Data Sets for Spoken Conversational Search

Johanne Trippas
RMIT University
Melbourne, Australia
johanne.trippas@rmit.edu.au

Paul Thomas
Microsoft
Canberra, Australia
pathom@microsoft.com

ABSTRACT

There is increasing interest in spoken conversational search—multi-turn interactions with a search engine, spoken in natural language—but until recently there was little public data to support research.

We describe our experiences building two data sets for spoken conversational search: the Microsoft Information-Seeking Conversation set ("MISC") and the Spoken Conversational Search set ("SCSdata"). Each data set contains recordings of spoken interactions between two people collaborating on web search tasks, but relatively small differences in protocol have led to observably different data. We discuss some consequences of these differences, and describe attempts to reproduce analyses from one set to the other.

1 DATA SETS OVERVIEW

The increasing capability for natural-language, voice interactions with computers poses a range of research and engineering questions. To address these questions we need corresponding data—for example, recordings of conversations with information-gathering agents. Unfortunately, current systems cannot maintain a lengthy exchange, have trouble tracking context, and are largely unaware of non-verbal communication and of users' emotional state. In 2016–17 two separate groups tried to bridge the gap by recording information-seeking conversations between people, looking for structures which would help build new systems or evaluate old ones [c.f. 5, 9, 19].

1.1 MISC

The Microsoft Information-Seeking Conversation data (MISC) is a set of recordings of spoken conversation between human "seekers" and "intermediaries" [21]. 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 is played by politeness or other conversational norms?; what tactics do we see in information-seeking conversation, and do particular structures help or impede progress or satisfaction? MISC has been used in unpublished work on these questions, in work on conversational style [20], on multimodal collaboration [14], and on conversational structures described below.

The study. The overall setup for both the MISC and SCSdata recordings is shown in Figure 1. Tasks were assigned to a "seeker", who was responsible for assembling information and deciding a final answer. They were connected over an audio link to an "intermediary", who stood in for a future software agent (SCSdata participants were located in the same room). The intermediary had unrestricted access to the web, including search engines. We recorded video and audio from both participants.

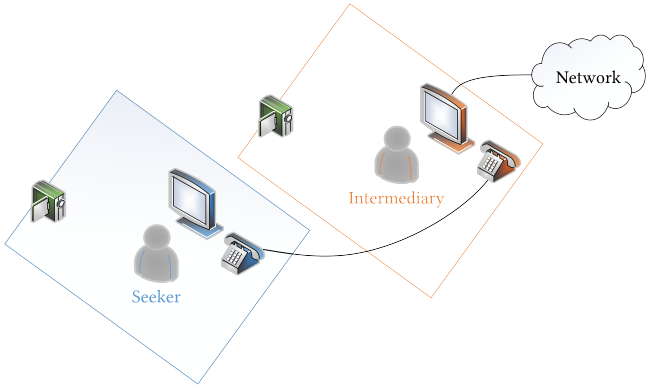

Figure 1: Recording setup for both MISC and SCSdata. Tasks were assigned to a "seeker", who communicated with an "intermediary" who had access to a browser. From Thomas et al. [21].

The data. 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.

Reuse and reusability. We designed the MISC data with regard to our own future research, but intended from the start that it could be used by other researchers. Our participants consented to possible reuse and sharing, and were informed of their right to withdraw consent at any time, including post-hoc. The study was approved by our internal ethics review board.

Although MISC includes a good deal of derived data, we have chosen to include the raw data wherever possible so as to enable (a) replication and (b) further unanticipated analyses. For example, we include the raw audio, from which we derived the included transcripts; and we include these transcripts, from which we derived data on word use. The only processing of the "raw" video and audio has been to segment by task. The full text of each pre-experiment and post-task question is also included. This policy has already enabled reuse inside our research group: for example, work by McDuff et al. [14], on the effect of facial expressivity and multimodel communication, was not anticipated when we collected MISC. We are not aware of any attempts to re-process the audio or video

Workshop on Barriers to Interactive IR Resources Re-use at the ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR 2019), 14 March 2019, Glasgow, UK

streams, but we hope this policy also makes reuse outside our own research group more likely.

We used standard instruments and standard processing tools where available:

- To help interpret physiological and affectual signals, we used the UPSS Impulsive Behaviour Scale [27] and Cohen et al.'s perceived stress scale [8][1]. These are commonly-used instruments and should be comparable across studies.
- MISC includes five tasks, one of which was used as a warmup. We believed there may be a difference in behaviour and self-reports depending on the complexity and difficulty of the task, so we varied these in a controlled manner. We also wanted tasks that elicited an emotional response, which ruled out those from most past collections; instead we selected tasks from the Repository of Assigned Search Tasks (RepAST)[2]. Participants addressed the tasks using the open web, which may make it hard to reproduce some results but did allow intermediaries to use the full range of web search features.
- We measured effort with the NASA task load index (TLX) [16]. This is a commonly-used and well-validated scale which we were able to adopt with minimal modification (only omitting the question on physical effort). Post-hoc tests validated this modified scale (Krippendorf's $\alpha = 0.84$ [21]).
- We measured engagement using a subset of the user engagement scale (UES) [17]. This proved very useful for our purposes, and again the modifications were validated post hoc ($\alpha = 0.85$ [21]).
- Questions on per-task emotion used a widely recognised set of basic emotions, as well as a separate question about other emotions which we considered more likely during our tasks.
- Processing used standard tools, both to reduce effort and to aid reproducability. We used OpenSMILE [11] for audio analysis; OpenFace [3] for coding facial actions; Microsoft Cognitive Services[3] to produce transcripts; and Linguistic Inquiry and Word Count (LIWC) [18] for lexical analysis.

We are happy to make many of our processing scripts available for other researchers—a small number use in-house tools—although again there have been no requests so far.

Reporting and availability. We described our protocol in detail in our first publication [21]. This paper includes details of participants, the wording for all tasks and questions, and descriptive statistics including reliability measures.

The MISC data is available online at http://aka.ms/MISCv1.

1.2 SCSdata

The Spoken Conversational Search data set (SCSdata) contains the utterance transcriptions of a spoken information seeking process between two actors. To the best of our knowledge, SCSdata was the first data set which was created in this experimental setup. It is also the first SCS data set which received labelling of the actions or utterances, albeit only for the first three turns [22]. However, the release of the fully labelled data set is planned.

[1]See also e.g. http://www.mindgarden.com/documents/PerceivedStressScale.pdf.
[2]https://ils.unc.edu/searchtasks/
[3]https://www.microsoft.com/cognitive-services/en-us/speech-api

The SCSdata was created to investigate the interaction behaviour between the two actors, including helping us to understand questions such as; what is the impact of audio-only interactions for search?; how are information-dense documents transferred in an audio-only setting?; what are the components or actions of an information-seeking process via audio, and what is the impact of query complexity on the interactions and interactivity in spoken conversational search? The SCSdata has been used in research published by the creators of the data set [22, 23] and also has been used recently in a study by the broader IR community [25].

The study. The SCSdata was created in a controlled laboratory study at RMIT University. We recorded the spoken interactions between seeker and intermediary (as explained in Section 1.1). We then transcribed the recordings with transcription principles and protocols described by Trippas et al. [24]. Much detailed work went into creating highly accurate transcriptions, with the vision to increase the reusability of the data set, including indexing [13].

The data. The data includes the transcriptions of the audio signals, the codebook and labels for the first three utterances, and the backstories used in the setup. Other data such as the audio, video, pre- and post-task questionnaires are not available due to ethics regulations.

The data is maintained by an author of this paper (Trippas).

Reuse and reusability. The SCSdata reuses nine backstories based on TREC Q02, R03, and T04 as described by Bailey, Moffat, Scholer, and Thomas [2]. These backstories follow the cognitive complexity framework of the Taxonomy of Learning [1].

Participants completed a pre-test questionnaire before starting the study. This pre-test questionnaire gathered demographic data such as age, gender, highest level of education, employment, and computer and search engine usage. Participants were also asked to complete a modified version of the Search Self-Efficacy Scale [4] and how they would rate their own overall search skills. Participants were asked if they had experience with intelligent personal assistants such as Google Now, Siri, Amazon Alexa, or Cortana. Seekers and intermediaries were asked to complete pre- and post-task questionnaires throughout the study measuring interest and knowledge about the task, experienced task difficulty, experienced conversational difficulty, experienced collaboration difficulty, experienced search presentation difficulty, overall difficulty, overall satisfaction, and open questions. Some of these questions were adapted or reproduced from Kelly, Arguello, Edwards, and Wu [12].

The SCSdata was designed with our own research questions in mind, while optimising the transcriptions and labelling for future use. We believe that the labelled data set is very valuable for the research community. The data set was recently updated and it is planned to update the data set with the full labelling annotations and label creation methodology in the near future.

Reporting and availability. We described our experimental setup in the preliminary data analysis paper [22]. Fully documented information is available on the transcription protocol and labelling process in Trippas et al. [24]. That paper aimed to establish a protocol for spoken search interaction transcription, minimising the likelihood that consequently produced transcripts are inconsistent with each other.

Other details such as the procedure of the study or questionnaire results have not yet been published.

The SCSdata is available online via https://jtrippas.github.io/Spoken-Conversational-Search/.

2 COMPARING MISC AND SCSdata

In recent, unpublished work, one of us (Trippas) has developed a code schema for annotating utterances in spoken conversational search. Initial development used SCSdata, but since MISC is very similar it has been reused to validate the schema. We offer below some observations on re-using MISC and SCSdata, based on this experience.

It is clearly valuable to have two data sets collected with such similar protocols, and for similar purposes. Coding conversations relies on having lengthy, naturalistic exchanges, and both SCSdata and MISC have several exchanges running to ten minutes. Both sets distinguish the "seeker" and "intermediary" roles, allowing direct comparison, and both include transcripts which could be coded more or less directly. However, some differences across the data sets did hamper reuse, or led to unexpected findings.

2.1 Protocol differences

First, while the SCSdata was manually transcribed, the MISC data is about ten times larger but has only been transcribed with a commercial speech-to-text system. Although the automatic speech recognition (ASR) system was state of the art, it was still prone to errors. (One common error was to inject "speech" when a participant was typing, as if the ASR was confused by keyboard noise.) These errors were discovered because a close reading was needed to label the MISC utterances for the validation of an annotation schema. The difference in transcription techniques also gave a different notion of utterance or turn: in SCSdata these are divided manually, while in MISC they are separated by pauses in the audio signal. Utterance-level statistics may not be directly comparable.

The sets also differ in the pre- and post-task questions. The MISC questions and responses are part of the released data, and the published description includes descriptive statistics and basic validity checks. We hope this is useful for future work. The SCSdata protocol also added many pre- and post-task items (see section 1.2), on overlapping themes but with different instruments. These have not been examined to date so they may or may not be useful or comparable. Future SCSdata releases will not include this data.

Some apparently small differences between the SCSdata and MISC protocols have led to observable differences in the collected data. SCSdata participants were expressly prohibited from reading out the task statement verbatim and had to verbalise their information request; MISC participants were given no instruction on this matter. As a result, the MISC data include seekers reading out and repeating the task statements, verbatim. More importantly, once both participants have the same statement, the roles of "seeker" and "intermediary" are blurred and the two act much more like peers. This has influenced the interactions in MISC, and the distribution of conversational moves.

The two protocols also differed at the end of each task. For MISC, "seekers" were asked to record an answer: this was meant partly to encourage participants to properly complete each task, and partly so researchers could look for differences in answer correctness or completeness[4]. SCSdata participants were not asked to record an answer, but were asked to say "stop search" when they were satisfied with the found information and could answer the information need. This again led to differences in behaviour, such as MISC "seekers" confirming spelling in order to write down the answer.

These differences were an unexpected nuisance, as even with such similar protocols it required some work to understand and account for the substantial differences in data. However, familiarity with the data meant that once we had observed the differences, they were easy to understand. A close reading of the published descriptions would have given the same hints. Further, it is likely that the differences were in fact *useful* for the validation, as they gave more variety and tested the coding schema in slightly different exchanges.

We also note some smaller differences. For the SCSdata recordings, a researcher was in the room with the participant. The MISC researchers were not. This may have led to some differences in the data, although we have not yet explored this. There is also a difference in audio quality. The audio files from the SCSdata are poor, because they were recorded through a video camera. Using those recordings was never part of the experimental setup.

Finally, there are details of the protocol which may have resulted in minor differences between the sets. MISC featured a warm-up task, while SCSdata did not; MISC participants used a Windows PC, while SCSdata participants used a Mac; and MISC intermediaries started with Bing, SCSdata intermediaries with Google, although all were allowed to switch to any other site.

2.2 Terminology

There has been some inconsistency in terminology. First the two actors of the SCSdata were referred to as the "user" (the participant with the search task) and "retriever" (the participant with the search engine) [22, 24]. In later publications describing the SCSdata, "user" became "seeker" and "retriever" became "intermediary" [23]. These latter terms match MISC.

Other terminology is not standard. Trippas et al. used "spoken conversational search" to emphasise the spoken channel, as opposed to multi-turn interactions with e.g. typing or selecting buttons. For the same scenario, Thomas et al. used the phrase "information-seeking conversation" to encourage a broader understanding encompassing negotiation and clarification, not just a traditional query/response "search" model. Other terms again are used elsewhere in the literature. Presumably in the near future this terminology, as well as the names of the different roles, will be standardised.

2.3 Task design

As explained in section 1.2, the tasks used for the SCSdata were reused from research by Bailey et al. [2] and are based on the Taxonomy of Learning. Three of the five cognitive dimensions were used: *Remember*, *Understand*, and *Analyse*. However, it has been suggested that there are no clear interaction differences between Understand and Analyse tasks [22], which is consistent with the difficulties Moffat et al. reported when classifying tasks [15].

[4]In the event, we have not been able to code the answers with any degree of reliability.

Table 1: MISC search tasks. These were controlled for complexity, difficulty, and likely emotional response.

	Difficulty	Complexity	Emotion	Task source
0	Warm-up	(NA)	(NA)	Buhi et al. [7], via RepAST
1	Low	Low	Positive	Modified TREC topic 442
2	Low	High	Negative	Broussard and Zhang [6], via RepAST
3	High	Low	(NA)	Newly created
4	High	High	Positive	White [26], via RepAST

The MISC tasks were gathered from different sources and one task was created specifically for this study (Table 1). More specifically, the tasks used in MISC were chosen to elicit positive and negative emotions and were based on two different levels of difficulty and complexity as seen in Table 1. Since MISC uses only two levels, it would perhaps make sense to consider Understand and Analyse as high complexity, and Remember as low complexity, if task-to-task comparisons were needed. Alternatively, differences in interaction patterns may let us align tasks across the two sets. We have not yet explored these possibilities.

3 OBSERVATIONS

Two sets of spoken conversational searches—SCSdata and MISC—were collected independently, by different teams, in different geographical locations, to support different research. It is fortunate that the data sets are similar enough so that we can make direct comparisons, and use one set to verify observations from the other.

Despite being collected with very similar goals and methods, relatively small differences in protocol made observable differences to the data and we have had to be careful with reuse and comparisons. This was made much easier by our familiarity with the data; another researcher could quite reasonably choose these two data sets, compare them, and have difficulty. That this is possible despite careful design and description, and despite close similarity in protocol, may perhaps caution us about reuse in interactive studies generally.

We were however helped by the decision to explicitly allow the release of MISC's raw data (not just, e.g., transcripts). Because audio was available, the transcription errors could be detected. Unfortunately ethical clearance precludes a similar release for SCSdata, and this may limit reuse.

Communication between two people is very culture-specific [10]. Even though both MISC and SCSdata were collected in English speaking countries, and all participants claimed native or high-level English, we do not exclude that cultural differences played a role in the differences in the two data sets. Similarly, the difference in participant populations (more uniform in SCSdata, more varied in MISC) may have resulted in differences in communication.

Spoken conversational search is still an immature field of inquiry, and we should exercise some caution re-using data sets. Nuances of data collection are not always easy to describe in a paper, but the protocols for SCSdata and MISC were relatively simple and the data can be re-used with care. It has been interesting and informative to compare the two sets of transcripts, and we hope to continue this to investigate other conversational questions.

ACKNOWLEDGMENTS

We thank Daniel McDuff, Mary Czerwinski, and Nick Craswell for their effort assembling MISC, and Penny Analytis for auditing the SCSdata transcriptions. We are grateful to our participants for their time.

REFERENCES

[1] L. W. Anderson, D. R. Krathwohl, and B. S. Bloom. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman, New York.
[2] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2016. UQV100: A test collection with query variability. In *Proc. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*. 725–728.
[3] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: An open source facial behavior analysis toolkit. In *Proc. IEEE Winter Conf. Applications of Computer Vision*. 1–10.
[4] Kathy Brennan, Diane Kelly, and Yinglong Zhang. 2016. Factor analysis of a search self-efficacy scale. In *Proc. ACM SIGIR Conf. on Human Information Interaction and Retrieval*. 241–244.
[5] H. M. Brooks and N. J. Belkin. 1983. Using discourse analysis for the design of information retrieval interaction mechanisms. In *Proc. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*. 31–47.
[6] Ramona Broussard and Yan Zhang. 2013. Seeking treatment options: Consumers' search behaviors and cognitive activities. *J. American Society for Information Science and Technology* 50, 1 (2013), 1–10.
[7] Eric R. Buhi, Ellen M. Daley, Hollie J. Fuhrmann, and Sarah A. Smith. 2009. An observational study of how young people search for online sexual health information. *J American College Health* 58, 2 (2009), 101–111.
[8] Sheldon Cohen, Tom Kamarck, and Robin Mermelstein. 1983. A global measure of perceived stress. *J. Health and Social Behavior* 24, 4 (Dec. 1983), 385–396.
[9] Penny J. Daniels, H. M. Brooks, and N. J. Belkin. 1985. Using problem structures for driving human-computer dialogues. In *RIAO-85: Actes: Recherche d'Informations Assistée par Ordinateur*. 645–660.
[10] Birgit Endrass, Matthias Rehm, and Elisabeth André. 2009. Culture-specific communication management for virtual agents. In *Proc. Int. Confón Autonomous Agents and Multiagent Systems—Volume 1*. 281–287.
[11] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. In *Proc. ACM Multimedia*. 835–838.
[12] Diane Kelly, Jaime Arguello, Ashlee Edwards, and Wan-ching Wu. 2015. Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework. In *Proc. Int. Conf. on the Theory of Information Retrieval*. 101–110.
[13] Martha Larson and Gareth JF Jones. 2012. Spoken content retrieval: A survey of techniques and technologies. *Foundations and Trends® in Information Retrieval* 5, 4–5 (2012), 235–422.
[14] Daniel McDuff, Paul Thomas, Mary Czerwinski, and Nick Craswell. 2017. Multimodal analysis of vocal collaborative search: a public corpus and results. In *Proc. ACM Int. Conf. on Multimodal Interaction*. 456–463.
[15] Alistair Moffat, Peter Bailey, Falk Scholer, and Paul Thomas. 2014. Assessing the cognitive complexity of information needs. In *Proc. Australasian Document Computing Symposium*. ACM, 97–100.
[16] National Aeronautics and Space Administration Human Systems Integration Division. 2016. TLX @ NASA Ames. (2016). Retrieved January 2017 from https://humansystems.arc.nasa.gov/groups/TLX/
[17] Heather L O'Brien and Elaine G Toms. 2010. The development and evaluation of a survey to measure user engagement. *J American Society for Information Science and Technology* 61, 1 (2010), 50–69.
[18] James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report. University of Texas at Austin.
[19] Rachel Reichman. 1985. *Getting computers to talk like you and me*. MIT Press, Cambridge, Massachusetts.
[20] Paul Thomas, Mary Czerwinksi, Daniel McDuff, Nick Craswell, and Gloria Mark. 2018. Style and alignment in information-seeking conversation. In *Proc. ACM SIGIR Conf. on Human Information Interaction and Retrieval*. 42–51.

[21] Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2017. MISC: A data set of information-seeking conversations. In *Proc. Int. Workshop on Conversational Approaches to Information Retrieval*.
[22] Johanne R. Trippas, Lawrence Cavedon, Damiano Spina, and Mark Sanderson. 2017. How do people interact in conversational speech-only search tasks: A preliminary analysis. In *Proc. ACM SIGIR Conf. on Human Information Interaction and Retrieval*. 325–328.
[23] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the design of spoken conversational search: Perspective paper. In *Proc. ACM SIGIR Conf. on Human Information Interaction and Retrieval*. 32–41.
[24] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, and Mark Sanderson. 2017. A conversational search transcription protocol and analysis. In *Proc. Int. Workshop on Conversational Approaches to Information Retrieval*.
[25] Svitlana Vakulenko, Kate Revoredo, Claudio Di Ciccio, and Maarten de Rijke. 2019. QRFA: A data-driven model of information-seeking dialogues. In *Proc. European Conf. on Information Retrieval*. To appear.
[26] Ryen W White. 2004. *Implicit feedback for interactive information retrieval*. Ph.D. Dissertation. University of Glasgow.
[27] Stephen P. Whiteside and Donald R. Lynam. 2003. Understanding the role of impulsivity and externalizing psychopathology in alcohol abuse: application of the UPPS impulsive behavior scale. 11, 3 (2003), 669–689.