lexical* level we investigate the number of one-word turns. All measures were computed at a session level.

TABLE 8.2: Interaction behaviour measures.

Level	Measure	Definition
Interaction behaviour	Time on Task	The amount of time in seconds participant pairs spent completing the search task. A maximum of 10 minutes per task was imposed.
Interaction behaviour	Turns per Task	The total number of interactions between the two participants when completing the search task.
Themes	Number of Task Level [°]	The total number of Task Level utterances of the two participants when completing the search task.
Themes	Number of Discourse Level [°]	The total number of Discourse Level utterances of the two participants when completing the search task.
Themes	Bigram interactions	The total number of bigram interaction chains on Theme Level between the two participants when completing the search task.
Sub-themes	Number of Information Requests [°]	The total number of Information Requests of the two participants when completing the search task.
Sub-themes	Number of Results Presentation [°]	The total number of Results Presentation of the Intermediary when completing the search task.
Sub-themes	Number of Discourse Management [°]	The total number of Discourse Management utterances of the two participants when completing the search task.
Lexical	Number of one-word turns	The total number of one-word utterances of the two participants when completing the search task.

NOTE: [°] Only utterances with single actions were used in this analysis. Multiple actions in one utterance were not included (See Section 5.8.2).

8.3 Results

We first present an overview of SCSdata descriptive statistics, number of words per turn, and one-word turns. We then answer the research questions relating to task complexity for search and discourse interactions.

8.3.1 Overall SCSdata Statistics

The SCSdata consists of 1044 turns between the 13 pairs of participants. Seekers took a total of 528 turns and Intermediaries took 516 turns. The observed discrepancy of 12 turns between the Seeker and Intermediary is because Seekers need to instigate the search and are the only actor who can conclude the search unless the 10-minute time limit is reached. An average of 80.30 turns per pairs and 26.76 turns per task were recorded (minimum turns per pair is two and maximum is 69).

The fill-word “uhm” was removed for analysis purposes. However, we deliberately did not remove any errors, false starts, or confirmations since these will likely occur in real case voice search scenarios.

8.3.1.1 Utterance Length

Participants exchanged 15.82 words per utterance (i.e., length of utterance) on average with a minimum of one word per turn and a maximum of 359 words per turn. Following stopword removal, the average length is 9.34 words (minimum length is zero and maximum is 219).² The number of words per turn for both actors is presented on log scale in Figure 8.1.

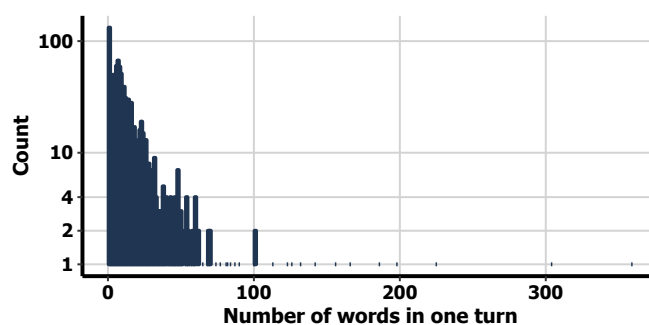


FIGURE 8.1: Number of words per turn for both actors ($N = 1044$).

8.3.1.2 One-word Turns

The dataset consists of 132 (12.64%) one-word utterances. These are utterances such as “yeah” or “OK”, for example,

P2 -INTERMEDIARY: How many people get caught at airport security checks yeah?

P1 -SEEKER: Yeah

[one-word turn]

One-word turns were more frequently produced by Seekers (82) than by Intermediaries (50).

A total of 127 (12.16%) of all turns in the dataset consisted of one word and were situated in the Discourse Level theme. All of these one-word Discourse Level utterances were located under the Discourse Management sub-theme.

²We used the SMART stopword list.

The five remaining one-word utterances in the dataset (0.47%) were located in the Task Level theme. Two of these utterances were found in the Results Presentation sub-theme and three in the Information Request sub-theme.

8.3.2 Data Analyses

A visual inspection of boxplots for all interaction behaviour measures showed that the data were not normally distributed and therefore one-way chi-square tests were used to investigate statistical significance ($\alpha = .05$) [74]. Bonferroni adjusted α -level (.017) was used for all post-hoc analyses [93].

8.3.2.1 Task Complexity and Search Interactions

Table 8.3 shows interaction behaviours per task complexity (Remember, Understand, and Analyse). A statistically significant difference for time on task over the three different task complexities was found, $\chi^2(2, N = 39) = 75.52, p < .01$. Post-hoc analyses revealed a statistically significant difference between the Remember and Understand task complexities, $\chi^2(1, N = 26) = 55.35, p < .017$ and Remember and Analyse task complexities, $\chi^2(1, N = 26) = 68.15, p < .017$. No statistically significant difference between the Understand and Analysis task complexities was found.

TABLE 8.3: Interaction behaviours per task complexity.

	Time on task*	Turns per task*	No. of Task Levels*	No. of Information Requests*	No. of Results Presentations
Remember	237 sec	248	158	85	67
Understand	429 sec	352	202	106	85
Analyse	454 sec	444	237	135	95

NOTE: * Statistically significant difference.

There was a statistically significant difference in the number of turns taken over the three different task complexities, $\chi^2(2, N = 1044) = 55.26, p < .01$. Post-hoc analyses revealed a statistically significant difference between all task complexities (Remember and Understand, $\chi^2(1, N = 600) = 18.03, p < .017$, Remember and Analyse, $\chi^2(1, N = 692) = 55.51, p < .017$, and Understand and Analyse, $\chi^2(1, N = 796) = 10.63, p < .017$).

A statistically significant difference in the number of Task Level turns taken over the three different task complexities was revealed, $\chi^2(2, N = 597) = 15.75, p < .01$. Post-hoc analyses revealed a statistically significant difference between the Remember and Analyse task complexities, $\chi^2(1, N = 395) = 15.8, p < .017$. No statistically

significant difference was found for the task complexities Remember and Understand, and Understand and Analyse.

A statistically significant difference was found in the number of Information Request utterances over the three different task complexities, $\chi^2(2, N = 326) = 11.60, p < .01$. Post-hoc analyses revealed a statistically significant difference between the Remember and Analyse task complexity, $\chi^2(1, N = 220) = 11.36, p < .017$. No statistically significant difference was found between Remember and Understand, and Understand and Analyse task complexities. A total of 31.22% (326) turns in the whole corpus are classified as Information Requests. The average Information Request length was 12.7 words ($SD=9.88, \text{min}=1, \text{max}=69$).

No statistically significant difference was found in the number of utterances as Results Presentation over the three different task complexities, $\chi^2(2, N = 247) = 4.89, p = .09$.

8.3.2.2 Task Complexity and Discourse Utterances

Table 8.4 shows the discourse behaviours per task complexity. A statistically significant difference in the number of Discourse Level turns taken over the three different task complexities was found, $\chi^2(2, N = 336) = 34.62, p < .01$. Post-hoc analyses revealed a statistically significant difference between all task complexities. With the Remember and Understand task complexities, $\chi^2(1, N = 181) = 12.2, p < .017$. The Remember and Analyse task complexities, $\chi^2(1, N = 222) = 34.88, p < .017$, and Understand and Analyse task complexities, $\chi^2(1, N = 269) = 6.25, p < .017$.

TABLE 8.4: Discourse behaviours per task complexity.

	No. of Discourse Level turns*	No. of Discourse Management turns*	No. of one-word turns*
Remember	67	58	29
Understand	114	86	35
Analyse	155	119	68

NOTE: * Statistically significant difference.

A statistically significant difference was found in the number of turns taken in the Discourse Management sub-theme over the three different task complexities, $\chi^2(2, N = 261) = 22.83, p < .01$. Post-hoc analyses revealed a statistically significant difference between the Remember and Understand task complexities, $\chi^2(1, N = 142) = 6.34, p < .017$ and Remember and Analyse task complexities, $\chi^2(1, N = 175) = 22.68, p < .017$. The remaining task complexity pair Understand and Analyse was not statistically significant different.

A statistically significant difference was revealed in the number of turns taken over the three different task complexities, $\chi^2(2, N = 132) = 20.04, p < .01$. Post-hoc analyses

revealed a statistically significant difference between the Remember and Analyse task complexities, $\chi^2(1, N = 97) = 15.68, p < .017$, and Understand and Analyse task complexities, $\chi^2(1, N = 103) = 10.57, p < .017$. No statistically significant difference was found between the Remember and Understand task complexities.

8.3.2.3 Bigram Interactions

We now investigate interaction bigrams at Theme Level (i.e., Task and Discourse Level) by task complexity. We take the top-five interaction bigrams which cover 842 turns (80.65%) in the SCSdata. Thus we look at the frequency of two interactions appearing in sequence. For example, the utterance below is an utterance classified as Task→Task, since utterance 1 and 2 belong to Task Level.

- (1) P13 -SEEKER: So which state in Australia consumes the most alcohol per person?
[Task Level]
- (2) P14 -INTERMEDIARY: Again 2016 or the most recent information?
[Task Level]

As shown in Table 8.5, no statistically significant difference was found in the number of chains of Task Level→Task Level turns taken over task complexity, $\chi^2(2, N = 317) = 2.83, p = .24$.

TABLE 8.5: Interaction bigrams for task complexity.

	Task →Task	Task →Discourse*	Discourse →Task*	Discourse →Discourse*	Task →Task+Task
Remember	95 (29.97%)	40 (21.28%)	34 (17.78%)	24 (21.43%)	8 (17.78%)
Understand	103 (32.49%)	65 (34.57%)	67 (37.22%)	36 (32.14%)	19 (42.22%)
Analyse	119 (37.54%)	83 (42.02%)	79 (43.89%)	52 (46.43%)	18 (40%)

NOTE: * Statistically significant difference.

A statistically significant difference was revealed in the frequencies of Task Level→Discourse Level chains over the three different task complexities, $\chi^2(2, N = 188) = 14.88, p < .01$. Post-hoc analyses revealed a statistically significant difference between the Remember and Understand task complexities, $\chi^2(1, N = 105) = 5.95, p < .017$ and Remember and Analyse task complexities, $\chi^2(1, N = 123) = 15.03, p < .017$. No statistically significant difference was found for the Understand and Analyse task complexities.

A statistically significant difference was revealed in the frequencies of Discourse Level→Task Level chains over the three different task complexities, $\chi^2(2, N = 180) = 18.1,$

$p < .01$. Post-hoc analyses revealed a statistically significant difference between the Remember and Understand task complexities, $\chi^2(1, N = 101) = 10.78, p < .017$ and Remember and Analyse task complexities, $\chi^2(1, N = 113) = 17.92, p < .017$. No statistically significant difference was found for the Understand and Analyse task complexities.

A statistically significant difference was revealed in the frequencies of Discourse Level \rightarrow Discourse Level over the three different task complexities, $\chi^2(2, N = 112) = 10.57, p < .01$. Post-hoc analyses revealed a statistically significant difference between the Remember and Analyse task complexities, $\chi^2(1, N = 76) = 10.31, p < .017$. No statistically significant difference was found for Remember and Understand, and Understand and Analyse task complexities.

No statistically significant difference was revealed in the number of chains of Task Level \rightarrow Task Level + Task Level turns taken and task complexity, $\chi^2(2, N = 45) = 4.93, p = .08$.

8.4 Discussion

This analysis aimed to examine interaction behaviour over tasks with different task complexities in SCS. We investigated the interaction patterns and frequencies of participants throughout their search process including search and discourse frequencies.

The first hypothesis was supported, and participants of our study interacted more when they engaged in complex search tasks. Results showed when completing tasks of different levels of cognitive complexity, participants had a significantly different number of interactions (overall and at Task Level), they spent more time on task, and they posed a higher number of information requests. Even though there was not always a significant difference detected in search behaviours between the mid-level cognitive complexity (i.e., from Remember versus Understand or Understand versus Analyse), we did show significant differences between the more extreme cognitive complex tasks (Remember versus Analyse). Our results are consistent with findings from Arguello et al. [11], Jansen et al. [97], Kelly et al. [108], and Liu et al. [125] where in general the number of interactions increased with greater cognitive complexity.

Concerning the task complexity and discourse utterances, the data support our second hypothesis, and more discourse interactions were observed in complex tasks. We showed that participants interacted more on Discourse Level when the cognitive complexity increased. When we investigated these discourse interactions, we found that Discourse Management utterances and one-word turns are more frequently used with the increase

of task complexity. In addition, an inspection of one-word turns revealed that these utterances are often used as confirmation actions, such as “yeah” or “OK”. We again found significant differences in interaction behaviours between the more extreme cognitive levels (Remember versus Analyse). Thus, our results support the suggestion that task complexity affects discourse interaction behaviour [59, 79]. Furthermore, our results also support that more discourse functions, in particular, meta-communication and confirm utterances, are used when the task complexity increases [59, 151, 153].

Finally, we investigated the differences in bigram interaction behaviour and task complexity. When we investigated the most frequent bigram interactions, we found no difference in the number of interactions between Task Level→Task Level interactions. However, a difference was found when bigram interactions involved Discourse Level utterances. We see that Discourse bigrams are used more frequently for more complex tasks. We speculate that when the task complexity increases, the cognitive resources of the Seeker are stretched and therefore communication failure is more frequent [79]. Furthermore, these discourse actions are crucial to overcome the imposed difficulty of the audio-only channel, not to mention the disruptions imposed by the task complexity in the communication flow between Seeker and Intermediary. Further investigation is needed to understand how this observed behaviour is related to users experiencing difficulties throughout their search process (i.e., search struggle) [16, 89, 140].

A limitation of our study is the small sample size which meant we were unable to perform data transformations (i.e., to normalise the data) [74]. Future work could expand our study to a larger pool of participants which may allow for more powerful analyses such as ANOVAs.

8.5 Conclusion

To the best of our knowledge, we are the first to examine task complexity in an audio-only search environment. We explored the relationship between different task complexities and interaction behaviours in SCS. Following previous research, we show that more complex queries relate to higher interaction counts (e.g., a higher number of turns, longer sessions, more information requests, and more results presentations) [11, 108, 218]. We also found that more complex tasks exhibited greater support utterances through meta-communication, such as Discourse Management (i.e., confirmations). Discourse interactions in bigrams were also further frequently found in more complex tasks.

Thus it appears that task complexity has an effect on interaction and discourse behaviour in SCS. One could speculate that the audio channel poses information transfer restrictions. Due to these restrictions, more complex tasks show further discussion to repair or confirm the information chain. We suggest that this increase in discourse interactions signifies the complexity increase that the audio-only channel imposes.

8.6 Chapter Summary

In this chapter, we investigated the interaction behaviours of participants in our observational study concerning task complexity, search interaction, and discourse utterances. We studied the interaction patterns of variables such as time on task, turns per task, number of Task and Discourse Level interactions, bigram interactions, amount of Information Requests, Results Presentations, or Discourse Management, and frequency of one-word turns. Our results showed that interactivity and discourse utterances, in particular meta-discourse, increased as tasks became more complex.

Overall, these results contribute to our suggestion that SCS likely involves greater complexity than the current browser-based search (see Chapter 6). The results in this chapter emphasise the need for further research in the usage and nature of meta-discourse functions in SCS and their role in search tasks and information seeking.

Part IV

Discussion

Chapter 9

Recommendations for the Design of Spoken Conversational Search Systems

In this chapter we bring together and discuss the results from Chapters 3–8 by triangulation.¹ That is, we combine the analyses and results from our multiple methods and datasets in this thesis to explore the new interaction paradigm of SCS. Each data source and method examines the SCS paradigm from a different angle to reduce deficiencies caused by only using one dataset or investigation method [154]. Thus, we aim to aggregate our analyses and results while reviewing these with more in-depth argumentation.

Firstly, we present practical outcomes. We introduce ten SCS design recommendations and discuss these design recommendations using insights from the SCoSAS annotation schema and SCSdata, RealSAM, and MISC datasets. Secondly, we suggest a schematic SCS model, based on the SCoSAS and SCSdata, enabling the evaluation of SCS processes against existing information seeking models. This examination demonstrates that most existing models insufficiently capture the system’s role as an active and responsible participant in the information seeking conversation. It also demonstrates the lack of discourse functions (i.e., meta-communication), particularly to overcome discourse errors or communication breakdowns in systems at present. Furthermore, our conceptual contributions suggest that SCS systems are complex, and need to be interactive and

¹This chapter consists of the following publications J. R. Trippas, D. Spina, L. Cavedon, H. Joho, and M. Sanderson. Informing the design of spoken conversational search. In *Proceedings of Conference on Information Interaction and Retrieval (CHIIR)*, pages 32–41, 2018 and J. R. Trippas, D. Spina, P. Thomas, H. Joho, M. Sanderson, and L. Cavedon. Towards a model for spoken conversational search. *Information Processing & Management*, 2019. (Submitted).

pro-active. Finally, we revisit and expand the SCS requirements presented in Chapter 2. The research and development of genuinely communicative SCS systems are still in the early stages and our practical, conceptual, and methodological contributions offer insights for future research.

This chapter is structured as follows. In Section 9.1, we discuss our design recommendations in relation to results in Chapters 3–8 and provide practical suggestions. Section 9.2 covers conceptual outcomes: the first schematic model of SCS in Section 9.2.1; theoretical implications on the interaction style in Section 9.2.2; and the evaluation of our schematic model against existing models in Section 9.2.3. We then redefine SCS requirements in Section 9.3. Finally, we present the chapter summary in Section 9.4.

9.1 SCS Design Recommendations

In this section, we present ten design recommendations for SCS systems to support natural user interactions. The recommendations follow the overall SCoSAS structure of Task and Discourse level including two other recommendations promoting effortless information-engagement beyond controlled user-system interactions (see Table 9.1). An explanation of each recommendation and a summary of how to address that recommendation are given. We suggest the design criteria with reference to the results from Chapters 3–8.

TABLE 9.1: SCS Design Recommendations (DR).

DR	Level	SCS Design Recommendations
1	Task	Be adaptive to accept information requests
2		Present relevant information and support flexible results exploration
3		Pro-actively provide search assistance
4		Accept search progression insights from users for personalisation and contextualisation
5	Discourse	Use discourse markers to improve communication
6		Exploit grounding and other dialogue dynamics
7		Support multi-dimensional navigation
8		Be able to communicate the system's state to the user
9		Advance beyond one-action paradigms
10		Support processes of information use outside the system

9.1.1 Task Level Design Recommendations

Be adaptive to accept information requests. (DR1)

The analysis of the SCSdata suggests that information requests were formed in many different ways (Section 6.1.1), from natural language expressions to detailed and carefully

crafted requests. Furthermore, how and when these requests were expressed varied. For example, Seekers were able to pose conditional information requests (i.e., if the search results are valid to a specific condition, then they used that result for further searches) as specified in Section 6.1.3. Other examples include gradual revealing of information needs over multiple turns (Section 6.1.3), searches being conducted within or about a document (Sections 6.1.3 and 7.2.1), and expressions of Seekers' uncertainty of how to go about a search (Section 7.4.6). The Intermediaries are also involved in eliciting or refining the information requests. This includes posing information request suggestions and fully developing the requests after extracting/obtaining the user's information need (Section 7.2.1).

We suggest that formulations of information needs may not always conform to the browser-based query since users can express their information need more naturally through speech (i.e., without formulating it as they would submit it to a browser-based search box). In a voice environment, users can use natural language to describe their search, which may contribute to the longer and more verbose information requests [85]. Furthermore, we observed Seekers disclosing that they had identified a knowledge gap with a need to specify it, but the information need was not yet formalised similar to the ASK hypotheses (Section 7.4.6) [22]. Thus, the information request may not always go through Taylor's four stages of information need (i.e., visceral, conscious, formalised, and compromised) before it is expressed [179]. The verbosity and ill-formed information requests together with the lack of one-word information requests strongly suggest that users find it more intuitive to express their information need in longer queries. We suggest that a SCS system needs to anticipate a range of information need expressions and be adaptive to support the users' different and changeable information requests within one session.

Ellis suggested that information requests may poorly express the users' underlying knowledge gap [73]. We extend this to the audio-sphere, by suggesting that users can first formulate vague information requests to the system, then navigate to a document and pose more specific information requests within documents (Section 6.1.3). Further investigation is needed to understand how to respond to the variety of these information requests and best support the elicitation of the user's need.

To summarise:

- DR1.1: Support a range of different (ill-formed) information need expressions;
- DR1.2: Anticipate for information need expressions to be given at any time;
- DR1.3: Allow for gradual discovery of the user's information need (over multiple turns or sessions);
- DR1.4: Allow for searching within content.

Present relevant information and support flexible results exploration. (DR2)

We observed a variety of results presentation techniques from Intermediaries in the SCS-data, such as summarising, comparing, or synthesising documents and SERPs into “information units” (Section 7.2.1). The category Reveal reported by Azzopardi et al. [19] has similar functions (i.e., summarise, compare). Besides conveying text information, Intermediaries also interpreted visual information for Seekers such as graphs, photos, or changes in link colour (i.e., already clicked), as discussed in Section 6.1.2. Such visual material will benefit from better descriptive information, enabling the full potential of audio-only interaction systems [1] which could be achieved through image description generation [104].

We suggested that different kinds of queries may benefit from an optimised summary (single-facet versus multi-facet queries) in Chapter 4. We indicated that query words should be put in the context of the found document, thus reflecting their relationship with the underlying document, which is in line with recommendations by Clarke et al. [54] (Section 3.3). For audio recordings, such as music, lecture recordings, or podcasts, this may mean that users listen to a snippet extracted from the podcast audio to understand the context of their query word [171]. Further investigation is needed to understand which kind of presentation (i.e., multi-document summarisation or comparing results against each other) and which interaction techniques (i.e., combination of results presentations) are suitable for different types of information need and query, or different contexts.

The SCSdata suggests that the boundaries between the kind of documents (e.g., news articles, blogs, or general web pages) is becoming undistinguishable (Section 6.1.2). This indicates that boundaries between SERPs and documents are not detectable in audio without creating modifications [51]. Furthermore, the credibility of documents or information units can be assessed quickly in a visual setting; however, this multi-dimensional credibility assessment is not easily accessible in an audio-only environment (Section 7.2.1). Thus, transparency is needed to indicate from where the information was extracted.

To summarise:

- DR2.1: Create adaptable results presentation styles in different contexts and information needs;
- DR2.2: Support non-text information interpretation (e.g., graphs and photos);
- DR2.3: Present information in the context of the document to reiterate the relationship;
- DR2.4: Be transparent as to which sources the information comes from if necessary.

Pro-actively provide search assistance. (DR3)

Intermediaries assisted Seekers throughout their search process, from providing specific search suggestions to requesting spelling (Sections 7.2.1 and 7.4.6). Seekers also explicitly asked for search assistance throughout the search session (Section 7.2.1). We suggested none of these search assistance techniques to participants. However, we believe that they were intuitively applied to overcome the challenge of transferring information through an audio-only communication channel. It is suggested that these techniques may be suitable for future SCS systems to adopt.

Many different techniques exist to unobtrusively integrate search assistance in browser-based search systems, for example, query suggestion during the query formulation stage or spelling suggestion after the query formulation. In our results, we extend these search assistance functions by including that the system: (1) provide and ask document relevance feedback or usefulness judgements to/from the user, and (2) suggest continuing from an information space (e.g., progressing to a new topic).

To summarise:

- DR3.1: Pro-actively provide assistance to the searcher;
- DR3.2: Accept and utilise relevance feedback from the user;
- DR3.3: Elicit relevance feedback from the user reasonably;
- DR3.4: Suggest to “move on” to a different information space when the topic has been exhausted or does not contribute to satisfying the information need.

Accept search progression insights from users for personalisation and contextualisation. (DR4)

The results indicate that Seekers actively shared their progression of the search task (Section 7.2.1). They provided this progress insight by specifying performance feedback (i.e., how the search is progressing), rejecting search results, or notifying the Intermediary if they had gathered enough information.

Seekers were not forced in any way to produce relevance feedback in our study (i.e., indicating the positive or negative relevance of a document or proposed query-reformulation); however, they offered it nonetheless. We suggest that SCS systems incorporate such feedback which can help with the personalisation of the system for the user and may lead to better system performance. The relevance feedback can also be used by the system to further contextualise the users’ information requests [62, 95, 182].

To summarise:

- DR4.1: Include user performance feedback into the search model;
- DR4.2: Allow for negative/positive relevance feedback as rejected or accepted results.

9.1.2 Discourse Level Design Recommendations

Use discourse markers to improve communication. (DR5)

Our results suggest that discourse markers (e.g., utterances which are concerned with the conversational coherence and cohesion between participants while dealing with conversational repairs) are a crucial component of SCS (Sections 7.2.2 and 8.3.2.2). Indeed, miscommunication, repairs of the miscommunication, or solving speech disambiguation (e.g., by confirmations) are frequent occurrences in speech [111, 132, 224]. These discourse utterances invariably help overcome difficulties imposed by the audio-only interaction channel [79]. The discourse markers and meta-communication may occur more frequently in a spoken setting versus a web-based search due to the temporal nature of speech, and are not part of the primary information seeking web-based actions (i.e., query, document recommendations, or item selection). However, the particular discourse utterances may become the fundamental support-actions for these primary information seeking steps.

Handling errors or miscommunications through dialogue directly may cultivate a more user-friendly or human-like conversation. Much research has been conducted in information transfer [165], conversational repair [162], or miscommunication [169] in verbal communication. This includes investigations of how conversational repair is organised, which possibilities of repair exist, and how to deal with miscommunications which are only realised later in the conversation. All the differences in these discourse management functions and comprehension of conversational repair need to be addressed for implementation in SCS.

Furthermore, we observed an increase in discourse interactions and meta-communication in more complex tasks, suggesting that task complexity affects SCS interaction behaviour (Sections 8.3.2.2 and 8.3.2.3). This is in line with previous research [59, 151]. We believe that such discourse interactions are vital in dealing with disruptions imposed by complex tasks which also interrupt the communication flow. Additionally, we suggest using discourse together with a measure of effective information transfer [165] to be included in the prospective evaluation of a SCS system.

To summarise:

- DR5.1: Express discourse markers to indicate problems such as miscommunication, uncertainty, or vagueness;
- DR5.2: Include the users' (meta-)discourse markers and uncertainty of effective information transfer expressions as an evaluation measure.

Exploit grounding and other dialogue dynamics. (DR6)

Grounding (i.e., discourse for the creation of mutual knowledge and beliefs) is when participants in a conversation engage in a specific discourse activity to share their mutually understood utterances [52]. We observed grounding actions in the SCSdata (Section 7.2.2). For example, Seekers provided indirect feedback by reciting their interpretation of the found results. This grounding process could enable a future SCS system to better understand a user's awareness of the results or information space, including helping the SCS system to disambiguate a users' information need. In particular, users' grounding utterances can be incorporated as a SCS feedback feature as investigated in SDS with research in Information State Update (ISU) [188]. ISU researchers attempted to identify these grounding utterances by characterising the dynamics of a dialogue [80, 190]. The ISU symbolises what is known at a given moment in a dialogue and can consists of two parts, the (1) *mental (or internal) states of the user* and (2) *information about the dialogue*. The user's mental state component includes the user's beliefs, obligations, intentions, commitments, or desires. The dialogue information component collects which utterances have been said, which dialogue moves were generated, and if the information was shared. The *information about the dialogue* component has recently been described by Radlinski and Craswell's memory model for conversational search [146]. They proposed that memory of the system has two specific roles; firstly, by recalling what has been said earlier in the conversation (this includes the information need) and secondly, by referencing explicitly to what has been said such as clarifications "What I meant with that...". Hence, Radlinski and Craswell's memory function can keep track of the conversation's context [146].

We suggest the implementation of a combined ISU and conversational search memory model and add two more components to form the SCS grounding model or *UMII*: (3) *interaction preferences* and (4) *information space coverage* (see Table 9.2) as described in Sections 4.4 and 6.1.2. Indeed, a system should adapt to the conversational style and preferences (e.g., search assistance preferences) of the user in their context and their given task. Imagine a user is read out confidential information in a public space. The system should have presented the results in the preferred mode in the context of the user. We also recommend including the *information space coverage*, keeping track of the materials already covered and the users' mental model of the information space (i.e., what the users believe or understand is part of the existing information) as discussed in Section 6.2.2. Finally, not all users provide grounding utterances; thus, a system cannot rely entirely on this measure and it needs to be dynamically updated throughout the search interactions.

TABLE 9.2: SCS grounding model components (or UMII).

	SCS	ISU [190]	Conversational Search [146]
1	User's mental states	User's mental states	
2	Memory	Information about dialogue	Memory
3	Interaction preference		
4	Information space coverage		

To summarise:

DR6.1: Use UMII to continuously update mental states, what has been said, information space coverage, and interaction preferences;

DR6.2: Do not rely exclusively on all aspects of UMII.

Support multi-dimensional navigation. (DR7)

The SCSdata suggests that users may navigate through non-linear navigational interactions (Section 7.2.2). That is, SCS users can express a “back-button click” in many different ways, do not need to navigate physically through documents or search system in a linear fashion, and can skip navigational steps altogether by simply referencing an item or document. This non-linear SCS navigational behaviour is in contrast to web-based information seeking patterns where the results page returned by the search engine is often seen as the central hub from which users explore documents (i.e., hub-and-spoke user interface design pattern) [98]. Furthermore, instead of interacting with lists in a spoken environment, as often done in a SDS or in the case of RealSAM, users can freely navigate in a multi-dimensional information space in SCS. This unlimited SCS navigational experience is due to the liberation of rigid web-based navigation. Nevertheless, lists can still be a (back-up) results presentation strategy. In the event that lists are needed, we suggest that techniques such as “infinite-reading” mode (i.e., seamlessly navigating) are implemented to mitigate interruptions in the output and the need for the user to repeat commands (Section 3.3). Other navigational support could be included through sonification of clustered search results to indicate proximity or similarity through sound-features (i.e., changing the pitch as the information space and orientation develop).

In future SCS systems, with the flexibility of navigation, we believe keeping navigational steps accessible and traceable to refer back to will be helpful for users [19]. Being able to present a traceable history also provides further transparency for the user and supports the explainability of the system. One could investigate the use of “breadcrumbs” to contribute to the users’ location awareness, the current document or information space,

and their interaction path. For example, breadcrumbs could refer to previous information spaces or provide summaries of information the user visited instead of titles of documents as in a browser-based back-button action.

To summarise:

DR7.1: Keep navigational steps at hand for traceback and explainability;

DR7.2: Avoid reading out lists, but when necessary implement techniques such as “infinite-reading” or sonification of the information space.

Be able to communicate the system’s state to the user. (DR8)

The SCSdata suggests that a system should be able to indicate which processes are happening inside the system through visibility of system status [136] so the user understands what is happening (Section 7.2.2), for example, when the system is listening or processing information. Visibility of system status also enables greater control, explainability, and transparency of the system processes and outputs [62]. At any point in time, the SCS system should be able to disclose its state to the user. For example, it should disclose how the system retrieved or computed specific information which contributes to the interaction process, or respond when a user wants to understand why particular results have been presented. This information may be stored in the system’s memory from past preferences (as discussed in DR6) or communication with the user, and it should be able to demonstrate where it was extracted from.

However, providing constant feedback on what is happening in a system may not be convenient in a spoken environment and could overload the user with too much (unnecessary) information. Instead, understanding which aspects should be given and in which mode (i.e., audio or screen based) to the user may be essential in the usability assessment of a SCS system.

To summarise:

DR8.1: Be ready to disclose and explain the steps or processes the system took.

Advance beyond one-action paradigms. (DR9)

On another practical note, we argue that progressing beyond the one-action search paradigm (action-response) is necessary for a user-friendly system (Section 6.2.1). We suggested that the *naturalness of the interaction* with a SCS can be an evaluation feature as in SDS [124, 208]. We recommend that one of the aspects of this measure could be users uttering multiple moves in one turn (i.e., one user-move can consist of a navigational command and feedback request). In a human–human interaction, this behaviour

is observed and expected, and the other actor can handle it. Therefore, allowing users to utter multiple moves in one turn which the system can process is likely to lead to positive interactions with the system. Recently naturalness was also proposed as an evaluation measure in TREC Conversational Assistance Track (CAST) [64].

Support process of information use outside the system. (DR10)

Finally, we also advocate that the SCS system actively supports the user to process the found information and usage beyond the system (Section 6.1.4). We suggest aiding the user in manipulating, integrating, or utilising the found information in their (physical) world [19, 126]. For example, by note-taking or summarising all the found information and present it on a desktop device.

9.2 Towards Models and Detectable Components of SCS

In this section, we present a visual overview of our annotation schema and all its components (i.e., themes and sub-themes). We then provide possible avenues for further research on how to extend the annotation schema. We finish this section by discussing existing search models considering our annotation schema.

9.2.1 Schematic SCS Themes Model

We present a nested schematic overview of our observed SCS interactions which is derived from the SCoSAS annotation schema built on the SCSdata (Figure 9.1). This schema presents the Task Level as the centre of the conversations with the utterances regarding the topical search task. The Discourse Level is positioned around the Task Level representing the statements which are about the mechanism (i.e., the function, not the task). Thus the Discourse Level would still exist if the search task is changed to a different task other than search. The figure also demonstrates which sub-theme is accessed by each actor.

Previous research in *communication goal studies* suggested a similar two-tiered model as our proposed schema [187]. Furthermore, the goal studies community argues that ordinary discourse is segmented in different types of goals such as communicative functions or interaction outcomes which is similar to our two themes of Task and Discourse. Bunt provided a two-tiered model where general information dialogues consist of two motivations, that is, one tier was concerned about the task communication and the second

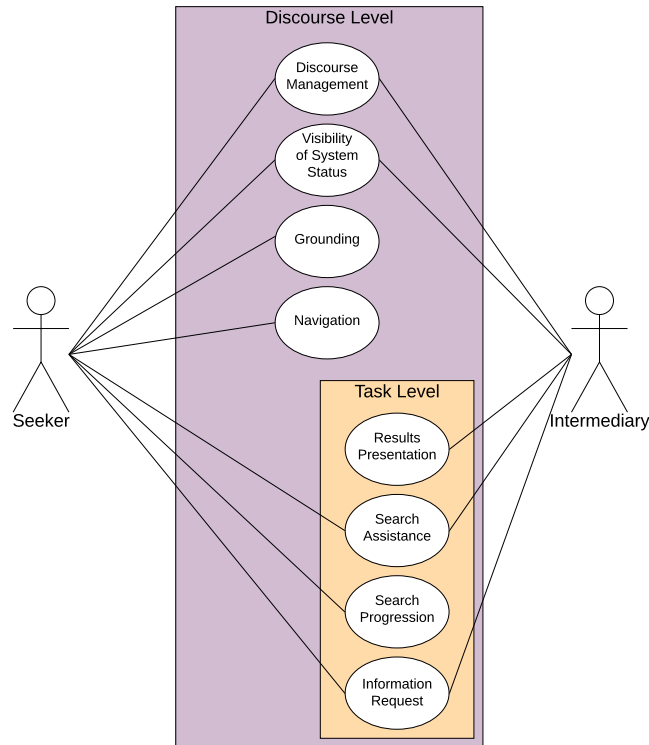


FIGURE 9.1: Schematic model of SCoSAS themes and sub-themes.

with driving the conversation [40]. Both two-tiered models strengthen our findings and our multi-level schema classification.

This is the first attempt to create an interaction model of two actors in a SCS setting. Further refinements of the model are not excluded. For example, possible extensions to the schematic model could include System Level functions such as user help functions, device functions, or personalisation functions. All interactions related to user guides, settings (i.e., WiFi, battery, or personalisation, Section 3.2.2), and discovering which device functions are available could potentially be covered in this System Level. The inclusion of this theme would not interfere with the existing themes as shown in Figure 9.2.

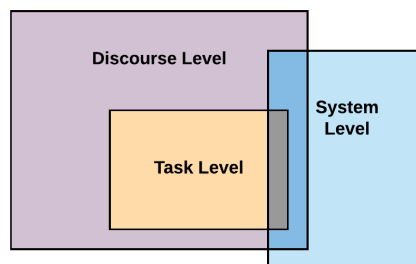


FIGURE 9.2: Possible schematic inclusion of System Level function.

9.2.2 Increased Complexity, Interactivity, and Pro-activity

SCS as a new interaction paradigm introduces opportunities. We have shown that changing the communication channel inherently introduces constraints to the search process. However, these difficulties can be alleviated with the use of conversations, interactivity, and pro-activity. Thus, our remaining conceptual implications for SCS research are the following: *(i)* increased system and interaction complexity, *(ii)* increased interactivity, and *(iii)* increased pro-activity (or agency) against existing browser-based systems. We have illustrated throughout our work that these components are intrinsic to SCS.

Increased system and interaction complexity. We illustrated the limitations of the restricted audio-only channel in this thesis: for example, by the range of different information requests, kinds of possible feedback loops, or different results presentation strategies (Chapter 7). These described complexities are based on observational interactions and do not include the conversational strategy management systems which need to be implemented [133]. Furthermore, our results suggest that systems should have more autonomy through SCS decision making (Section 6.2.3). Enabling decision making by the system increases the system–action possibilities and thus promotes the complexity. However, increasing system–actions leads to more complicated user- and system-models and expectations, including the users’ cognitive models [23], resulting in greater complexity.

Increased interactivity (collaboration). Our results show that interactivity and collaboration through dialogue are important for mitigating communication breakdowns in more complex tasks, as discussed in Section 8.4, and supports previous research [59, 151]. Furthermore, we illustrated the necessity of extra non-search interactions to update each other’s mental state via grounding (Section 7.2.2). Thus the development of SCS has to include all these possible interaction action-pairs. Future investigation of limitations, negative use-cases, contexts in which this system will be used, and asking the users for their needs of this system, will be invaluable.

Increased pro-activity or agency. The SCS system needs to be actively involved in the search process with the user to fulfil all SCS design recommendations as presented in Table 9.1. Furthermore, the audio-only interaction channel imposes limitations on the amount of information which can be transferred in one utterance (or turn) without cognitively overloading the user. This is in contrast to browser-based search, where all

information can be presented at once and where the user can determine which information is relevant to them at that moment. In SCS, the system must decide whether presenting information or providing assistance is worth the bandwidth cost, with a possibility of cognitively overloading the user (Section 7.2.1). The SCS system needs to take responsibility for assessing this cost–benefit determination.

Some examples of agency are: the system’s obligation to make decisions or cost–benefit analyses on how to best present search results for the user in their context at a given time (Section 6.2.3); suggesting relevant search assistance for a particular problem (Section 7.2.1). Thus, the system has to adapt, accommodate, and support the user so that the user has to expend as little effort as possible. This illustrates the pro-activity required and the system’s capacity to act independently. For example, future research can investigate how comfortable users are with offloading their decisions for particular tasks and in which contexts.

9.2.3 Evaluating Existing Search Behaviour Models with SCoSAS

One objective of our observational study was to explore whether any existing information seeking models fit SCS. However, to our knowledge many well-known models such as Belkin’s ASK [22] or Marchionini’s ISP [126] do not include the system’s “responsibility” of interacting with the user and thus do not capture all SCS behaviours.

Other models, such as Sitter and Stein’s COR model [168], Belkin et al.’s scripts [25], or the recently proposed QRFA model by Vakulenko et al. [205] encompass the interaction between two actors. However, these models either lack the flexibility of the speech aspect, such as multiple moves in one turn, or are based on broad DA categorisations. Furthermore, meta-discourse utterances are also lacking in those existing models, and these utterances appear to be a substantial aspect of SCS. It is important to include these discourse markers because incorporating them inherently creates a system which interacts in a mixed-initiative information seeking communication (the system can ask for clarification and thus takes initiative). Such mixed-initiative dialogue is a requirement of what makes a SCS system truly conversational. Additionally, the broad DA categorisation only provides a high level insight of the actions users take while the SCoSAS discloses more refined details of the users’ and systems’ state in each turn.

Finally, Saracevic’s stratified model includes the system as an active participant in the information seeking process [158]. Furthermore, Saracevic specifies that the process consists of a dialogue between the two actors. He also mentions that the dialogue can be used for not only “searching” utterances but also for a number of “other engagements” beyond the searching, for example, obtaining and providing different types of

feedback, judgements, or states. In the SCS model, we also identify the system as an active participant throughout the search process, which is in itself a conversation. In addition, the “other engagements” Saracevic mentions could be interpreted as our Discourse Level interactions, such as our identified grounding utterances. Furthermore, the stratified model could be used to illustrate the effect of the audio-only interaction channel limitation. That is, Saracevic says that a weak point in the system could hamper the desirable outcome for the search process [158]. The stratified model and the schematic SCS themes model may be complementary for the abstraction of a SCS process.

9.3 Expanding SCS Requirements

We outlined the requirements of a SCS system in Chapter 2. A SCS is concerned with dialogue-like information seeking exchanges through spoken language between users and system. The system is pro-actively involved with eliciting, displaying, and helping the user to satisfying their information need through multi-turn transactions which can be over multiple sessions (see Table 9.3).

TABLE 9.3: Predefined SCS requirements.

	SCS
1 <i>Analogy</i>	Human intelligible dialogue-like, beyond command and control
2 <i>Language</i>	Spoken natural language, conversational
3 <i>System participation</i>	Pro-active, mixed-initiative (implies listening)
4 <i>Information request length</i>	Longer, more natural
5 <i>Results presentation mechanism</i>	Adaptive to users' need and context (ranked list is inadequate)
6 <i>Turn-taking</i>	Multi-turn
7 <i>History</i>	Over (multiple) sessions

During our research we identified further suggested requirements. We add five new requirements: multi-moves (Section 6.2.1), errors (Section 7.2.2), turn-time (Section 8.3.2.1), semantics (Section 6.1.1), and navigation (Section 7.2.2) as presented in Table 9.4. We refine the requirements of SCS systems by: A SCS system supports the users' input which can include multiple actions in one utterance, is more semantically complex, and thus turn-time is less predictable. Moreover, the SCS system helps users navigate a non-linear information space and can overcome standstill-conversations due to errors or communication breakdown by including meta-communication as part of the interactions.

The methodologies in this thesis, namely the use of thematic analysis and crowdsourcing, have been described in detail [196, 198]. As such, they are readily replicable by future researchers. Indeed, the crowdsourcing framework has already been used for additional

TABLE 9.4: Redefined SCS requirements.

		SCS
1	<i>Multi-moves</i>	From user and system
2	<i>Errors</i>	Intelligent problem solving and anticipation of errors (through meta-communication)
3	<i>Turn-time</i>	Less predictable
4	<i>Semantics</i>	Complex, more discourse
5	<i>Navigation</i>	Multi-dimensional

purposes by others [51, 171]. The SCSdata created as a result of our methodology has also been utilised [205].

9.4 Chapter Summary

We discussed the analyses and results of Chapters 3–8. We combined the outcomes from those chapters and formed an in-depth discussion.

We provided practical contributions as design recommendations for SCS which were derived from the triangulation of results. These design recommendations were also discussed in more detail with further references for possible future investigation avenues. Our conceptual contributions are the first step towards a SCS model, including features of SCS, and a discussion of how SCS varies from existing search behaviour models. Finally, we proposed extensions to SCS requirements.

Chapter 10

Conclusion and Future Work

Voice search is increasingly used with the rise in popularity of a number of speech-based search applications. Within this area, voice-input has been previously researched but limited work has explored voice-output. That work has suggested that a number of difficulties may be found because of the limitations inherent in the narrow channel of speech. We believe that conversational interactions can alleviate some of these speech-imposed difficulties. Thus, the aim of this thesis was to explore SCS, and in particular, to examine *(i)* the interaction behaviours and *(ii)* the results presentation in SCS.

To explore the interaction behaviours, we started with a log analysis of an audio-only communication channel system, RealSAM, which is used for accessing media by people with a visual impairment. The RealSAM logs enabled us to conduct an initial exploration of interaction behaviours (Chapter 3). This interaction log analysis informed methodological decisions for the second step in our SCS exploration, the development of SCSdata in Chapter 5. SCSdata is a unique dataset with extensive documentation for reproducibility. We also developed the data analysis methodology for the SCSdata, including the annotation and validation processes. During the formalisation of the SCSdata, we started accumulating and assembling observations unique to the experiment and these are outlined in Chapter 6. We then identified and classified the atomic interactions observed in the SCSdata in Chapter 7. We derived the first annotation schema for SCS from these classifications which we called the SCoSAS. The SCoSAS was then thoroughly validated and shown to be generalisable and replicable for a SCS setting. Finally, we demonstrated the extensive use of the SCSdata and SCoSAS by further investigating task complexity, interaction, and discourse behaviour in Chapter 8.

To explore results presentation for SCS, we first created a crowdsourcing framework to investigate different results presentation strategies in an interactive environment (Chapter 4). The framework has been used by several other results presentation studies which

confirm the transferability of our setup.

Our results suggested that allowing users to express their information need verbally increases the complexity of SCS. We found that two major processes are involved in SCS, namely utterances which are either related to the task or related to the meta-discourse. Thus, it appears that different interaction behaviours occur in SCS which are not found in a browser-based search, such as asking for repetition (meta-discourse utterances) to resolve a communication breakdown. Our results also suggested that translating text to audio is insufficient to support users' needs. To address some of these challenges, design recommendations have been outlined in Chapter 9. This chapter also outlines conceptual and methodological suggestions to support future research.

The remainder of this chapter is divided into the following sections: in Section 10.1 we provide the summary of our contributions, and in Section 10.2 we state some extensions to our observational study and SCoSAS creation. Finally, in Section 10.3 we specify possible future experiments.

10.1 Summary of Contributions

We now provide a summary of the thesis contributions by chapter.

Part I – Thesis Overview and Background

Chapter 1 – Introduction: We outlined our motivations for this thesis and the research scope, as well as providing an overview of the challenges in SCS and our contributions.

Chapter 2 – Background: We provided background to this thesis including reviewing the development of Spoken Conversational Systems and SCS in particular. We outlined interactivity in IIR and the impact of task complexity and discourse. We reviewed information seeking processes and models including information seeking through dialogue. We discussed search actions through audio and review speech user interfaces. Finally, we provided background information on SDS.

Part II – User Preferences in Results Presentation and Access over an Audio-Only Communication Channel

Chapter 3 – Accessing Media Via an Audio-only Communication Channel:

- We illustrated the importance of thorough experiment setup and analysis protocol for future audio-only studies.
- We highlighted the need to meticulously log audio-only interactions, including where possible the preservation of the audio input and output for closer examination.
- We outlined the influence of pre-defined system categories on interaction behaviour in audio-only interactions.

Chapter 4 – Results Presentation for Audio-only Communication:

- We designed a novel crowdsourcing framework to investigate results presentation in an interactive audio-only communication setting.
- We provided further confirmation that text snippets cannot simply be translated into audio without consequences for user preference in an audio-only environment.
- We showed that different kinds of queries benefit from a different optimised summary.

Part III – Towards a New Model of Spoken Conversational Search

Chapter 5 – Methods:

- We proposed a methodology for creating a SCS dataset, including the data collection setup, questionnaires, semi-structured interviews, and transcription methodology.
- We introduced the data analysis methodology, annotation schema conception, and validation process.
- We created SCSdata.

Chapter 6 – Observing Spoken Conversational Search Interaction Behaviour:

- We illustrated with empirical evidence that interactions with SCS can be divided into search or non-search communication.
- We indicated that many different interaction behaviours can be observed in SCS which are not found in a browser-based search environment.
- We demonstrated that complexity and interactivity are fundamental components of SCS.

- We highlighted the importance that existing information seeking models are not sufficient to cover foundational communicative functions.

Chapter 7 – Identifying, Classifying, and Validating the Interaction Space for Spoken Conversational Search:

- We identified the possible actions taken in SCSdata and divide them into two themes and eight sub-themes which provide insight into the characteristics of SCS.
- We established a multi-tiered classification or annotation schema based on these identified actions, the SCoSAS, including both actors in the seeking process (the Seeker and the Intermediary) as equals, leveraging multi-turn activities and multi-move utterances. The SCoSAS facilitates further research in the conceptual understanding of human search dialogue behaviour, by enabling the researcher to select particular points of interest to investigate.
- We produced evidence that our SCoSAS is generalisable and replicable for a SCS setting by validating the SCoSAS with inter-rater reliability and code overlap with the SCSdata including coverage and code overlap with the MISC.

Chapter 8 – Task Complexity and Interactivity for Spoken Conversational Search:

- We contributed that task complexity has an effect on interaction and meta-discourse behaviour in SCS.
 - We showed that more complex queries relate to higher interaction counts.
 - We demonstrated that more complex tasks exhibited greater support utterances such as Discourse Management (i.e., confirmations).

Part IV – Discussion

Chapter 9 – Recommendations for the Design of Spoken Conversational Search Systems:

- We produced ten practical SCS design recommendations from triangulation of multiple data sources and methods, and discuss these recommendations concerning the SCoSAS action space. An objective of these recommendations is to help focus the SCS research.
- We created the first schematic model of SCS based on the SCoSAS which highlights that existing models do not sufficiently include discourse actions and the system as an active agent with its responsibilities.
- We included further conceptual avenues for SCS.

- We illustrated the importance of multi-disciplinary teamwork to advance in SCS while contributing new research avenues for SCS research.

10.2 Extensions

This thesis presents exploratory work and it is an initial investigation of SCS. As such, there are several limitations which could be addressed with the following extensions:

Human to human interaction: We are aware that human to human interaction may differ from the intended human-machine interactions [71]. However, since this is an exploratory study to understand the interactions which may lead to hypotheses formation, it was not a significant drawback and focus in our study. Nevertheless, we plan to conduct further studies to test our hypotheses in a human-machine interaction setting.

Lab setting: Participating in a lab setting influences the participants' behaviours [106]. Nevertheless, we believe that, even though this study was conducted in this setting, the overall findings will apply to a general day-to-day environment. In addition, the tasks tested in this study were created from TREC tasks and have traditionally been developed with the usage of a graphical interface in mind. Intuitively, we could investigate whether task design (i.e., whether it was developed to be completed in a browser-based or audio-only setting) impacted on search behaviour. Thus, investigating the information needs for SCS which arise in a natural setting will be necessary to develop natural systems. This will include understanding the different information needs and creating new taxonomies for these needs.

Taking initiative equals one turn: Our coding schema allows for coding per turn since we segmented the users' utterances with the idea that taking the initiative equals one turn. This means that slight subtleties inside a turn such as long pauses may be lost. However, we believe this was necessary to understand the broader context of SCS.

10.3 Informing Future Experiments

Our exploratory research into SCS allowed us to think broadly to expand knowledge in this new search interaction paradigm. We now present future research directions for SCS.

Interaction model and Wizard of Oz to test hypotheses: As seen in Figure 10.1, this thesis covers the data collection, creation of the first SCS annotation schema, analysis and validation of this schema, and design recommendations for SCS. Future work

includes creating detailed interaction models for this search paradigm which will inform the design of these systems, for example by the evaluation of particular features. The evaluation and hypotheses testing can be done in a WOZ setting.

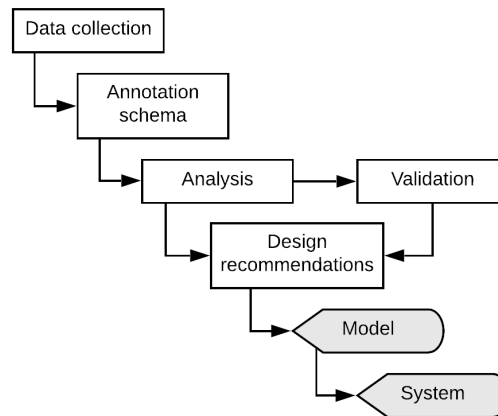


FIGURE 10.1: Future work includes creating models and systems.

Automatic utterance categorisation: Spoken utterance classification is a unique form of spoken language understanding and involves determining the function of the utterance in a dialogue [202]. These classification techniques can range from simple keyword detection to more sophisticated semantic classifications. Recent advances such as understanding the characteristic of particular utterances will be useful in this categorisation process [144]. However, Bunt et al. showed that dialogue utterances are often multi-functional or have a communicative function in more than one dimension, which makes utterance recognition and labelling still a complicated task [43]. Nevertheless, to scale the classification process of SCS's user input, automatic categorisation of these utterances will be necessary.

Cognitive level detection from user's speech: Tools are available to measure the cognitive levels from a person's speech unobtrusively [211]. For example, if a SCS system presents information to the user and the user responds, the SCS system could examine the cognitive load implied by that utterance. This utterance analysis could then form a basis of whether the system needs to adjust those responses depending on the cognitive load from the user or even suggest switching to a different device when a user is cognitively overloaded [50, 109, 212]. Other more suitable lab-based techniques such as fNIRS could also be used [214].

Suggesting device switching and adaptive conversational systems: It is desirable to overcome the cognitive overload of information while leveraging the interactions and conversations between the system and the user. Nevertheless, each user has different cognitive limitations at different times due to numerous external or internal factors.

Thus, conversational systems need to adapt continuously to the users' cognitive interaction abilities. For example, if the system detects from the user's voice that they are struggling or cognitively overloaded, the system should adapt the interaction strategies. This may require that the system should suggest searching at another time or on other device and assist the user in this change. Thus, device or search strategy switching allows for a dynamic user–system interaction which will be experienced as more natural [62].

Refinding and revisitation of search results: Understanding how users refind or revisit their results in an audio setting may provide us with another opportunity to understand the user's cognition or search process while searching. Studying this phenomenon through dialogue offers a unique opportunity to investigate particular utterances, for example, grounding, and their role in the refinding process [47].

Understanding conversational search turn-taking: Turn-taking is an essential phenomenon in dialogue [110]. The system incrementally releases information, and the user can process these data without having to handle all information in one go and is referred to as turn-taking phenomena [110] or user revelation [137]. Analysing this turn-taking and revelation behaviour while incorporating these results in the dialogue can increase dialogue efficiency.

Results presentation, response generation strategies, and results organisation: One major usability factor of these spoken systems is how it presents results and information units, including how to structure the output of these results (i.e., the organisation such as clustering). It is well understood that just reading out a results list or text is not sufficient and that one cannot translate a graphical user interface into an audio one [51, 224]. However, it is still unclear *which* search results should be presented and *how* this should be done over an audio-only communication channel. Our observational study provided an initial natural approach to learn how people express, structure, or summarise found information units. Next, we can test these approaches in a crowdsourcing framework as specified in Chapter 4. Furthermore, investigating comprehensive readability aspects for audio such as the listenability of a document are new avenues of multi-disciplinary research [217] including understanding if readability can help as a measure of the quality of a spoken summary [226].

Identifying interaction cost: Much research has been devoted to understanding the costs associated with interactions throughout a search process [141, 142]. These costs have been identified in many different ways, including temporal, physical, or mental demands. For example, the interaction costs could be viewed as an economic problem in which different costs are assigned to different conversational interaction behaviours [18]. More precisely, by investigating the interaction cost as an economics problem, we can

use labelled datasets such as the SCSdata to understand which conversational strategies are effective in terms of user effort and gain.

Evaluation: The extensive list of possible future experiments listed above demonstrates the complexity of SCS as an end-to-end problem. Thus, the definition of evaluation frameworks for SCS and, more generally, conversational IR, is also challenging [62, 172]. Recent evaluation initiatives such as the TREC 2019 Conversational Assistance Track (CAST) [64] aim to create reusable test collections for text-based information-centric conversational dialogues. We believe that novel evaluation methodologies and frameworks are needed for SCS to leverage the knowledge of SDS evaluation [90]. The creation of resources such as SCoSAS, MISC, and other testbeds is arguably the first step to inform the evaluation of SCS.

Concerning our future extensions, we will examine the collected questionnaire data from the observational study and the semi-structured interviews. We are also investigating ideal paths as a dialogue schema and completing the sequence mining of SCSdata.

10.3.1 Informing Wider Research Agendas

We believe our research implications extend to other research areas. For example, existing systems and models have difficulties with multi-turn actions, utterances which consist of multiple moves, or intent extraction. In this thesis, we attempted to better understand these unique features of SCS by creating a labelling schema and schematic model of these labels. Our model and annotation schema provide a novel extension of prior preliminary SCS models [197] and the conceptual framework Azzopardi et al. [19]. While the labelling schema developed in this thesis provides insight into the interaction space of SCS and possible actions taken by both actors we also instigated a new understanding of non-search related or discourse actions. These non-search related actions provide much-needed information for future SCS systems and are likely to transform search interactions.

Appendix A

Ethics Approvals and Participant Information Statement

A.1 Ethics Approval BSEHAPP 10-14



College of Science, Health & Engineering

College Human Ethics Advisory Network (CHEAN)

Email: [REDACTED]

Tel: [61 3] 9925 4620

Building 91, Level 2, City Campus/Building 215, Level 2, Bundoora West Campus

29 January 2019

Professor Mark Sanderson
School of Science
RMIT University

Dear Professor Sanderson

RE: BSEHAPP 10-14 An investigation into browsing web search results over a speech-only communication channel

Thank you for submitting the above amendment to your approved ethics application for consideration by the Science Engineering & Health College Human Ethics Advisory Network (CHEAN).

The application was considered and reviewed by the CHEAN in January 2019. Your ethics approval has now been extended to 31 January 2020.

Status: Approved

The CHEAN reviewed the above amendment application and agreed that it meets the requirements of the *National Statement on Ethical conduct in Human Research, NHMRC, 2007* (NS) guidelines and approves the requested amendment.

If there is anything in this letter that you are unclear about or require further clarification upon then please contact the CHEAN secretary, Ms Mary Duffy.

Yours sincerely



**Associate Professor Barbara Polus
Chair, Science Engineering & Health
College Human Ethics Advisory Network**

Cc: Student Investigator/s:
Other Investigator/s:

Joanne Trippas [REDACTED] School of Computer Science & IT RMIT University
Lawrence Cavedon School of Computer Science & IT RMIT University



A.2 Ethics Approval ASEHAPP 08-16



College Human Ethics Advisory Network (CHEAN)
College of Science, Engineering and Health

Email: [REDACTED]
Phone: [61 3] 9925 4620
Building 91, Level 2, City Campus/Building 215, Level 2, Bundoora West Campus

21 February 2017

Associate Professor Lawrence Cavedon
School of Science
RMIT University

Dear A/Prof Cavedon

ASEHAPP 08-16 Investigating Search Behaviour over Audio

Thank you for requesting an amendment to your Human Research Ethics project titled: **Investigating Search Behaviour over Audio**, which was originally approved by Science Engineering and Health CHEAN in 2016 for a period of 2 years.

I am pleased to inform you that the CHEAN has **approved** your amendment as outlined in your request.

The CHEAN notes and thanks you for providing all documentation that incorporates these amendments. This documentation will be appended to your file for future reference and your research may now continue.

The committee would like to remind you that:


All data should be stored on University Network systems. These systems provide high levels of manageable security and data integrity, can provide secure remote access, are backed up on a regular basis and can provide Disaster Recover processes should a large scale incident occur. The use of portable devices such as CDs and memory sticks is valid for archiving; data transport where necessary and for some works in progress; The authoritative copy of all current data should reside on appropriate network systems; and the Principal Investigator is responsible for the retention and storage of the original data pertaining to the project for a minimum period of five years.


Please Note: Annual reports are due on the anniversary of the commencement date for all research projects that have been approved by the CHEAN. Ongoing approval is conditional upon the submission of annual reports failure to provide an annual report may result in Ethics approval being withdrawn.

Final reports are due within six months of the project expiring or as soon as possible after your research project has concluded.

The annual/final reports forms can be found at:
www.rmit.edu.au/staff/research/human-research-ethics

Yours faithfully,


Associate Professor Barbara Polus
Chair, Science Engineering & Health
College Human Ethics Advisory Network

Cc Other Investigator/s: Johanne Trippas  School of Science
Professor Mark Sanderson School of Science

A.3 Participant Information Statement



INVITATION TO PARTICIPATE IN A RESEARCH PROJECT

PARTICIPANT INFORMATION

Project Title: Investigating Search Behaviour over Audio

Dear participant,

You are invited to participate in a research project being conducted by RMIT University. Please read this sheet carefully and be confident that you understand its contents before deciding whether to participate. If you have any questions about the project, please ask one of the investigators.

Who is involved in this research project?

Researchers at RMIT are conducting the project. This research is performed by Johanne Trippas as part of her PhD in Computer Science. She is under the supervision of Prof. Mark Sanderson, Assoc. Prof. Lawrence Cavedon and Dr. Damiano Spina of RMIT. RMIT Human Research Ethics Committee has approved this research project.

If you have any questions, please contact Johanne Trippas on [REDACTED]

What is the project about?

We are conducting this project to gain better insight into how people communicate with a search engine (such as Google or Bing) by using speech, when no keyboard or screen is available. We seek a better understanding of how to present search results over audio while not overwhelming the users with information, nor leaving users uncertain as to whether what they covered the information space. We expect to form new hypotheses and research questions from this study.

If I agree to participate, what will I be required to do?

You will be shown a scenario with an underlying information need. You will then have to communicate this information need to the other participant who has access to a search engine. The other participant can use the search engine to help you with finding the information you need from the scenario. The only way to communicate this information need is through talking with each other. The roles of the person with the access to the scenario and the search engine will be reversed. We will put something between you and the other participant so you cannot see each other and really need to focus on what the other participant says without picking up on facial expressions.

Short post-task questionnaires will be provided after each scenario. At the end of the experiment, we will conduct a short interview where you can provide any feedback.

If you have questions or comments during the experiment, please ask the investigators, we are here to help you. You may leave at any time.

What are the possible risks or disadvantages?

There are no perceived risks outside your normal day-to-day activities.

Reading scenarios and trying to complete the task is not the most exciting work. The scenarios have been screened and do not tackle culturally sensitive issues. If, however, you prefer not to judge a particular query for any reason, just skip the query and move to the next one. You can stop the participation any time.

What are the benefits associated with participation?

There are no direct benefits to you for participating in this study. However, the data collected in the study may help to contribute to public knowledge of how search engines can be made more user friendly for a wide audience.

To thank you for your time, you will receive a \$20 Coles Group & Myer gift card for your participation in the study.

What will happen to the information I provide?

The recordings from the conversations between you and the other participant and the interviews will be transcribed for analysis. This allows us to de-identify the recordings and conduct analysis from the transcriptions instead of the recordings.

The data from the questionnaires will be analysed. The data will be stored on a password-protected computer at RMIT for five (5) years and will not be shared with others. The research conducted using this data will be published in a PhD thesis and refereed journal or conference. We hope the publication will happen sometime between 2016 and 2018.

We will keep the data safely locked away, however, there might be a possibility that we want to revisit the data later on. This means that the data might be used in a future project either by us or by another researcher.

What will happen to the video recordings?

The video recordings will be transcribed in text format which allows us to analyse the results. Once the recordings are transcribed, they will be stored on a password-protected computer at RMIT for five (5) years. No personal identifiers will be stored as we will de-identify all the recordings and transcriptions with IDs. No images from these recordings will be altered, copied or used for publication.

What are my rights as a participant?

- You have the right to withdraw from participation at any time.
- You have the right to have any unprocessed data withdrawn and destroyed.
- You have the right to request that any recording cease.
- You have the right to be de-identified in any photographs intended for public publication, before the point of publication.
- You have the right to ask questions (via email or in person) at any time.

Whom should I contact if I have any questions?

Please contact Johanne Trippas ([REDACTED]).

What other issues should I be aware of before deciding whether to participate?

You will be working with another participant in this study. If you don't feel comfortable conducting a search with the other participant, please feel free to leave at any time.

Yours sincerely,

Prof. Mark Sanderson ([REDACTED])

Assoc. Prof. Lawrence Cavedon ([REDACTED])

Dr. Damiano Spina ([REDACTED])

Johanne Trippas ([REDACTED])

If you have any concerns about your participation in this project, which you do not wish to discuss with the researchers, then you can contact the Ethics Officer, Research Integrity, Governance and Systems, RMIT University, GPO Box 2476V VIC 3001. Tel: (03) 9925 2251 or email human.ethics@rmit.edu.au

CONSENT FORM

1. I have had the project explained to me, and I have read the information sheet
2. I agree to participate in the research project as described
3. I agree:
 - to undertake the tests or procedures outlined
 - to be interviewed and/or complete a questionnaire
 - that my voice will be audio recorded
 - that my image will be taken and no images from these recordings will be altered, copied or used for publication.
4. I acknowledge that:
 - (a) I understand that my participation is voluntary and that I am free to withdraw from the project at any time and to withdraw any unprocessed data previously supplied (unless follow-up is needed for safety).
 - (b) The project is for the purpose of research. It may not be of direct benefit to me.
 - (c) The privacy of the personal information I provide will be safeguarded and only disclosed where I have consented to the disclosure or as required by law.
 - (d) The security of the research data will be protected during and after completion of the study. The data collected during the study may be published. Any information which will identify me will not be used.

Participant's Consent

Participant: _____ Date: _____
(Signature)

Appendix B

Questionnaires and Semi-structured Observational Study Interview Questions

Unless otherwise indicated all items are evaluated with a five-point scale, where 1=Not at all, 2=Slightly, 3=Moderately, 4=Very, and 5=Extremely.

B.1 Pre-task questionnaire for the Seeker

TABLE B.1: Pre-task questionnaire for the Seeker.

Measure	Question
Interest and Knowledge	How many times have you searched for information about this task?* [1=7 times or more, 2=5-6 times, 3=3-4 times, 4=1-2 times, 5=Never] I am interested to learn more about the topic of the task.* How knowledgeable are you about the topic of the task?*
Task Complexity	How defined is this task in terms of the types of information needed to complete it?*" How defined is this task in terms of the steps required to complete it?*" How defined is this task in terms of its expected solution?*"
Expected Task Difficulty	In this simulated search environment, how easy do you think it will be to search for information for this task?*" In this simulated search environment, how easy do you think it will be to understand the information found?*" In this simulated search environment, how easy do you think it will be to decide if the information found is useful for completing the task?*" In this simulated search environment, how easy do you think it will be to determine when you have enough information to finish the task?*"

NOTE: * Adapted from Kelly et al. [108].

B.2 Post-task questionnaire for the Seeker

TABLE B.2: Post-task questionnaire for the Seeker.

Measure	Question
Interest and Knowledge	I am interested to learn more about the topic of the task.* In this simulated environment, how much did your knowledge of the task increase as you searched?*
Experienced Task Difficulty	In this simulated search environment, how easy was it to search for information for this task? In this simulated search environment, how easy was it to understand the information found? In this simulated search environment, how easy was it to decide if the information found was useful for completing the task? How easy was it to determine when you had enough information to finish the task?
Experienced Conversational Difficulty	Thinking about the content of the information, how understandable was the information given by your partner? Thinking about the content of the information, how logical was the information given by your partner? How easy did you find verbalising the information need compared to typing it?
Experienced Collaboration Difficulty	How would you rate the collaboration between you and your partner? [Where 1=Very poor and 5=Very good] I gave clear instructions as to what my partner had to search for. My partner gave me clear directions to help him/her with the search task.
Experienced Search Presentation Difficulty	My partner presented a good overview of the search results. My partner presented the search results in a way that was easy to understand. My partner gave me enough information to select the most relevant result. My partner provided enough information to help me solve the search task.
Overall Difficulty	Overall, how easy was this task?*
Overall Satisfaction	Overall, how satisfied are you with your solution to this task?* Overall, how satisfied are you with the search strategy you took to solve this task?*
Open question	What would you have done differently to accomplish this search task? To what extent did you achieve your search goal?

NOTE: * Adapted from Kelly et al. [108].

B.3 Post-task questionnaire for the Intermediary

TABLE B.3: Post-task questionnaire for the Intermediary.

Measure	Question
Experienced Conversational Difficulty	Thinking about the content of the information, how understandable was the information given by your partner? Thinking about the content of the information, how logical was the information given by your partner? How well was the search query formulated by your partner? How well did you understand what your partner was searching for? How easy did you find completing the search task with the the information your partner gave you?
Experienced Collaboration Difficulty	How would you rate the collaboration between you and your partner? [Where 1=Very poor and 5=Very good] My partner gave me clear search directions. I gave clear instructions to my partner in order to conduct the search with the search engine.
Experienced Search Presentation Difficulty	How easy did you find verbalising the information that you read on the screen? How well do you think your partner understood what you verbalised from the screen? I presented a good overview of the available options. I presented the search results in a way that was easy to understand. I presented the search results in a way that gave my partner enough information to select the most relevant result.
Overall Difficulty	Overall, how easy was this task?*
Overall Satisfaction	Overall, how satisfied are you with your solution to this task?*
	Overall, how satisfied are you with the search strategy you took to solve this task?*
Open question	What would you have done differently to accomplish this search task? To what extent did you achieve your search goal?

NOTE: * Adapted from Kelly et al. [108].

B.4 Exit questionnaire for the Seeker

TABLE B.4: Exit questionnaire for the Seeker.

Measure	Question
Experienced Conversational Difficulty	<p>The length of my partner's statements was appropriate to complete the task.</p> <p>I found the general conversation flow with my partner comfortable.</p> <p>I felt overloaded with information from my partner.</p> <p>My partner spoke too quickly.</p> <p>I gave clear instructions about what my partner had to do.</p> <p>My partner gave me clear instructions as to what I had to do.</p> <p>The information spoken by my partner was too complicated to understand what had actually been said.</p> <p>My partner understood the meaning of what I said.</p>
Experienced Collaboration Difficulty	<p>My partner worked together with me in the search task.</p> <p>My partner encouraged me to give clear search directions.</p> <p>My partner was disruptive in the search task.</p> <p>My search would have been faster if I had used the search engine by myself.</p> <p>My search would have been more efficient if I had used the search engine by myself.</p>
Experienced Search Presentation Difficulty	<p>Hearing an overview of all possible options is important to me.</p> <p>I found not having visual information from the search engine difficult.</p>
Open Questions	<p>What did you like about this study?</p> <p>What did you dislike about this study?</p> <p>How could we improve this study?</p>

B.5 Exit questionnaire for the Intermediary

TABLE B.5: Exit questionnaire for the Intermediary.

Measure	Question
Experienced Conversational Difficulty	The length of my partner's statements was appropriate to complete the task.
	I found the general conversation flow with my partner comfortable.
	I felt overloaded with information from my partner.
	My partner spoke too quickly.
	I gave clear instructions about what my partner had to do.
	My partner gave me clear instructions as to what I had to do.
	The information spoken by my partner was too complicated to understand what had actually been said.
My partner understood the meaning of what I said.	
Experienced Collaboration Difficulty	My partner worked together with me in the search task.
	My partner encouraged me to give clear information about what I found on the search engine.
	My partner was disruptive in the search task.
	My search would have been faster if I had known the search tasks and used the search engine by myself.
	My search would have been more efficient if I had known the search task and used the search engine by myself.
I found searching without knowing the scenario difficult.	
Open Questions	What did you like about this study?
	What did you dislike about this study?
	How could we improve this study?

B.6 Semi-structured Observational Study Interview Questions

Task Complexity:

- Before you started your search, did you have any expectations of how your partner would react on your query (results)?
 - What were the expectations?

Expected and experienced task difficulty:

- Can you recall a time when you felt engaged in a search task in this simulated search environment?
 - What topic were you searching?
 - Which website were you looking at?
- Can you recall a time when you felt frustrated in a search task in this simulated search environment?
 - What topic were you searching?
 - Which website were you looking at?
 - Was there something specific that made you felt frustrated?

Interest and Knowledge:

- What were the key points or moments that triggered your interest in the search task?

Experienced conversational difficulty:

- How did you find the general conversation flow between you and your partner?
 - Which were moments you understood each other and you had a common understanding of what you were searching for (“aha moment”)?
 - Which were moments you did not understand each other?
 - What were the strategies to make sure you understood your partner correctly?
- Thinking of the conversation, if you had to refind a result, how would you do this?

- If you need to refind something when you are searching on your own computer on a search engine, how do you do this? What would be different in this setting where you do not see the screen and you cannot type the search query?
- How did you find verbalising your search or search results?
 - What would you have done differently if you were in control of the search engine by yourself via a keyboard and screen?

Experienced collaboration difficulty:

- How did you find the conversation about the search task went between you and your partner?
 - Did you find any useful sentences or probes to receive more information from your partner?
 - What were the probes your partner did not respond on the way you anticipated?
 - Imagine if you did not understand your partner, there was noise or you were not paying attention to what he/she was saying. In what way would you try to understand what your partner had said?

Experienced search presentation difficulty:

- What were the techniques your partner used to present you with the search results in a way that was easy to understand?
 - Can you think of a moment that you clearly understood what your partner was saying in the aspect of what he/she found on the search engine results page?
 - Which techniques would you have used to present the search results to your partner if you were using the search engine?

Overall difficulty:

- What would you do differently to make this search process easier?
- How do you see this kind of search work in the future?
- Thinking about a technique of presenting the search results how do you think clustering would impact your way of searching?

Appendix C

SCSdata

Released data can be found on http://bit.ly/SCSdata_thesis and has received an ACM SIGIR badge for having the dataset publicly available.

Artefact type: Dataset

ACM SIGIR badge: Artefacts Available¹

C.1 Provided Files

We provide all the releasable data in different files:

- Transcripts (ConversationalSearchDataSet.csv) and (SCSdataset.csv)
- Backstories (backstories_ConversationalSearchDataSet.csv)
- Code book (CodeBook_CHIIR.pdf)

C.2 Acknowledgments

This research is partially supported by Australian Research Council Project LP130100563 and Real Thing Entertainment Pty Ltd. The data collection and release was reviewed and approved by RMIT University's Ethics Board (ASEHAPP 08-16).

The authors were employed by RMIT University when these transcripts were created.

¹https://openreview.net/forum?id=rJgGxq1_z4

Appendix D

Spoken Conversational Search Interaction Themes

D.1 Theme 1: Task Level

TABLE D.1: Information Request (Seeker).

Theme	Sub-theme	Actor	Code	Frequency
Task Level	Information Request	Seeker	Automated repetitive search	3
		Seeker	Definition explanation	1
		Seeker	Definition lookup or person	1
		Seeker	Information about document	6
		Seeker	Information about SERP overview	2
		Seeker	Information request	67
		Seeker	Information request within document	80
		Seeker	Information request within SERP	15
		Seeker	Initial information request	39
		Seeker	Intent clarification	52
		Seeker	Query embellishment	20
		Seeker	Spells (query or query word)	2
		Intermediary	Definition clarification	1
		Intermediary	Enquiry for further information	11
		Intermediary	Google query expansion suggestion	3
		Intermediary	Query refinement offer	57
		Intermediary	Query rephrase	12
		Intermediary	Requests more details about information request	5
		Intermediary	Query formulation for information found in document	1
		Intermediary	Asking what they are looking for	2
Intermediary	Within-Document search result entity lookup request	1		

TABLE D.2: Results Presentation (Intermediary).

Theme	Sub-theme	Actor	Code	Frequency
Task Level	Results Presentation	Intermediary	Source information	8
		Intermediary	Image overview on SERP	2
		Intermediary	Interpretation of photos	1
		Intermediary	Multi-document summary	3
		Intermediary	Paraphrasing from document which is not in front of them	1
		Intermediary	Scanning document with modification	51
		Intermediary	Scanning document without modification	79
		Intermediary	Scanning document without modification but with interpretation of photos	1
		Intermediary	SERP Card	16
		Intermediary	SERP overview without modification	1
		Intermediary	SERP with modification	19
		Intermediary	SERP without modification	72
		Intermediary	Within SERP search result	4
		Intermediary	Within-Document command response	1
		Intermediary	Within-Document search result	60
		Intermediary	Interpretation biased towards information request or clarification given by the User	1
		Intermediary	Comparing results against each other	1
		Intermediary	Interpretation	22

TABLE D.3: Search Assistance (Seeker and Intermediary).

Theme	Sub-theme	Actor	Code	Frequency
Task Level	Search Assistance	Seeker	Recommendations	1
		Seeker	Requests “enough information” judgement	1
		Intermediary	Asking about usefulness	4
		Intermediary	Requests spelling	2
		Intermediary	Suggestion to move on	2
		Intermediary	Relevance judgement	6
		Intermediary	Suggestion to search more	1
		Intermediary	Requests to access search engine	1
		Intermediary	Search suggestion based on info encountered in document	1

TABLE D.4: Search Progression (Seeker).

Theme	Sub-theme	Actor	Code	Frequency
Task Level	Search Progression	Seeker	Enough information	6
		Seeker	Performance feedback	18
		Seeker	Rejects	9

D.2 Theme 2: Discourse Level

TABLE D.5: Discourse Management (Seeker and Intermediary).

Theme	Sub-theme	Actor	Code	Frequency
Discourse Level	Discourse Management	Seeker	Asks to repeat	31
		Seeker	Asks to repeat first search result	6
		Seeker	Asks to repeat Nth search result	1
		Seeker	Confirms	114
		Seeker	Query repeat	14
		Intermediary	Asks to repeat	38
		Intermediary	Checks navigational command	13
		Intermediary	Confirms	46
		Intermediary	Repeats	12
		Intermediary	Repeats the query back	9

TABLE D.6: Grounding (Seeker).

Theme	Sub-theme	Actor	Code	Frequency
Discourse Level	Grounding	Seeker	Creating bigger picture	1
		Seeker	Interpretation	12

TABLE D.7: Navigation (Seeker).

Theme	Sub-theme	Actor	Code	Frequency
Discourse Level	Navigation	Seeker	Access link within document	1
		Seeker	Access search engine	2
		Seeker	Access source	29
		Seeker	Access source (implicit)	2
		Seeker	Between-document navigation	1
		Seeker	Is there more information	6
		Seeker	Leave document	1
		Seeker	Next	3
		Seeker	Read more from the document	1
		Seeker	Within-document command	3

TABLE D.8: Visibility of System Status (Seeker and Intermediary).

Theme	Sub-theme	Actor	Code	Frequency
Discourse Level	Visibility of system status	Seeker	Access source feedback-request	3
		Seeker	Feedback on what is happening	1
		Seeker	Results?	10
		Intermediary	Feedback on what is happening	13
		Intermediary	Misheard	1
		Intermediary	Previously seen results	2
		Intermediary	Wayfinding	3

D.3 Theme 4: Other Level

TABLE D.9: Other Level (Seeker).

Theme	Sub-theme	Actor	Code	Frequency
Other Level		Seeker	Utter (“So I’m” and “Well so they are saying”)	2
		Seeker	Provides information about the Search Engine (“So it’s [a] search engine”)	1
		Seeker	Asks if allowed to query embellish (“Actually can I add something else to that?”)	1
		Seeker	Offers to spell (“[...] would you like me to spell it?”)	1

Bibliography

- [1] A. Abdolrahmani and R. Kuber. Should I trust it when I cannot see it?: Credibility assessment for blind web users. In *Proceedings of International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*, pages 191–199, 2016.
- [2] J. Ajmera, A. Joshi, S. Mukherjea, N. Rajput, S. Sahay, M. Shrivastava, and K. Srivastava. Two-stream indexing for spoken web search. In *Proceedings of World Wide Web Conference (WWW)*, pages 503–512, 2011.
- [3] F. N. Akinnaso. On the differences between spoken and written language. *Language and Speech*, 25(2):97–125, 1982.
- [4] W. Albert, T. Tullis, and D. Tedesco. *Beyond the Usability Lab: Conducting Large-Scale Online User Experience Studies*. Morgan Kaufmann, 2009.
- [5] M. Aliannejadi, M. Hasanain, J. Mao, J. Singh, J. R. Trippas, H. Zamani, and L. Dietz. ACM SIGIR student liaison program. *ACM SIGIR Forum*, 51(3):42–45, 2018.
- [6] J. Allan, B. Croft, A. Moffat, and M. Sanderson. Frontiers, Challenges, and Opportunities for Information Retrieval: Report from SWIRL 2012 the Second Strategic Workshop on Information Retrieval in Lorne. *ACM SIGIR Forum*, 46(1):2–32, 2012.
- [7] J. Allen and M. Core. DAMSL: Dialogue act markup in several layers (draft 2.1). Technical report, Multiparty Discourse Group, Discourse Resource Initiative, 1997.
- [8] J. Altmann. Observational study of behavior: sampling methods. *Behaviour*, 49(3-4):227–266, 1974.
- [9] L. W. Anderson, D. R. Krathwohl, and B. S. Bloom. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom’s Taxonomy of Educational Objectives*. Longman, 2001.

- [10] H. Ando, R. Cousins, and C. Young. Achieving saturation in thematic analysis: Development and refinement of a codebook. *Comprehensive Psychology*, 3(4):1–7, 2014.
- [11] J. Arguello, W.-C. Wu, D. Kelly, and A. Edwards. Task complexity, vertical display and user interaction in aggregated search. *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*, pages 435–444, 2012.
- [12] J. Arguello, S. Avula, and F. Diaz. Using query performance predictors to improve spoken queries. In *Proceedings of the European Conference on Information Retrieval (ECIR)*, pages 309–321, 2016.
- [13] J. Arguello, B. Choi, and R. Capra. Factors influencing users’ information requests: Medium, target, and extra-topical dimension. *ACM Transactions on Information Systems (TOIS)*, 36(4):41:1–41:37, 2018.
- [14] B. Arons. SpeechSkimmer: interactively skimming recorded speech. In *Proceedings of User Interface Software and Technology (UIST)*, pages 187–196. ACM Press, 1993.
- [15] B. Arons. SpeechSkimmer: a system for interactively skimming recorded speech. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 4(1):3–38, 1997.
- [16] A. Aula, R. M. Khan, and Z. Guan. How does search behavior change as search becomes more difficult? *Proceedings of Conference on Human Factors in Computing Systems (CHI)*, pages 35–44, 2010.
- [17] S. Avula, G. Chadwick, J. Arguello, and R. Capra. Searchbots: User engagement with chatbots during collaborative search. In *Proceedings of Conference on Information Interaction and Retrieval (CHIIR)*, pages 52–61, 2018.
- [18] L. Azzopardi. The economics in interactive information retrieval. In *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*, pages 15–24, 2011.
- [19] L. Azzopardi, M. Dubiel, M. Halvey, and J. Dalton. Conceptualizing agent-human interactions during the conversational search process. In *SIGIR 2nd International Workshop on Conversational Approaches to Information Retrieval (CAIR’18)*, 2018. 8 pages.
- [20] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. User variability and IR system evaluation. In *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*, pages 625–634, 2015.

- [21] R. Barzilay, K. R. McKeown, and M. Elhadad. Information fusion in the context of multi-document summarization. In *Proceedings of Association for Computational Linguistics (ACL)*, pages 550–557, 1999.
- [22] N. J. Belkin. Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, 5(1):133–143, 1980.
- [23] N. J. Belkin. Cognitive models and information transfer. *Social Science Information Studies*, 4(2-3):111–129, 1984.
- [24] N. J. Belkin, H. M. Brooks, and P. J. Daniels. Knowledge elicitation using discourse analysis. *International Journal of Man-Machine Studies*, 27(2):127–144, 1987.
- [25] N. J. Belkin, C. Cool, A. Stein, and U. Thiel. Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert Systems with Applications*, 9(3):379–395, 1995.
- [26] D. J. Bell and I. Ruthven. Searcher’s assessments of task complexity for web searching. In *Proceedings of the European Conference on Information Retrieval (ECIR)*, pages 57–71, 2004.
- [27] J. Belsky. Mother–father–infant interaction: A naturalistic observational study. *Developmental Psychology*, 15(6):601, 1979.
- [28] L. Benaquisto. Codes and coding. In L. Given, editor, *The SAGE Encyclopedia of Qualitative Research Methods*, pages 86–88. SAGE Publications, 2008.
- [29] T. Berners-Lee, J. Hendler, O. Lassila, et al. The semantic web. *Scientific American*, 284(5):28–37, 2001.
- [30] D. Biber. Are there linguistic consequences of literacy? Comparing the potentials of language use in speech and writing. In D. R. Olson and N. Torrance, editors, *Cambridge Handbook of Literacy*, pages 75–91. Cambridge University Press, 2009.
- [31] P. Borlund. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research. An International Electronic Journal*, 8(3), 2003.
- [32] P. Borlund. Interactive information retrieval: An introduction. *Journal of Information Science Theory and Practice*, 1(3):12–32, 2013.
- [33] H. Bota, K. Zhou, and J. M. Jose. Playing your cards right: The effect of entity cards on search behaviour and workload. In *Proceedings of Conference on Information Interaction and Retrieval (CHIIR)*, pages 131–140, 2016.

- [34] R. E. Boyatzis. *Transforming Qualitative Information: Thematic Analysis and Code Development*. SAGE Publications, 1998.
- [35] S. J. Boyce. User interface design for natural language systems: From research to reality. In *Human Factors and Voice Interactive Systems*, pages 43–80. Springer, 2008.
- [36] V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.
- [37] V. Braun and V. Clarke. *Successful Qualitative Research: A Practical Guide for Beginners*. SAGE Publications, 2013.
- [38] K. Brennan, D. Kelly, and Y. Zhang. Factor analysis of a Search Self-Efficacy scale. In *Proceedings of Conference on Information Interaction and Retrieval (CHIIR)*, pages 241–244, 2016.
- [39] S. Buchholz, J. Latorre, and K. Yanagisawa. Crowdsourced assessment of speech synthesis. *Crowdsourcing for Speech Processing*, pages 173–216, 2013.
- [40] H. Bunt. Dynamic interpretation and dialogue theory. *The Structure of Multimodal Dialogue*, 2:1–8, 1999.
- [41] H. Bunt. The DIT++ taxonomy for functional dialogue markup. In *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, pages 13–24, 2009.
- [42] H. Bunt, J. Alexandersson, J. Carletta, J. Choe, A. C. Fang, K. Hasida, K. Lee, V. Petukhova, A. Popescu-Belis, L. Romary, C. Soria, and D. Traum. Towards an ISO standard for dialogue act annotation. In *Seventh Conference on International Language Resources and Evaluation (LREC’10)*, pages 2548–2555, 2010.
- [43] H. Bunt, V. Petukhova, D. Traum, and J. Alexandersson. Dialogue act annotation with the ISO 24617-2 standard. In *Multimodal Interaction with W3C Standards*, pages 109–135. Springer, 2017.
- [44] K. Byström and K. Järvelin. Task complexity affects information seeking and use. *Information Processing & Management*, 31(2):191–213, 1995.
- [45] C. Callison-Burch and M. Dredze. Creating speech and language data with amazon’s mechanical turk. In *Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk (NAACL HLT CSLDAMT’10)*, pages 1–12, 2010.

- [46] R. Capra, J. Arguello, A. Crescenzi, and E. Vardell. Differences in the use of search assistance for tasks of varying complexity. In *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*, pages 23–32, 2015.
- [47] R. G. Capra and M. A. Pérez-Quiñones. Re-finding found things: An exploratory study of how users re-find information. *arXiv preprint cs/0310011*, 2003. 9 pages.
- [48] D. Case. *Looking For Information. A Survey of Research on Information Seeking, Needs and Behavior*. Emerald Group Publishing Limited, 2012.
- [49] E. Chang, F. Seide, H. M. Meng, C. Zhuoran, S. Yu, and L. Yuk-Chi. A system for spoken query information retrieval on mobile devices. *Transactions on Speech and Audio Processing*, 10(8):531–541, 2002.
- [50] F. Chen, N. Ruiz, E. Choi, J. Epps, M. A. Khawaja, R. Taib, B. Yin, and Y. Wang. Multimodal behavior and interaction as indicators of cognitive load. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(4):22, 2012.
- [51] A. Chuklin, A. Severyn, J. R. Trippas, E. Alfonseca, H. Silen, and D. Spina. Prosody modifications for question-answering in voice-only settings. *arXiv preprint arXiv:1806.03957*, pages 1–5, 2018.
- [52] H. H. Clark and S. E. Brennan. Grounding in communication. In *Perspectives on Socially Shared Cognition*, pages 222–233, 1991.
- [53] C. L. Clarke, N. Craswell, and E. M. Voorhees. Overview of the TREC 2012 Web track. Technical report, National Institute of Standards and Technology Gaithersburg MD, 2012.
- [54] C. L. A. Clarke, E. Agichtein, S. Dumais, and R. W. White. The influence of caption features on clickthrough patterns in web search. In *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*, pages 135–142, 2007.
- [55] C. W. Cleverdon, J. Mills, and E. M. Keen. Factors determining the performance of indexing systems, (Volume 1: Design). *Cranfield: College of Aeronautics*, 1966.
- [56] P. R. Cohen and S. L. Oviatt. The role of voice input for human-machine communication. *Proceedings of the National Academy of Sciences*, 92(22):9921–9927, 1995.
- [57] E. Coiera and V. Tombs. Communication behaviours in a hospital setting: an observational study. *The BMJ*, 316(7132):673–676, 1998.

- [58] K. Collins-Thompson, Paul Bennett, F. Diaz, C. L. a. Clarke, and E. M. Voorhees. TREC 2014 Web track overview. *Proceedings of Text REtrieval Conference Proceedings (TREC)*, pages 1–21, 2013.
- [59] S. L. Condon, W. R. Edwards, et al. Measuring conformity to discourse routines in decision-making interactions. In *Proceedings of Association for Computational Linguistics (ACL)*, pages 238–245, 1999.
- [60] F. Crestani and H. Du. Written versus spoken queries: A qualitative and quantitative comparative analysis. *Journal of the American Society for Information Science and Technology*, 57(7):881–890, 2006.
- [61] W. B. Croft and R. H. Thompson. I3R: A new approach to the design of document retrieval systems. *Journal of the American Society for Information Science*, 38(6):389–404, 1987.
- [62] J. S. Culpepper, F. Diaz, and M. D. Smucker. Research Frontiers in Information Retrieval: Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018). *ACM SIGIR Forum*, 52(1):34–90, 2018.
- [63] E. Cutrell and Z. Guan. What are you looking for?: an eye-tracking study of information usage in web search. In *Proceedings of Conference on Human Factors in Computing Systems (CHI)*, pages 407–416, 2007.
- [64] J. Dalton, C. Xiong, and J. Callan. The TREC 2019 Conversational Assistance Track (CAsT). <http://treccast.ai/>, 2018. (Last accessed January 24, 2019).
- [65] M. C. Day and S. J. Boyce. Human factors in human–computer system design. In M. C. Yovits, editor, *Advances in Computers*, volume 36, pages 333–430. Elsevier, 1993.
- [66] V. Demberg and J. D. Moore. Information presentation in spoken dialogue systems. In *Proceedings of European Chapter of the Association for Computational Linguistics (EACL)*, pages 65–72, 2006.
- [67] V. Demberg and A. Sayeed. Linguistic cognitive load: implications for automotive UIs. In *Proceedings of AutomotiveUI*, 2011. 4 pages.
- [68] V. Demberg, A. Winterboer, and J. D. Moore. A strategy for information presentation in spoken dialog systems. *Computational Linguistics*, 37(3):489–539, 2011.
- [69] B. Dervin and P. Dewdney. Neutral questioning: A new approach to the reference interview. *RQ*, pages 506–513, 1986.

- [70] T. A. Dilorenzo, J. Becker-Feigeles, J. Halper, and M. A. Picone. A qualitative investigation of adaptation in older individuals with multiple sclerosis. *Disability and Rehabilitation*, 30(15):1088–1097, 2008.
- [71] L. Dybkjaer, N. O. Bernsen, and W. Minker. Evaluation and usability of multimodal spoken language dialogue systems. *Speech Communication*, 43(1):33–54, 2004.
- [72] D. Ellis. A behavioural approach to information retrieval system design. *Journal of Documentation*, 45(3):171–212, 1989.
- [73] D. Ellis. The dilemma of measurement in information retrieval research. *Journal of the American Society for Information Science*, 47(1):23–36, 1996.
- [74] A. Field, J. Miles, and Z. Field. *Discovering Statistics Using R*. SAGE Publications, 2012.
- [75] M. W. Firmin. Themes. In L. Given, editor, *The SAGE Encyclopedia of Qualitative Research Methods*, page 869. SAGE Publications, 2008.
- [76] N. M. Fraser and G. N. Gilbert. Simulating speech systems. *Computer Speech and Language*, 5(1):81–99, 1991.
- [77] N. Gao, D. W. Oard, and M. Dredze. Support for interactive identification of mentioned entities in conversational speech. In *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*, pages 953–956, 2017.
- [78] D. Gibbon, R. Moore, and R. Winski. *Handbook of Standards and Resources for Spoken Language Systems*. de Gruyter, 1997.
- [79] R. Gilabert, J. Barón, and À. Llanes. Manipulating cognitive complexity across task types and its impact on learners’ interaction during oral performance. *International Review of Applied Linguistics in Language Teaching*, 47(3-4):367–395, 2009.
- [80] J. Ginzburg. Interrogatives: Questions, facts and dialogue. In S. Lappin, editor, *The Handbook of Contemporary Semantic Theory*, pages 385–422. Blackwell, Oxford, 1996.
- [81] K. Go and J. M. Carroll. The blind men and the elephant: Views of scenario-based system design. *ACM Interactions*, 11(6):44–53, 2004.
- [82] A. Gorin, G. Riccardi, and J. Wright. How may I help you? *Speech Communication*, 23(1):113–127, 1997.

- [83] J. D. Gould, J. Conti, and T. Hovanyecz. Composing letters with a simulated listening typewriter. *Communications of the ACM*, 26(4):295–308, 1983.
- [84] D. Griol, J. Carbo, and J. M. Molina. Bringing context-aware access to the web through spoken interaction. *Applied Intelligence*, 38(4):620–640, 2013.
- [85] M. Gupta and M. Bendersky. Information retrieval with verbose queries. *Foundations and Trends in Information Retrieval*, 9(3-4):209–354, 2015.
- [86] I. Guy. Searching by talking: Analysis of voice queries on mobile web search. In *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*, pages 35–44, 2016.
- [87] I. Guy. The characteristics of voice search: Comparing spoken with typed-in mobile web search queries. *ACM Transactions on Information Systems (TOIS)*, 36(3):30:1–30:28, 2018.
- [88] E. Hagen. An approach to mixed initiative spoken information retrieval dialogue. *User Modeling and User-Adapted Interaction*, 9(1):167–213, 1999.
- [89] A. Hassan, R. W. White, S. T. Dumais, and Y.-M. Wang. Struggling or exploring?: Disambiguating long search sessions. In *Proceedings of Web Search and Data Mining (WSDM)*, pages 53–62, 2014.
- [90] H. Hastie. Metrics and evaluation of spoken dialogue systems. In *Data-Driven Methods for Adaptive Spoken Dialogue Systems*, pages 131–150. Springer, 2012.
- [91] M. Hearst. *Search User Interfaces*. Cambridge University Press, 2009.
- [92] R. Higashinaka, K. Imamura, T. Meguro, C. Miyazaki, N. Kobayashi, H. Sugiyama, T. Hirano, T. Makino, and Y. Matsuo. Towards an open-domain conversational system fully based on natural language processing. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 928–939, 2014.
- [93] A. M. Hills. *Foolproof Guide to Statistics Using IBM SPSS*. Pearson, 2011.
- [94] J. Huang, R. W. White, and S. Dumais. No clicks, no problem: Using cursor movements to understand and improve search. In *Proceedings of Conference on Human Factors in Computing Systems (CHI)*, pages 1225–1234, 2011.
- [95] P. Ingwersen and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer, 2005.

- [96] B. J. Jansen and A. Spink. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management*, 42(1):248–263, 2006.
- [97] B. J. Jansen, D. Booth, and B. Smith. Using the taxonomy of cognitive learning to model online searching. *Information Processing & Management*, 45(6):643–663, 2009.
- [98] C. Jenkins, C. L. Corritore, and S. Wiedenbeck. Patterns of information seeking on the web: A qualitative study of domain expertise and web expertise. *IT & Society*, 1(3):64–89, 2003.
- [99] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*, pages 154–161, 2005.
- [100] G. Jones. *An Introduction to Crowdsourcing for Language and Multimedia Technology Research*, pages 132–154. Springer, 2013.
- [101] F. Jurčiček, S. Keizer, M. Gašić, F. Mairesse, B. Thomson, K. Yu, and S. Young. Real user evaluation of spoken dialogue systems using Amazon Mechanical Turk. In *Proceedings of INTERSPEECH*, pages 3061–3064, 2011.
- [102] M. Kaisser, M. A. Hearst, and J. B. Lowe. Improving search results quality by customizing summary lengths. *Proceedings of Association for Computational Linguistics (ACL)*, pages 701–709, 2008.
- [103] T. Kanungo and D. Orr. Predicting the readability of short web summaries. In *Proceedings of Web Search and Data Mining (WSDM)*, pages 202–211, 2009.
- [104] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137, 2015.
- [105] M. Kellar, C. Watters, and M. Shepherd. A field study characterizing web-based information-seeking tasks. *Journal of the Association for Information Science and Technology (JASIST)*, 58(7):999–1018, 2007.
- [106] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1–2):1–224, 2009.
- [107] D. Kelly and L. Azzopardi. How many results per page?: A study of serp size, search behavior and user experience. In *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*, pages 183–192, 2015.

- [108] D. Kelly, J. Arguello, A. Edwards, and W.-c. Wu. Development and Evaluation of Search Tasks for IIR Experiments using a Cognitive Complexity Framework. In *Proceedings of International Conference on the Theory of Information Retrieval (ICTIR)*, pages 101–110, 2015.
- [109] M. A. Khawaja, F. Chen, and N. Marcus. Measuring cognitive load using linguistic features: implications for usability evaluation and adaptive interaction design. *International Journal of Human-Computer Interaction*, 30(5):343–368, 2014.
- [110] H. Khouzaimi, R. Laroche, and F. Lefevre. Turn-taking phenomena in incremental dialogue systems. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1890–1895, 2015.
- [111] J. Kiesel, A. Bahrami, B. Stein, A. Anand, and M. Hagen. Toward voice query clarification. In *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1257–1260, 2018.
- [112] J. Kim, G. Chern, D. Feng, E. Shaw, and E. Hovy. Mining and assessing discussions on the web through speech act analysis. In *Proceedings of the Workshop on Web Content Mining with Human Language Technologies at the 5th International Semantic Web Conference*, 2006. 10 pages.
- [113] J. Kim, J. R. Trippas, M. Sanderson, Z. Bao, and W. B. Croft. How do computer scientists use Google Scholar?: A survey of user interest in elements on SERPs and author profile pages. In *Proceedings of the 8th Workshop on Bibliometric-enhanced Information Retrieval (BIR 2019)*. *CEUR-WS*, pages 64–75, 2019.
- [114] S. R. Klemmer, A. K. Sinha, J. Chen, J. A. Landay, N. Aboobaker, and A. Wang. Suede: a Wizard of Oz prototyping tool for speech user interfaces. In *Proceedings of User Interface Software and Technology (UIST)*, pages 1–10, 2000.
- [115] B. P. Knijnenburg, M. C. Willemsen, and A. Kobsa. A pragmatic procedure to support the user-centric evaluation of recommender systems. In *Proceedings of RecSys*, pages 321–324, 2011.
- [116] C. C. Kuhlthau. Developing a model of the library search process: Cognitive and affective aspects. *RQ*, 28(2):232–242, 1988.
- [117] J. Lai and N. Yankelovich. *Speech Interface Design*, pages 764–770. Elsevier, 2006.
- [118] J. Lai, C. Karat, and N. Yankelovich. Conversational speech interfaces and technologies. *Human-Computer Interaction: Design Issues, Solutions, and Applications*, pages 53–63, 2009.

- [119] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174, 1977.
- [120] M. Larson and G. J. F. Jones. Spoken content retrieval: A survey of techniques and technologies. *Foundations and Trends in Information Retrieval*, 5(4–5):235–422, 2012.
- [121] H. Lausberg and H. Sloetjes. Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods*, 41(3):841–849, 2009.
- [122] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in web search. In *Proceedings of World Wide Web Conference (WWW)*, pages 391–400, 2005.
- [123] J. Liono, J. R. Trippas, D. Spina, M. S. Rahaman, Y. Ren, F. D. Salim, M. Sanderson, F. Scholer, and R. W. White. Building a benchmark for task progress in digital assistants. In *Proceedings of WSDM’19 Task Intelligence Workshop (TI@WSDM19)*, 2019. 6 pages.
- [124] D. J. Litman and S. Pan. Designing and evaluating an adaptive spoken dialogue system. *User Modeling and User-Adapted Interaction*, 12(2):111–137, 2002.
- [125] J. Liu, M. J. Cole, C. Liu, R. Bierig, J. Gwizdka, N. J. Belkin, J. Zhang, and X. Zhang. Search behaviors in different task types. In *Proceedings of Joint Conference on Digital Libraries (JDCL)*, pages 69–78, 2010.
- [126] G. Marchionini. *Information Seeking in Electronic Environments*. Cambridge University Press, 1997.
- [127] G. Marchionini and R. White. Find what you need, understand what you find. *International Journal of Human-Computer Interaction*, 23(3):205–237, 2007.
- [128] F. J. Massey. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.
- [129] D. Maxwell, L. Azzopardi, and Y. Moshfeghi. A study of snippet length and informativeness: Behaviour, performance and user experience. In *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*, pages 135–144, 2017.
- [130] D. McDuff, P. Thomas, M. Czerwinski, and N. Craswell. Multimodal analysis of vocal collaborative search: A public corpus and results. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI’17)*, pages 456–463, 2017.

- [131] E. McLellan, K. M. MacQueen, and J. L. Neidig. Beyond the Qualitative Interview: Data Preparation and Transcription. *Field Methods*, 15(1):63–84, 2003.
- [132] M. McTear, Z. Callejas, and D. Griol. *The Conversational Interface*. Springer, 2016.
- [133] M. F. McTear. Spoken dialogue technology: enabling the conversational user interface. *ACM Computing Surveys*, 34(1):90–169, 2002.
- [134] A. Moffat, P. Bailey, F. Scholer, and P. Thomas. Assessing the cognitive complexity of information needs. In *Proceedings of the Australasian Document Computing Symposium (ADCS)*, pages 97–100, 2014.
- [135] G. B. Newby. Cognitive space and information space. *Journal of the American Society for Information Science and Technology*, 52(12):1026–1048, 2001.
- [136] J. Nielsen. Ten usability heuristics. <https://www.nngroup.com/articles/ten-usability-heuristics/>, 2005. (Last accessed January 2, 2019).
- [137] R. Nordlie. “User revelation” - a comparison of initial queries and ensuing question development in online searching and in human reference interactions. In *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*, pages 11–18, 1999.
- [138] D. W. Oard. Query by babbling: A research agenda. In *Proceedings of International Conference on Information and Knowledge Management (CIKM)*, pages 17–21, 2012.
- [139] R. N. Oddy. Information retrieval through man-machine dialogue. *Journal of Documentation*, 33(1):1–14, 1977.
- [140] D. Odijk, R. W. White, A. Hassan Awadallah, and S. T. Dumais. Struggling and success in web search. In *Proceedings of International Conference on Information and Knowledge Management (CIKM)*, pages 1551–1560, 2015.
- [141] K. Ong, K. Järvelin, M. Sanderson, and F. Scholer. Qwerty: The effects of typing on web search behavior. In *Proceedings of Conference on Information Interaction and Retrieval (CHIIR)*, pages 281–284, 2018.
- [142] T. Pääkkönen, J. Kekäläinen, H. Keskustalo, L. Azzopardi, D. Maxwell, and K. Järvelin. Validating simulated interaction for retrieval evaluation. *Information Retrieval Journal*, 20(4):338–362, 2017.
- [143] T. Pica. Research on negotiation: What does it reveal about second-language learning conditions, processes, and outcomes? *Language Learning*, 44(3):493–527, 1994.

- [144] C. Qu, L. Yang, W. B. Croft, J. R. Trippas, Y. Zhang, and M. Qiu. Analyzing and characterizing user intent in information-seeking conversations. In *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*, pages 989–992, 2018.
- [145] C. Qu, L. Yang, W. B. Croft, Y. Zhang, J. R. Trippas, and M. Qiu. User intent prediction in information-seeking conversations. In *Proceedings of Conference on Information Interaction and Retrieval (CHIIR)*, pages 25–33, 2019.
- [146] F. Radlinski and N. Craswell. A theoretical framework for conversational search. In *Proceedings of Conference on Information Interaction and Retrieval (CHIIR)*, pages 117–126, 2017.
- [147] N. Reithinger and E. Maier. Utilizing statistical dialogue act processing in Verbomobil. In *Proceedings of Association for Computational Linguistics (ACL)*, pages 116–121, 1995.
- [148] P. D. Reynolds. *Primer in Theory Construction: An A&B Classics Edition*. Routledge, 2015.
- [149] V. Rieser, O. Lemon, and X. Liu. Optimising information presentation for spoken dialogue systems. In *Proceedings of Association for Computational Linguistics (ACL)*, pages 1009–1018, 2010.
- [150] J. Ritchie, J. Lewis, C. M. Nicholls, and R. Ormston. *Qualitative Research Practice: A Guide for Social Science Students and Researchers*. SAGE Publications, 2013.
- [151] P. Robinson. Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied linguistics*, 22(1):27–57, 2001.
- [152] P. Robinson. The cognitive hypothesis, task design, and adult task-based language learning. *The Second Language Studies*, 21(2):4–105, 2003.
- [153] P. Robinson. Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *IRAL-International Review of Applied Linguistics in Language Teaching*, 45(3):193–213, 2007.
- [154] P. M. Rothbauer. Triangulation. In L. Given, editor, *The SAGE Encyclopedia of Qualitative Research Methods*, pages 893–894. SAGE Publications, 2008.
- [155] I. Ruthven. Interactive information retrieval. *Annual Review of Information Science and Technology*, 42(1):43–91, 2008.

- [156] N. G. Sahib, D. Al Thani, A. Tombros, and T. Stockman. Accessible information seeking. *Proceedings of Digital Futures*, 12:1–3, 2012.
- [157] N. G. Sahib, A. Tombros, and T. Stockman. A comparative analysis of the information-seeking behavior of visually impaired and sighted searchers. *Journal of the Association for Information Science and Technology (JASIST)*, 63(2): 377–391, 2012.
- [158] T. Saracevic. The stratified model of information retrieval interaction: Extension and applications. In *Proceedings of the Annual Meeting-American Society for Information Science*, volume 34, pages 313–327, 1997.
- [159] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope. “Your word is my command”: Google search by voice: a case study. In *Advances in Speech Recognition*, pages 61–90. Springer, 2010.
- [160] R. Schaller, M. Harvey, and D. Elsweler. Out and about on museums night: Investigating mobile search behaviour for leisure events. In *Searching4Fun workshop at ECIR2012*, 2012. 4 pages.
- [161] E. A. Schegloff. Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, 29(1):1–63, 2000.
- [162] E. A. Schegloff, G. Jefferson, and H. Sacks. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382, 1977.
- [163] D. Schiffrin. Conversational coherence: The role of well. *Language*, 61(3):640–667, 1985.
- [164] J. R. Searle. *Speech Acts: An Essay in the Philosophy of Language*, volume 626. Cambridge University Press, 1969.
- [165] C. E. Shannon. *The Mathematical Theory of Communication*. University of Illinois Press, 1949.
- [166] S. Shiga, H. Joho, R. Blanco, J. R. Trippas, and M. Sanderson. Modelling information needs in collaborative search conversations. In *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*, pages 715–724, 2017.
- [167] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. Analysis of a very large web search engine query log. *ACM SIGIR Forum*, 33(1):6–12, 1999.

- [168] S. Sitter and A. Stein. Modeling the illocutionary aspects of information-seeking dialogues. *Information Processing & Management*, 28(2):165–180, 1992.
- [169] G. Skantze. *Error Handling in Spoken Dialogue Systems-Managing Uncertainty, Grounding and Miscommunication*. PhD thesis, KTH, Stockholm, 2007.
- [170] D. Spina, J. R. Trippas, L. Cavedon, and M. Sanderson. SpeakerLDA: Discovering topics in transcribed multi-speaker audio contents. In *Proceedings of the Third Edition Workshop on Speech, Language & Audio in Multimedia (SLAM)*, pages 7–10, 2015.
- [171] D. Spina, J. R. Trippas, L. Cavedon, and M. Sanderson. Extracting audio summaries to support effective spoken document search. *Journal of the Association for Information Science and Technology (JASIST)*, 68(9):2101–2115, 2017.
- [172] D. Spina, J. Arguello, H. Joho, J. Kiseleva, and F. Radlinski. CAIR’18: Second International Workshop on Conversational Approaches to Information Retrieval at SIGIR 2018. *ACM SIGIR Forum*, 52(2):111–116, 2019.
- [173] A. Stein and E. Maier. Structuring collaborative information-seeking dialogues. *Knowledge-Based Systems*, 8(2-3):82–93, 1995.
- [174] A. Stent. Rhetorical structure in dialog. In *Proceedings of the First International Conference on Natural language Generation*, pages 247–252, 2000.
- [175] A. Stent and J. Allen. Annotating argumentation acts in spoken dialog. Technical report, University of Rochester, Rochester, NY, 2000.
- [176] D. F. Stroup, J. A. Berlin, S. C. Morton, I. Olkin, G. D. Williamson, D. Rennie, D. Moher, B. J. Becker, T. A. Sipe, S. B. Thacker, and for the Meta-analysis Of Observational Studies in Epidemiology (MOOSE) Group. Meta-analysis of observational studies in epidemiology: A proposal for reporting. *JAMA*, 283(15):2008–2012, 2000.
- [177] P.-H. Su, N. Mrkšić, I. Casanueva, and I. Vulić. Deep learning for conversational AI. In *Proceedings of Association for Computational Linguistics (ACL)*, pages 27–32, 2018.
- [178] H. Sugiyama, T. Meguro, R. Higashinaka, and Y. Minami. Open-domain utterance generation for conversational dialogue systems using web-scale dependency structures. In *Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 334–338, 2013.
- [179] R. S. Taylor. The process of asking questions. *American Documentation*, 13(4):391–396, 1962.

- [180] J. Teevan, C. Alvarado, M. S. Ackerman, and D. R. Karger. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *Proceedings of Conference on Human Factors in Computing Systems (CHI)*, pages 415–422, 2004.
- [181] J. Teevan, E. Cutrell, D. Fisher, S. M. Drucker, G. Ramos, P. André, and C. Hu. Visual snippets: Summarizing web pages for search and revisitation. In *Proceedings of Conference on Human Factors in Computing Systems (CHI)*, pages 2023–2032, 2009.
- [182] J. Teevan, S. T. Dumais, and E. Horvitz. Potential for personalization. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 17(1):4, 2010.
- [183] J. Teevan, D. Ramage, and M. R. Morris. #TwitterSearch: A comparison of microblog search and web search. In *Proceedings of Web Search and Data Mining (WSDM)*, pages 35–44, 2011.
- [184] Telephone Speech Collection Group. Transcription Guidelines (NQTR). Technical report, Linguistic Data Consortium, 2006. URL https://catalog.ldc.upenn.edu/docs/LDC2010S01/trans_guide_nqrt_span.doc. (Last accessed May 5, 2019).
- [185] P. Thomas, D. McDuff, M. Czerwinski, and N. Craswell. MISC: A data set of information-seeking conversations. In *SIGIR 1st International Workshop on Conversational Approaches to Information Retrieval (CAIR'17)*, 2017. 6 pages.
- [186] P. Thomas, M. Czerwinski, D. McDuff, N. Craswell, and G. Mark. Style and alignment in information-seeking conversation. In *Proceedings of Conference on Information Interaction and Retrieval (CHIIR)*, pages 42–51, 2018.
- [187] K. Tracy and N. Coupland. Multiple goals in discourse: An overview of issues. *Journal of Language and Social Psychology*, 9(1-2):1–13, 1990.
- [188] D. R. Traum. Computational models of grounding in collaborative systems. In *Psychological Models of Communication in Collaborative Systems-Papers from the AAAI Fall Symposium*, pages 124–131, 1999.
- [189] D. R. Traum. Speech acts for dialogue agents. In *Foundations of Rational Agency*, pages 169–201. Springer, 1999.
- [190] D. R. Traum and S. Larsson. The information state approach to dialogue management. In *Current and New Directions in Discourse and Dialogue*, pages 325–353. Springer, 2003.

- [191] J. R. Trippas. Spoken conversational search: Information retrieval over a speech-only communication channel. In *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*, page 1067, 2015.
- [192] J. R. Trippas. Spoken conversational search: Speech-only interactive information retrieval. In *Proceedings of Conference on Information Interaction and Retrieval (CHIIR)*, pages 373–375, 2016.
- [193] J. R. Trippas and P. Thomas. Data sets for spoken conversational search. In *Proceedings of the CHIIR 2019 Workshop on Barriers to Interactive IR Resources Re-use (BIIRRR 2019). CEUR-WS*, pages 14–18, 2019.
- [194] J. R. Trippas, D. Spina, M. Sanderson, and L. Cavedon. Results presentation methods for a spoken conversational search system. In *CIKM'15 First International Workshop on Novel Web Search Interfaces and Systems (NWSearch'15)*, pages 13–15, 2015.
- [195] J. R. Trippas, D. Spina, M. Sanderson, and L. Cavedon. Towards understanding the impact of length in web search result summaries over a speech-only communication channel. In *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*, pages 991–994, 2015.
- [196] J. R. Trippas, D. Spina, L. Cavedon, and M. Sanderson. Crowdsourcing user preferences and query judgements for speech-only search. In *SIGIR 1st International Workshop on Conversational Approaches to Information Retrieval (CAIR'17)*, 2017. 3 pages.
- [197] J. R. Trippas, D. Spina, L. Cavedon, and M. Sanderson. How do people interact in conversational speech-only search tasks: A preliminary analysis. In *Proceedings of Conference on Information Interaction and Retrieval (CHIIR)*, pages 325–328, 2017.
- [198] J. R. Trippas, D. Spina, L. Cavedon, and M. Sanderson. A conversational search transcription protocol and analysis. In *SIGIR 1st International Workshop on Conversational Approaches to Information Retrieval (CAIR'17)*, 2017. 5 pages.
- [199] J. R. Trippas, D. Spina, L. Cavedon, H. Joho, and M. Sanderson. Informing the design of spoken conversational search. In *Proceedings of Conference on Information Interaction and Retrieval (CHIIR)*, pages 32–41, 2018.
- [200] J. R. Trippas, D. Spina, F. Scholer, A. H. Awadallah, P. Bailey, P. N. Bennett, R. W. White, J. Liono, Y. Ren, F. D. Salim, and M. Sanderson. Learning about work tasks to inform intelligent assistant design. In *Proceedings of Conference on Information Interaction and Retrieval (CHIIR)*, pages 5–14, 2019.

- [201] J. R. Trippas, D. Spina, P. Thomas, H. Joho, M. Sanderson, and L. Cavedon. Towards a model for spoken conversational search. *Information Processing & Management*, 2019. (Submitted).
- [202] G. Tur and L. Deng. *Intent Determination and Spoken Utterance Classification*, pages 93–118. John Wiley & Sons, 2011.
- [203] M. Turunen, J. Hakulinen, N. Rajput, and A. A. Nanavati. *Evaluation of Mobile and Pervasive Speech Applications*, pages 219–262. John Wiley & Sons, 2012.
- [204] K. Tuuri, T. Eerola, and A. Pirhonen. Design and evaluation of prosody-based non-speech audio feedback for physical training application. *International Journal of Human-Computer Studies*, 69(11):741–757, 2011.
- [205] S. Vakulenko, K. Revoredo, C. Di Ciccio, and M. de Rijke. QRFA: A data-driven model of information-seeking dialogues. In *Proceedings of the European Conference on Information Retrieval (ECIR)*, pages 541–557, 2019.
- [206] S. Vargas, F. Weng, and H. Pon-Barry. Interactive question answering and constraint relaxation in spoken dialogue systems. In *Proceedings of Association for Computational Linguistics (ACL)*, pages 28–35, 2006.
- [207] A. Vtyurina, D. Savenkov, E. Agichtein, and C. L. Clarke. Exploring conversational search with humans, assistants, and wizards. In *Proceedings of Conference on Human Factors in Computing Systems (CHI)*, pages 2187–2193, 2017.
- [208] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella. PARADISE: A framework for evaluating spoken dialogue agents. In *Proceedings of European Chapter of the Association for Computational Linguistics (EACL)*, pages 271–280, 1997.
- [209] M. A. Walker, R. Passonneau, and J. E. Boland. Quantitative and qualitative evaluation of darpa communicator spoken dialogue systems. In *Proceedings of Association for Computational Linguistics (ACL)*, pages 515–522, 2001.
- [210] M. Weiser. The computer for the 21st century. *Scientific American*, 265(3):94–105, 1991.
- [211] F. Weng, L. Cavedon, B. Raghunathan, D. Mirkovic, H. Cheng, H. Schmidt, H. Bratt, R. Mishra, S. Peters, S. Upson, E. Shriberg, C. Bergmann, and L. Zhao. A conversational dialogue system for cognitively overloaded users. In *Proceedings of INTERSPEECH*, pages 233–236, 2004.
- [212] R. W. White. *Interactions with Search Systems*. Cambridge University Press, 2016.

- [213] B. Wildemuth, L. Freund, and E. G. Toms. Untangling search task complexity and difficulty in the context of interactive information retrieval studies. *Journal of Documentation*, 70(6):1118–1140, 2014.
- [214] M. L. Wilson, N. Alsuraykh, and H. A. Maior. Measuring mental workload in IIR user studies with fNIRS. In *NeuroIIR'17*, 2017. 2 pages.
- [215] T. D. Wilson. Models in information behaviour research. *Journal of Documentation*, 55(3):249–270, 1999.
- [216] A. K. Winterboer, M. I. Tietze, M. K. Wolters, and J. D. Moore. The user model-based summarize and refine approach improves information presentation in spoken dialog systems. *Computer Speech & Language*, 25(2):175–191, 2011.
- [217] D. L. Worthington and G. D. Bodie. *The Sourcebook of Listening Research: Methodology and Measures*. John Wiley & Sons, 2017.
- [218] W.-C. Wu, D. Kelly, A. Edwards, and J. Arguello. Grannies, tanning beds, tattoos and NASCAR. In *Proceedings of International Conference on Information Interaction in Context (IIiX)*, pages 254–257, 2012.
- [219] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke. The microsoft 2017 conversational speech recognition system. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5934–5938, 2018.
- [220] L. Yang, M. Qiu, C. Qu, J. Guo, Y. Zhang, W. B. Croft, J. Huang, and H. Chen. Response ranking with deep matching networks and external knowledge in information-seeking conversation systems. In *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*, pages 245–254, 2018.
- [221] N. Yankelovich. Using natural dialogs as the basis for speech interface design. In *Human Factors and Voice Interactive Systems*, pages 255–290. Springer, 2008.
- [222] N. Yankelovich and J. Lai. Designing speech user interfaces. In *Proceedings of Conference on Human Factors in Computing Systems (CHI)*, pages 131–132, 1998.
- [223] N. Yankelovich and J. Lai. Designing speech user interfaces. In *Proceedings of Conference on Human Factors in Computing Systems (CHI)*, pages 124–125, 1999.
- [224] N. Yankelovich, G. Levow, and M. Marx. Designing speechacts: Issues in speech user interfaces. In *Proceedings of Conference on Human Factors in Computing Systems (CHI)*, pages 369–376, 1995.

- [225] J. Yi and F. Maghoul. Mobile search pattern evolution. In *Proceedings of World Wide Web Conference (WWW)*, pages 165–166, 2011.
- [226] E. Yulianti, R.-C. Chen, F. Scholer, W. B. Croft, and M. Sanderson. Ranking documents by answer-passage quality. In *Proceedings of Conference on Research and Development in Information Retrieval (SIGIR)*, pages 335–344, 2018.
- [227] E. Zarisheva and T. Scheffler. Dialog act annotation for twitter conversations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 114–123, 2015.
- [228] Y. Zhang, X. Chen, Q. Ai, L. Yang, and W. B. Croft. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of International Conference on Information and Knowledge Management (CIKM)*, pages 177–186, 2018.
- [229] I. Zuckerman, E. Segal-Halevi, A. Rosenfeld, and S. Kraus. First steps in chat-based negotiating agents. In *Next Frontier in Agent-based Complex Automated Negotiation*, pages 89–109. Springer, 2015.