MISC: A data set of information-seeking conversations

Paul Thomas
Microsoft
Canberra, Australia
pathom@microsoft.com

Daniel McDuff
Microsoft
Redmond, WA, USA
damcduff@microsoft.com

Mary Czerwinski
Microsoft
Redmond, WA, USA
marycz@microsoft.com

Nick Craswell
Microsoft
Bellevue, WA, USA
nickcr@microsoft.com

ABSTRACT

Conversational interfaces to information retrieval systems, via software agents such as Siri or Cortana, are of commercial and research interest. To build or evaluate these *software* interfaces it is natural to consider how *people* act in the same role, but there is little public, fine-grained, data on interactions with intermediaries for web tasks.

We introduce the Microsoft Information-Seeking Conversation data (MISC), a set of recordings of information-seeking conversations between human "seekers" and "intermediaries". MISC includes audio and video signals; transcripts of conversation; affectual and physiological signals; recordings of search and other computer use; and post-task surveys on emotion, success, and effort. We hope that these recordings will support conversational retrieval interfaces both in engineering (how can we make "natural" systems?) and evaluation (what does a "good" conversation look like?).

KEYWORDS

Conversational information retrieval, information seeking behaviour

ACM Reference format:
Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2016. MISC: A data set of information-seeking conversations. In *Proceedings of International Workshop on Conversational Approaches to Information Retrieval, Tokyo, August 2017 (CAIR'17)*, 6 pages.

1 INFORMATION-SEEKING CONVERSATION

Voice-enabled agents such as Siri, Cortana, and Alexa are rapidly growing in popularity[1]. Such agents support device control—making telephone calls, starting applications, checking calendars—but also support information-seeking tasks including via web search and specialised services.

These new interfaces pose new constraints. For example, over an audio link it is not possible to offer a screenful of search options or interactive widgets. They offer new opportunities: for example, requests longer than a few terms, differences in tone of voice, and more "natural" back-and-forth. Conversational interfaces also suggest new standards. We each have a lifetime's experience of conversation, and deeply- (if unconsciously-) held ideas of what is "good" and "bad" in a conversation, independent of the information which is transferred [see e.g. 6, 17].

[1]For example over 10M installations of Google Allo (Google, http://bit.ly/1qvt4QP); 133M Cortana users per month (Reported in Tech Radar, http://bit.ly/29TYevk); over 5M Echo sales (Consumer Intelligence Research Partners, http://bit.ly/2j33gYO); 24.5M "voice-first" devices predicted to ship in 2017 (Voicelabs, http://bit.ly/2jJlyk7).

Current systems are not able to maintain lengthy exchanges, and they provide only basic tools for tracking context, non-verbal signals, and emotion. It is natural therefore to look to information-seeking conversations between people: either for insights to help conversational interfaces for information retrieval, or for ideas to evaluate existing or proposed systems [c.f. 4, 8, 25, 30]. If we had a set of information-seeking conversations, recorded and annotated with self-reports and descriptions of the process, we could start to address questions such as:

- How do intermediaries' behaviours relate to seekers' satisfaction? Are there behaviours that a software agent should copy, or should avoid?
- Are there signals in seekers' behaviour which correlate with (e.g.) task success, engagement, satisfaction, or emotion? Can we use these as online measures [16]?
- What tactics are used in conversational information retrieval [30]? Are they similar to those observed and postulated e.g. by Brooks and Belkin [4], Daniels et al. [8], Hennoste et al. [12], or Reichman [25]?
- Do particular conversational structures promote or impede progress in the task, or promote or impede engagement and satisfaction? For example, are the RUSA or IFLA guidelines, which were developed for reference libraries [27], useful in these broader scenarios?
- How important in this context, and how often observed, are conversational norms such as those of Grice [11] or politeness conventions [6, 17]? Does their observance, or otherwise, lead to a better or worse experience?

Public data sets play an important role in advancing research. They allow for benchmarking of new methods, transparency in analysis, and reduce the burden on researchers to collect their own data. Some noteworthy data sets capture natural conversation [e.g. 10], but with no notion of task. Other data sets capture conversation during collaborations on natural tasks [13, 29], or collaboration on assigned tasks [1, 2]: however, there is an asymmetry of role, information, and tools when talking with an agent and this asymmetry is not present in these tasks.

Asymmetries, and conversation between information seekers, are designed in systems by Shah, Pickens, and collaborators [24, 28]. The assignment of roles differs however, in that collaborators are assumed to share an information need and to share a set of resources (e.g. documents and search engine). This is not true when working

with conversational search agents. To the best of our knowledge, nor are transcripts or other corpora available.

The Microsoft Information-Seeking Conversation (MISC) data set describes information-seeking conversations with a human intermediary[2], in a setup designed to mimic software agents such as Siri or Cortana. The release includes audio, video, transcripts, prosodic signals, detected and reported emotion, and data on demographics, stress, satisfaction, success, and engagement.

2 RELATED WORK

A number of conversation transcripts are available, covering different modes (telephone, text, or face-to-face) and different types of task, and which are relevant for digital assistants and other software agents.

For example, SRI have made available transcripts of telephone calls between their employees and travel agents, recorded in the late 1980s and manually transcribed with deletions to protect privacy and confidentiality [29]. These conversations are a combination of information-seeking and transactional needs—both "how can I get to Chicago?" and "book me a flight"—and unlike in MISC or other data, one party is both a domain and a task expert. Although not purely information-seeking, these transcripts may therefore be a good analogue for conversations with a software agent.

A similar data set was used in the Dialog State Tracking Challenge [15]. This includes transcripts of 42 hours of conversation between tourists and travel agents, with the agents recommending accommodation and tourist attractions in Singapore. The transcripts are segmented and annotated for speech acts, and some semantics, on a turn-by-turn basis.

The Ubuntu Dialogue Corpus [18] records IRC exchanges in a linux help forum. This is much larger than the sets above (1M dialogues), again with a division of roles (expert and questioner), but in some regards shallower: interactions are based on text only, and there are no annotations.

Several other corpora are widely used but have more focus on general conversation. Godfrey et al.'s Switchboard data, for example, records general conversation between paired volunteers [10]. Slightly more structured, although with more participants, are the Meeting Recorder transcripts [13, 14], which describe professional discussions between researchers.

Further corpora focus on a single task, although not information seeking. The HCRC Map Task set [2] includes pairs of participants, one giving instructions to the other in each case. Although these dialogues naturally include clarifying questions, the focus is on instruction and task completion. The Verbmobil data [1] likewise focusses on a different task, that of coordinating meetings, and there is less asymmetry of role.

To the best of our knowledge, none of the above come with data derived from video; or substantial data on emotion, affect, satisfaction, stress, or engagement.

The work most closely related to ours is from Trippas et al. [30]. As in MISC, participants were paired into "users" (seekers) and "retrievers" (intermediaries), with tasks assigned only to users. Early analysis has found a narrow variety of moves used by users to communicate their needs, and a larger variety of moves used by retrievers to communicate progress or information found. Trippas et al.'s retrievers were however given slightly different instructions to ours—whereas ours were asked just to "search for the answer" and "give them the information", Trippas et al.'s were explicitly asked to transcribe the queries they heard. Trippas et al.'s protocol perhaps more closely mimics current software, but the retriever has a smaller role in search strategy and in decisions on relevance. The two studies are similar enough that we expect either one could be used to help validate findings from the other.

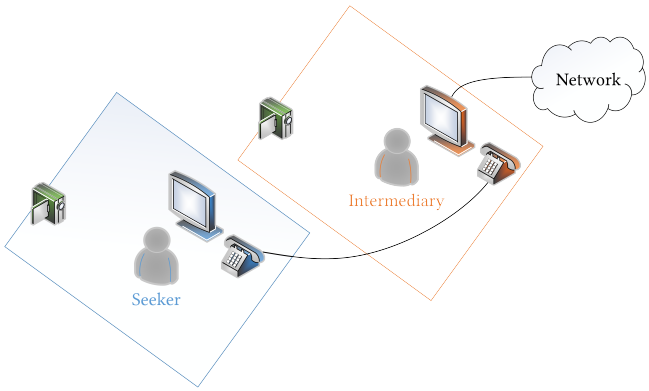

Figure 1: Recording setup. Tasks were assigned to a "seeker", who communicated with an intermediary only over an audio link. The intermediary had access to the internet through a standard browser. Both participants were recorded.

3 METHOD

To help address gaps in previous publicly released datasets we designed a method using pairs of volunteers (Figure 1). Each pair included a "seeker", who was given information needs but no direct way to answer them, and an "intermediary", who had access to a networked computer and standard software but who we did not tell the task. This roughly models situations such as reference interviews; conversations with subject-matter experts such as pharmacists or mechanics; and interactions with software agents such as future versions of Siri or Cortana.

Recruitment and screening. We recruited N=44 volunteers aged 24 to 64 years (median=45, st. dev.=12), 24 female, through Microsoft's in-house support services[3]. Participants were all fluent speakers of North American English (to minimise errors in speech-to-text processing), with self-reported familiarity with and expertise in internet use. Participants were recompensed with a $150 gift card for completing the two hour study.

Protocol. Participants formed 22 pairs and a random member of each pair was assigned the "seeker" role. Seekers were given

[2]Ross et al. [26] call this sort of purposeful interaction an "interview". "Conversation" is the de facto standard term in the technology industry, however, so we use it here.

[3]With $N = 44$ individuals and 22 pairs we have adequate power to recognise moderate differences between users and intermediaries (for example effects of $d \geq 0.63$ for paired t tests when $\alpha = 0.05$, $\beta = 0.8$) or between tasks (for example $f \geq 0.34$ for ANOVA with five groups), or simple correlations between variables (for example $r \geq 0.41$ for 44 individuals).

a sequence of information-seeking tasks and asked to record an answer for each.

The seekers did not have access to any on- or offline resources to complete the task: while they had a computer, this was used only to record their final answer and for post-task questions. Instead they asked questions of, and gave instructions to, the other participant ("intermediary"). The intermediary had the use of a computer with web browser and Internet access (the participants were asked not to use other software, only the web browser). All communication between seeker and intermediary was over an audio link.

After ten minutes on a task, seekers were asked to move on. Out of the 88 tasks completed, the pairs reached the ten-minute limit on 46 occasions (42%). Our participants spent on average 8 minutes 20 seconds on each task (st. dev. 2 minutes 38 seconds).

Tasks. The first task was a simple warm-up, and the remaining four tasks cover a range of difficulty and complexity. The four tasks were assigned with a Latin square to balance order effects. Participants were not informed which task was a warm-up, or informed about expected difficulty or complexity.

Our tasks were adapted from those in the literature, and selected in part to elicit positive (tasks 1, 4) and negative (task 2) emotion:

(0) *(warm-up)* Mary has been hearing a lot about the HPV vaccine, a vaccine that protects against several types of the human papillomavirus, a common sexually transmitted infection (STI). Mary is considering getting the vaccine. Find out who can get the HPV vaccine. (From Buhi et al. [7].)

(1) *(low difficulty, low complexity)* Recently you had dinner with your cousin. She is very cynical and kept telling you that nobody ever helps others unless there's something in it for them. You'd love to prove her wrong, so you want to find accounts of selfless heroic acts by individuals or small groups for the benefit of others or for a cause. (Modified from TREC topic 442.)

(2) *(low difficulty, high complexity)* Imagine that you recently began suffering from migraines. You heard about two possible treatments for migraine headaches, beta-blockers and/or calcium channel blockers, and you decided to do some research about them. At the same time, you want to explore whether there are other options for treating migraines without taking medicines, such as diet and exercise. (From Broussard and Zhang [5].)

(3) *(high difficulty, low complexity)* For a work project you're looking at international sport in the developing world. You're making a list of Olympic venues to see how well different areas are represented. Find the venues of the 2024 Olympic Games and the 2026 Winter Olympic Games. (New for this study. At the time of writing, these venues were not yet announced, so although not cognitively complex this task was impossible as given.)

(4) *(high difficulty, high complexity)* This summer, during your vacation, you are planning to go on a touring trip of North America. You want information to help you plan your journey and there are many tourist attractions you would be interested in visiting. You have set aside 3 months for the trip and hope to see as much of the continent as you can. As you cannot drive, you will have to use public transport, but are unsure which type to take. Bearing in mind this context, your task is to decide on the best form of transportation between cities in North America that would be suitable for you. (From White [32].)

Figure 2: Example video stills from the MISC data. As distributed, there are separate video files for each participant and task.

4 DATA RECORDED AND DERIVED

MISC includes data which was recorded during the tasks; pre- and post-task, self-reported data; and data which we have derived from the raw signal. This data has been processed to include common timestamps, where possible.

4.1 Raw Data

Pre-test: Before, completing the search tasks the participants completed the Positive and Negative Affect Schedule (PANAS) [31] and "Big Five" personality traits questionnaire. These act as a reference against which to interpret affect, emotion, and physiological data. We also collected information on gender, age range, ethnicity, and education as demographics influence expressivity [19].

During the task: During each task, we recorded the intermediary's search use: queries, pagination, and a screen recording of both participants computer screens. This will enable researchers to identify the specific part of the page being viewed. We also recorded audio and video of each participant, separately using cameras and microphones in each room (Figure 2). As the participants were wearing headphones the audio channels can be separated. The cameras were positioned to capture the full face of the individual; however, the participants were free to move and part of the face may be excluded in some frames.

Post-task: After each task, we asked both participants questions on emotion, effort, and engagement. The emotion questions were based on the most commonly used categorization of emotions (the so-called "basic emotions"—anger, contempt, disgust, fear, joy, sadness, and surprise). We also added interest, frustration and boredom to this list as they were deemed very relevant to information searching tasks:

(1) What was the dominant emotion you experienced during the task you just completed? *(Participants had to choose one.)*

(2) How intense was the dominant emotion you experienced? *(Likert-like, seven points from "very mild" to "very strong".)*

(3) Which of the following emotions did you experience during the task you just completed? *(Participants could choose one or more.)*

Participants were also asked about their effort. This used the NASA Task Load Index (TLX) [20], excluding physical demand, and each of the five items was on a seven-point scale:

(4) How mentally demanding was the task you just completed?
(5) How successful were you at accomplishing what you were asked to do?
(6) How hurried or rushed was the pace of the task?
(7) How hard did you have to work to accomplish your level of performance?
(8) How insecure, discouraged, irritated, stressed, and annoyed were you?

Item 5 was reverse coded.

A further set of items were adapted from the User Engagement Scale (UES) [21]. The UES includes several sub-scales; we used items from the combined "novelty", "felt involvement", and "endurability" sub-scale identified by O'Brien and Toms [22] for exploratory search. This sub-scale measures users' feelings of success, reward, and willingness to engage again, which we believe are relevant for any new technology. Other items from the UES were not relevant to our setup (the "aesthetic", "focussed attention", and "usability" subscales) or were dropped to save space as they partially overlapped with questions asked elsewhere.

Each item was on a seven-point Likert scale, and wording was somewhat adapted from the original:

(9) This experience was worthwhile.
(10) I consider my experience a success.
(11) This experience did not work out the way I had planned.
(12) My experience was rewarding.
(13) I felt interested in my task.
(14) This experience was fun.

Item 11 was reverse-coded.

It is of course possible that the wording changes above have introduced their own effects [22]. We leave this for later investigation.

Three final items asked after the participants' opinion of their partner:

(15) The other participant helped me work on this task.
(16) The other participant understood what I needed.
(17) The other participant communicated clearly.

Exit survey: After all tasks were completed, we asked participants two open-ended questions.

(18) What did you *like* about using another human to search?
(19) What did you *not like* about using another human to search?

Finally, the participants were asked to (20) rank the tasks in order of difficulty.

4.2 Automatic coding

The MISC set includes features automatically derived from the audio-visual recordings. Timestamped transcripts were produced using the Microsoft speech-to-text toolkit[4], and the transcripts are included as plain text. Linguistic Inquiry and Word Count (LIWC) outputs [23] are included for each participant and task, based on the transcripts. Basic prosodic signals—F0, voicing, and loudness—were extracted from the raw audio using OpenSMILE [9] and are also included, timestamped, for each participant.

We used the OpenFace toolkit [3] coding of head pose, 18 facial action units and seven emotion expressions. The released data includes time series of recognition confidence for each facial action unit and basic expression.

[4]https://www.microsoft.com/cognitive-services/en-us/speech-api

5 SUMMARY STATISTICS

A basic analysis of the MISC data strongly suggests that it is useful for understanding "natural" conversations, and as a guide to automatic evaluation. Further analysis is ongoing, and the data is available for other researchers.

5.1 Conversations

Participants exchanged 857 words per task, on average, although this varies with task and participant pair (s.d. 352 words per pair per task, range 210–1881), and on average intermediaries spoke slightly more than seekers (456 vs 397 words, one-sided t test $p < .01$).

5.2 Emotion

Participants' responses indicate a variety of emotions during the tasks. The most common were interest (185 reports from 217 responses), frustration (67 reports), surprise (56), joy (33), and boredom (27); sadness, contempt, disgust and fear were seldom reported and anger not at all. Emotions were typically mixed, with more than one emotion in 53% of cases. Participants' dominant emotion was interest in 156 cases (from 217 responses), frustration in 31 cases, and others in 30.

5.3 Effort

Cronbach's α over the five TLX items was good at 0.84 ± 0.03, which is consistent with experience elsewhere, so we computed a composite "effort" score as the mean of all five items. These effort scores are reasonably well distributed with mean 2.95/7, and standard deviation 1.3 (see right-hand margin of Figure 3). They do however vary considerably between participants, with one participant reporting mean effort only 1.36 on the 1–7 scale and another reporting mean effort 5.36 across the same tasks.

The seekers' effort correlates slightly with the length of conversations, as measured by word count: they reported more effort for tasks where they talked more (Pearson's $r = 0.30 \pm 0.18$, $p < 0.1$) and where the intermediary talked more ($r = 0.43 \pm 0.17$, $p \ll 0.01$). Further analysis is needed, but we may imagine simply counting words as a proxy for effort with a software agent.

5.4 Engagement

As we took a subset of the larger user engagement scale, we have also considered the internal consistency of these six items. Cronbach's α on this group was 0.85 ± 0.03, representing good reliability without redundancy, and no single item was uncorrelated with the others. Again, this allows us to compute a composite score. The engagement score is reasonably well distributed (mean 4.96, standard deviation 1.25; see top margin of Figure 3). Again there is a large range across participants (lowest average 3.10, highest 6.90).

We might expect a relationship between word count and engagement, and this is borne out in the data although the effect is less

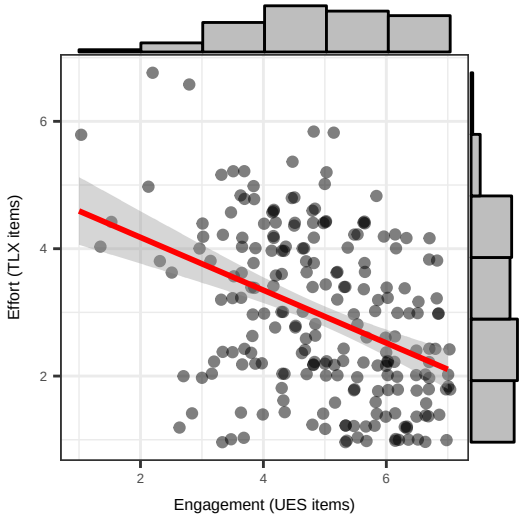

Figure 3: Distribution of self-reported effort (mean of five TLX items) and engagement (mean of six UES items). Pearson's $r = -0.40$.

than on effort ($r = -0.28 \pm 0.16$ for seekers' words, $r = -0.16 \pm 0.18$ for intermediaries').

We may also expect some correlation between effort and engagement, and again there is a relationship, as illustrated in Figure 3. Easier tasks correspond with greater engagement, but as before the correlation is only moderate (Pearson's $r = -0.40 \pm 0.11$). This suggests other factors are partly responsible for feelings of engagement. Understanding what else promotes engagement in conversational retrieval, and whether it can be measured or predicted from observed behaviour, is a key use for MISC.

5.5 Exit questions

Participants consistently rated the transportation task hard (22 participants rated it the hardest of the five, out of 42 participants responding) and the HPV task easy (22/42 rated it easiest). Relative orderings of the other three tasks varied.

In post-experiment comments, seekers often reported appreciating their partners' differing views and ideas: "it was interesting to have my partner's input" (participant 25); "it was helpful to bounce ideas off of someone and brainstorm together. She had ideas that I had not considered" (participant 29); "having her perspective was very helpful." (participant 37). Seekers also appreciated discussing search strategies, and simply working with another person.

Seekers also reported that they did not like giving up direct control of the search process: "it was harder to get ideas across" (participant 32); "I didn't have control over the proceedings" (participant 27); "not being able to see or focus on the search results of my choice" (participant 33). There was also some annoyance caused by difference in styles: "our thought processes do not necessarily align" (participant 39); "I like to visually scan for search results that stand out to me. She may not have chosen what I would have" (participant 37); "there were some searches I would have conducted very differently from her and it was frustrating to have to let her do it her way" (participant 34).

We anticipate that these comments will motivate futher investigations, with MISC or with follow-up experiments.

6 AVAILABILITY

The MISC data is available at http://aka.ms/MISCv1. It includes our audio, video, and questionnaire data, as well as the derived data described above.

Participants provided informed consent for use of their data for research purposes. Distribution of the dataset is governed by the terms of their consent. Approval to use the data does not allow redistribution, and is covered by citation terms.

7 CONCLUDING REMARKS

The Microsoft Information-Seeking Conversation data records pairs of volunteers working together to solve information-seeking tasks. One participant acts as "seeker" and one as "intermediary": this models interactions with, e.g., digital assistants such as Siri or Cortana instead of a conventional collaborative task. The set includes audio/video, as well as transcripts and other derived data; affect and physiological signals derived from video; and responses to post-task questions on effort, engagement, and satisfaction. Measures of effort and engagement are internally consistent and have a usable range of responses, and although there is some correlation between the two engagement is only partly explained by effort. As well as behaviours and physiological data, participants' comments provide some insight in this regard.

We hope that the MISC data can be used to support a range of investigations, including for example the understanding the relationship between intermediaries' behaviours and seekers' satisfaction; mining seekers' behavioural signals for correlations with success, engagement, or satisfaction; examining the tactics used in conversational information retrieval and how they differ from tactics in other circumstances; the importance of conversational norms or politeness; or investigating the relationship between conversational structure and task progress. You are invited to download and use the data, and to contact the authors with any comments.

ACKNOWLEDGMENTS

We thank our participants for their time. We also thank Microsoft's User Experience Central group for their support, and Heather O'Brien for her advice on the User Engagement Scale.

REFERENCES

[1] Jan Alexandersson, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Elisabeth Maier, Norbert Reithinger, Birte Schmit, and Melanie Siegel. 1997. *Dialogue acts in VERBMOBIL-2.* Verbmobil report 204.

[2] Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. 1991. The HCRC map task corpus. *Language and Speech* 34, 4 (1991), 351–366.

[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*. IEEE, 1–10.

[4] H M Brooks and N J Belkin. 1983. Using discourse analysis for the design of information retrieval interaction mechanisms. In *Proc. SIGIR*. 31–47.

[5] Ramona Broussard and Yan Zhang. 2013. Seeking treatment options: Consumers' search behaviors and cognitive activities. *Proc. American Society for Information Science and Technology* 50, 1 (2013), 1–10.

[6] Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language use.* Cambridge University Press, Cambridge.

[7] E R Buhi, E M Daley, H J Fuhrmann, and S 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] P 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.

[9] 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.* ACM Press, 835–838.

[10] John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proc. IEEE Conf. Acoustics, Speech, and Signal Processing*, Vol. 1. 517–520.

[11] H Paul Grice. 1975. Logic and conversation. In *Syntax and Semantics*, Peter Cole and Jerry L Morgan (Eds.). Vol. 3. Academic Press, New York, 41–58.

[12] T Hennoste, O Gerassimenko, R Kasterpalu, M Koit, A Rääbis, K Strandson, and M Valdisoo. 2005. Information-sharing and correction in Estonian information dialogues: Corpus analysis. In *Proc. Second Baltic Conf. on Human Language Technologies.* 249–254.

[13] International Computer Science Institute. 2004. The ICSI meeting corpus. (2004). Retrieved June 2016 from http://www1.icsi.berkeley.edu/Speech/mr/

[14] International Computer Science Institute. 2004. Meeting Recorder Dialog Act (MRDA) database. (2004). Retrieved June 2016 from http://www1.icsi.berkeley.edu/~ees/dadb/

[15] Seokhwan Kim, Luis Fernando DfiHaro, Rafael E Banchs, Jason Williams, Matthew Henderson, and Koichiro Yoshino. 2016. Dialog state tracking challenge 5 handbook. (2016). Retrieved February 2017 from https://github.com/seokhwankim/dstc5/raw/master/docs/handbook_DSTC5.pdf

[16] J Kiseleva, K Williams, J Jiang, A H Awadallah, A C Crook, I Zitouni, and T Anastasakos. 2016. Understanding user satisfaction with intelligent assistants. In *Proc. Conference on Human Information Interaction and Retrieval.* 121–130.

[17] Geoffrey N. Leech. 1983. *Principles of pragmatics.* Longman, London.

[18] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proc. SIGDIAL.* 285–294.

[19] Daniel McDuff, Jeffrey M. Girard, and Rana el Kaliouby. 2017. Large-scale observational evidence of cross-cultural differences in facial behavior. *Journal of Nonverbal Behavior* 41, 1 (2017), 1–19.

[20] 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/

[21] 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.

[22] Heather L O'Brien and Elaine G Toms. 2013. Examining the generalizability of the User Engagement Scale (UES) in exploratory search. *Info. Proc. Mgmt.* 49 (2013), 1092–1007.

[23] J W Pennbaker, R L Boyd, K Jordan, and K Blackburn. 2015. *The development and psychometric properties of LIWC2015.* Technical Report. University of Texas at Austin.

[24] Jeremy Pickens, Gene Golovchinsky, Chirag Shah, Pernilla Qvarfordt, and Maribeth Back. 2008. Algorithmic mediation for collaborative exploratory search. In *Proc. SIGIR.* 315–322.

[25] Rachel Reichman. 1985. *Getting computers to talk like you and me.* MIT Press, Cambridge, Massachusetts.

[26] Catherine Sheldrick Ross, Kirsti Nilsen, and Patricia Dewdney. 2002. *Conducting the reference interview: A how-to-do-it manual for librarians.* Facet Publishing, London.

[27] Pnina Shachaf and Sarah M Horowitz. 2008. Virtual reference service evaluation: Adherence to RUSA behavioral guidelines and IFLA digital reference guidelines. *Library and Information Science Research* 30 (2008), 122–137.

[28] Chirag Shah, Jeremy Pickens, and Gene Golovchinsky. 2010. Role-based results redistribution for collaborative information retrieval. 46 (2010), 773–781.

[29] SRI International. 2011. SRI's Amex Travel Agent Data. (2011). Retrieved June 2016 from http://www.ai.sri.com/~communic/amex/amex.html

[30] 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. Human Information Interaction and Retrieval.*

[31] D Watson, L A Clark, and A Tellegan. 1988. Development and validation of brief measures of positive and negative affect: The PANAS scales. *J. Personality and Social Psychology* 54, 6 (1988), 1063–1070.

[32] Ryen W White. 2004. *Implicit feedback for interactive information retrieval.* Ph.D. Dissertation. University of Glasgow.