Factors Affecting Users' Information Requests

Jaime Arguello, Bogeum Choi, Robert Capra
School of Information and Library Science
University of North Carolina at Chapel Hill
jarguello,choiboge,rcapra@unc.edu

ABSTRACT

Conversational search interfaces have two important characteristics: (1) they can accept voice requests from users and (2) they aim to provide users with more human-like interactions. In this paper, we investigate how two factors influence users' information requests. Our first factor, *medium*, considers whether the request is produced using text or voice. Our second factor, *target*, considers whether the request is intended for a search engine or a human search intermediary. In particular, we study how these two factors influence users' requests during search tasks that have a *domain knowledge* constraint—the user wants information for a domain novice or expert. We analyze information requests collected using crowdsourcing and address three research questions. We study the effects of our two factors (medium and target) on: (RQ1) participants' perceptions about their information requests, (RQ2) the different characteristics of their information requests, and (RQ3) participants' strategies when requesting information appropriate for a novice or expert. Our results show that *both* factors had a strong effect on participants' requests (including retrieval performance), and that *target* had a stronger effect on participants' perceptions and their choice of strategy in requesting novice- or expert-appropriate information.

1 INTRODUCTION

Conversational search interfaces are increasingly common and include intelligent mobile assistants such as Cortana, Google Now, and Siri, as well as intelligent home assistants such as Amazon Alexa and Google Home. Conversational search interfaces have two important characteristics. First, they are able to accept spoken (rather than textual) information requests from users. Second, they aim to engage with users using more human-like interactions. As conversational search interfaces evolve, it is important to understand how their differences from traditional text-based search systems may influence the way users request information.

A large body of prior work has already found important differences between voice and textual queries [3–6, 14, 19]. However, with respect to conversational search interfaces, open questions remain. For example, what happens when users increase their expectations about the system's ability to understand and respond to requests? Will users change the way they formulate their queries? Will users contribute more information about their information need? Will this additional information actually improve retrieval results? And, if not, why not?

In this paper, we investigate how two different factors influence users' information requests. Our first factor, referred to as the *medium* of the information request, considers whether the request is produced using text or voice. Our second factor, referred to as the *target* of the information request, considers whether the request is intended for a search engine or a human search intermediary (i.e., someone who will search on the user's behalf). Our goal in the "human intermediary" condition was to study information requests in the extreme case where users have a very high expectation about the system's ability to understand and respond to requests.

In addition to investigating our two factors of medium and target, we are interested in cases where the information need has an extra-topical dimension. It has long been recognized that relevance from a user's perspective is a multi-dimensional concept [12]. Prior studies clearly show that users consider different criteria (other than topic) when deciding that a document is relevant to an information need. Important criteria include the user's domain knowledge and the readability or understandability of the information [2, 10, 11, 13, 16–18]. In this paper, we focus on information needs that have a domain knowledge constraint—the user wants information that is appropriate for a domain novice or expert. While there are other extra-topical dimensions to consider (e.g., temporal, geographical), we focus on domain knowledge as a first step.

In this paper, we analyze a large number of information requests gathered using crowdsourcing and address the following three research questions:

RQ1: In our first research question, we investigate how our two factors of medium and target influence users' perceptions about their own information request. Specifically, we focus of users' perceived level of difficulty in producing the information request, their level of satisfaction with the request, and their level of confidence that the request will yield high-quality search results.

RQ2: In our second research question, we investigate how our two factors of medium and target influence the characteristics of users' information requests. We address this question from two perspectives. First, following prior research on the differences between textual and spoken queries [6], we consider different surface-level characteristics such as length (in words), part-of-speech composition, and level of grammatical complexity. Second, we compare the retrieval effectiveness between information requests produced under different conditions.

RQ3: In our third and final research question, we investigate how our two factors of medium and target influence the way people request information appropriate for a domain novice or expert (a type of extra-topical constraint). For example, do people use more elaborate or complex strategies when the request is produced using speech versus text? Or when the request is intended for a human intermediary versus a search engine? Again, we address this question from two perspectives. First, we investigate whether certain strategies are more common depending on the medium and/or target condition. Second, we investigate each strategy's retrieval performance. Ultimately, we are interested in whether our two factors of medium and target influence people to adopt strategies that impact retrieval performance given an existing state-of-the-art retrieval system.

2 RELATED WORK

Our research builds on two areas of prior work: analyzing the differences between spoken and textual queries and research on the different criteria that influence relevance from a user's perspective.

Spoken vs. Textual Queries: Early work by Du and Crestani [4, 5] explored the differences between spoken versus textual queries for the *same* information needs (based on TREC topics). Spoken queries were found to be longer than textual queries in English [4] and Mandarin [5]. In English, spoken queries also had more varied language—the total number of unique terms from all participants for the same information need was greater. In a follow-up analysis of the same English data, Crestani and Du [3] also found that spoken queries had fewer nouns and more of other parts of speech, and that spoken queries outperformed textual queries *only* after removing all words except for nouns, adjectives, and verbs.

Prior work has also investigated the differences between spoken and textual queries in the mobile search domain. Schalkwyk *et al.* [14] found that spoken mobile queries tend to be about "on-the-go" topics and are less likely to be about sensitive topics such as adult themes and health, and tend to have less interaction with the search results. Schalkwyk *et al.* [14] also found spoken queries to be *shorter* than textual queries, although the authors note that the differences in topics between spoken and textual mobile queries might have been a confounding factor. Yi and Maghoul [19] analyzed 79K spoken mobile queries from 2010, and found spoken queries to be slightly longer than textual queries and, consistent with the previous paper [14], more often associated with topics that require less interaction (i.e., good abandonment). More recently, Guy [6] compared 500K spoken versus 500K textual queries from 2015. Spoken queries were found to be longer than textual queries and have a greater resemblance to natural language. For example, spoken queries had more wh-words, more parts of speech other than nouns, and more full-sentence grammatical structure. Furthermore, consistent with previous work, spoken queries were more often associated with "quick search" topics that a require less interaction (e.g., more often triggered entity cards and returned multimedia content), and were less often associated with sensitive subjects such as adult themes. Finally, spoken queries more often had terms that are easy to pronounce, but difficult to spell, while textual queries more often had terms that are easy to type, but difficult to speak (e.g., abbreviations and calendar years).

Prior research has also investigated how users reformulate spoken queries in response to different types of system errors (i.e., recognition error, term addition/deletion, or system interruption). Jiang *et al.* [8] found that users respond to system errors using a combination of lexical and phonetic reformulation pattern. Results also found that certain reformulation patterns are more effective than others in yielding a better retrieval. Shokouhi *et al.* [15] investigated reformulation patterns across modalities (text and speech) and found that users rarely switch modalities during a reformulation. Hassan *et al.* [7] focused on the task of predicting whether a pair of consecutive voice queries have the same information need and, if so, whether the reformulation was due to a recognition error or non-relevant search results. The authors used a combination of session, query-similarity, and speech features (derived from the ASR system output), and found these to be complementary.

Relevance Criteria: A large body of research shows that relevance is a multi-dimensional concept, often involving different extra-topical preferences or constraints. Saracevic's stratified model argues that relevance from a user's perspective is influenced by different cognitive and affective factors, such as the user's state of knowledge, goals, and motivations, as well as different situational factors [12]. Prior research, mostly from the information and library science communities, has studied the different criteria users consider when making relevance decisions. Several studies have cited as important factors the content's readability and understandability, as well as the user's prior knowledge in the domain [2, 10, 11, 13, 16–18]. In general, these studies have analyzed how different relevance criteria are used to explain or justify document relevance decisions, either retrospectively [10], or in real-time using think-aloud protocols [17]. In this study, we investigate how users request information in the presence of a domain knowledge preference (a type of extra-topical dimension), and how our two factors of medium and target influence the strategies adopted.

3 METHODOLOGY

To explore our three research questions, we gathered a large set of information requests using Amazon Mechanical Turk (MTurk). As described in more detail later, we constructed 10 search tasks with a domain knowledge dimension present in the background story. Five of tasks were designed to influence participants to request information for a domain *novice* and five were designed to influence participants to request information for an *expert.* As previously mentioned, we investigate the influence of two factors on participants' information requests: medium (text vs. speech) and target (search engine vs. human). MTurk workers were randomly assigned to *one* combination of medium and target condition (i.e. a between-subjects design).

Our MTurk Human Intelligent Tasks (HITs) proceeded as follows. First, participants were given a set of instructions describing the HIT. The instructions were different depending on the medium and target condition. Then, participants were directed to a webpage where they could access the search task and enter their information request. We wanted to discourage participants from using parts of the search task description verbatim in their information requests. To this end, participants were required to click a "view task" button that displayed the search task description in a pop-up browser window. Participants had to close the pop-up window to continue interacting with the page. In other words, participants were not able to see the search task description and produce the request at the same time. Finally, after submitting their information request, participants were directed to a post-task questionnaire.

We gathered 14 information requests from different MTurk workers for each combination of search task, medium condition, and target condition ($10 \times 2 \times 2 \times 14 = 560$ total requests). Individual participants were not able to "see" the same search task more than once (whether or not they completed the HIT). We restricted our HITs to U.S. MTurk workers with a 95% acceptance rate or greater, and gathered information requests from 182 individual workers.

Search Tasks: We designed 10 search tasks aimed to influence participants to request information appropriate for either a domain novice or expert (5 tasks each). Each search task included a background story that added the domain knowledge dimension to the task. One of our goals was to avoid influencing participants to use specific terms from the task description in their own information requests (i.e., language priming). To this end, each of our tasks described a scenario in which the participant had already done some searching, and now needed to find more appropriate information.

Novice Example: For her Health class at school, your 13-year old nephew is writing a school paper on the metabolic processes of living organisms: the process through which living organisms produce energy. You want to help your nephew find information. So far, you have ONLY found highly technical scientific papers intended for medical professionals and scientists. You want to find information that is more suitable for your nephew.

Expert Example: Your 24-year old cousin is studying to become an occupational therapist. For one of her college courses, she is writing a research paper on the different health effects of UV rays on the eyes. You want to help her find information for his paper. So far, you have ONLY found information for people who want quick tips on how to protect their eyes and avoid exposure. You want to find information that is more suitable for your cousin.

Medium and Target Conditions: We investigate the influence of two factors on participants' information requests. The *medium* considers whether the information request is produced using text or speech, and the *target*, considers whether the information request is intended for a search engine or a human intermediary.

Medium: In the *text* condition, participants entered their information requests using a textbox of the same size as most commercial search engines. In the *voice* condition, we provided participants with Javascript tools to record their information request using their own microphone, save the recording as a WAV file, and upload the WAV file to our server.

Target: In the *search engine* condition, participants were asked: "How would you request this information from a search engine such as Google, Bing, or Yahoo?" In the *human intermediary* condition, participants were told that the goal of the HIT was to pilot-test a service called SearchForMe, which allows users to submit information requests to search intermediaries who can search for information on their behalf. Participants were asked: "How would you request this information from a SearchForMe intermediary?"

Post-task Questionnaire: After submitting their information request, participants were directed to a post-task questionnaire. Participants were asked about: (1) the level of difficulty in producing the information request, (2) the level of satisfaction with the information request, (3) the level of confidence that a search engine (or human intermediary, depending on the target condition) would be able to "understand" the information request, and (4) the level of confidence that a search engine (or human intermediary) would be able to find useful information in response to the information request. All four questions were asked using agreement statements with a 7-point scale with labeled endpoints (from strongly disagree (1) to strongly agree (2)).

Voice Requests and Relevance Judgments: We used MTurk for two additional data collection efforts. First, all information requests submitted in the voice condition (submitted as WAV files) were transcribed using MTurk. Initially, we tried using several speech-to-text APIs, but the ASR error rate to be too high. MTurk workers were paid $0.20 USD to listen to a WAV file and produce a textual transcription. Second, as part of RQ2 and RQ3, we were interested in measuring the retrieval effectiveness of information requests. To this end, we issued all requests (in textual form) to a commercial web search API and cached the top-5 search results. Then, we used MTurk to gather relevance judgments for these results. MTurk workers were given the original search task description (not the information request) and were asked to judge a set of webpages as as being *useful* or *not useful* for the given search task. MTurk workers were paid $0.20 USD to judge a bundled set of 10 web results for the same task. Each web result was judged by five redundant workers. The Fleiss Kappa agreement between workers was $\kappa_f = 0.366$, which is considered "fair agreement" [9].

Strategy Annotation: In our third research question (RQ3), we investigate the different strategies used by participants to request information appropriate for a domain novice or expert. To address this question, we manually coded participants' information requests into different strategies. Two of the authors conducted two rounds of qualitative coding. During the first round, both authors independently coded 200 information requests (100 from novice search tasks and 100 from expert search tasks), and then resolved their codes to form a closed set of codes. During the second round, both authors independently (re-)coded the 200 requests and all remaining requests using the closed set of codes. Six different codes were identified and are described below. For each code, we include one example information requests from a novice search task and one example information request from an expert search task.

(1) **Genre:** The participant mentioned the desired genre.
- Novice: Robotics research mainstream articles
- Expert: Find scientific studies on [...]

(2) **Complexity:** The participant used adjectives to describe the complexity of the information desired.
- Novice: The human metabolic process in simple English.
- Expert: Find detailed information regarding [...]

(3) **Source:** The participant requested information from a particular online publisher or source.
- Novice: I'm looking for YouTube videos that explain [...]
- Expert: U.S. census projections and past data

(4) **Purpose:** The participant described their purpose in requesting the information.
- Novice: [...] for a middle school report
- Expert: data for college paper on [...]

(5) **User:** The participant described attributes of the person who would ultimately use the information.
- Novice: [...] for teens
- Expert: Give me search results for [...] for people in their 20s

(6) **Not want:** The participant explicitly described the types of information he/she did not want.
- Novice: I want to find information [...] for a teenager to understand rather than highly technical articles
- Expert: [...] but I don't want opinionated partisan news articles

Table 1 shows the Cohen's Kappa agreement (κ_c) between the two coders for each strategy. The agreement level was "almost perfect" ($0.80 < \kappa_c \leq 1.00$), "substantial" ($0.60 < \kappa_c \leq 0.80$), and "moderate" ($0.40 < \kappa_c \leq 0.60$) [9].

4 RESULTS

In this section, we present results for our three research questions (RQ1-RQ3). We discuss major trends and implications in Section 5.

Table 1: Cohen's Kappa (κ_c) Agreement

Strategy	κ_c
User	0.878
Purpose	0.839
Not Want	0.768
Complexity	0.684
Source	0.679
Genre	0.558

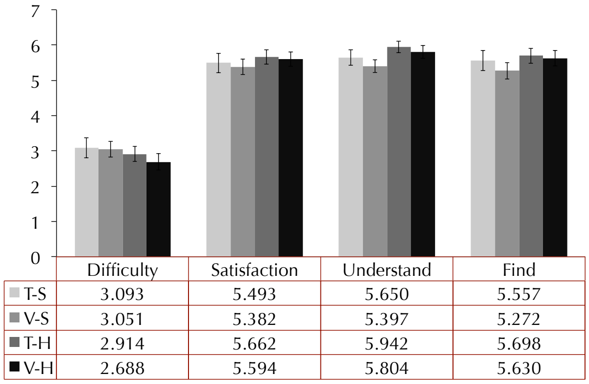
4.1 RQ1: Participants Perceptions

In our first research question (RQ1), we investigate the effects of our two factors (medium and target) on participants' perceptions about their information request. To address this question, we analyzed participants' responses to our four post-task questions about: (1) *difficulty* in producing the information request, (2) *satisfaction* with the request, (3) confidence that a search engine or intermediary (depending on the target condition) would be able to *understand* the request, and (4) confidence that a search engine or intermediary would be able to *find* useful information in response to the request. Figure 1 shows the means of participants' responses across medium and target conditions.

We performed two-way ANOVAs to analyze the effects of medium and target on each measure. Medium had a *marginally* significant main effect for 'understand' ($F(1, 559) = 3.583$, $p = .059$), with participants reporting higher levels of confidence in the text versus voice condition.

Target had a significant main effect for 'difficulty' ($F(1, 559) = 3.943$, $p < .05$), 'understand' ($F(1, 559) = 11.473$, $p < .05$), and 'find' ($F(1, 559) = 5.527$, p =< .05), and had a *marginally* significant effect main for 'satisfaction' ($F(1, 559) = 3.101$, $p = .079$). Participants reported lower levels of difficulty, as well as higher levels of satisfaction and confidence ('understand' and 'find'), in the human versus search engine condition.

The interaction effect between medium and target was not significant for any of the four measures. That being said, participants reported the *lowest* levels of satisfaction and confidence ('understand' and 'find') in the voice and search engine (V-S) condition.



	Difficulty	Satisfaction	Understand	Find
■ T-S	3.093	5.493	5.650	5.557
■ V-S	3.051	5.382	5.397	5.272
■ T-H	2.914	5.662	5.942	5.698
■ V-H	2.688	5.594	5.804	5.630

Figure 1: Mean of post-task questionnaire responses across conditions: text/search engine (T-S), voice/search engine (V-S), text/human (T-H), and voice/human (V-H)

4.2 RQ2: Information Request Characteristics

In our second research question (RQ2), we investigate the effects of our two factors (medium and target) on the characteristics of participants' information requests. We address this question from two perspectives. First, we compare the surface-level characteristics of information requests gathered under different conditions. Second, we consider the retrieval effectiveness of participants' information requests when issued to a commercial web search API.

Surface-level Characteristics: We computed the following measures: number of words, percentage of non-stopwords, and percentage of different parts of speech (nouns, adjectives, verbs, prepositions, determiners, pronouns, adverbs, WH-words). Moreover, similar to Guy [6], we used the Stanford NLP Toolkit to compute different parse tree characteristics: the parse tree depth (a measure of grammatical complexity), and whether the parse tree of the request has an "S-root" (full-sentence structure) or an "NP-root" (noun-phrase structure). Table 2 shows the mean values of these measures across medium and target conditions.

We conducted two-way ANOVAs to analyze the effects of medium and target on each of the first 11 *real*-valued measures in Table 2. We used a logistic regression to analyze the effects of medium and target on the last two *binary*-valued measures. The last column in Table 2 indicates whether there was a significant main effect of medium (m), target (t), and/or a significant interaction effect of medium and target (m*t) on each measure.

The results in Table 2 show three important trends. First, medium had a significant main effect on all but five measures. Compared to textual requests, spoken requests were significantly longer, had a greater percentage of stopwords, had fewer nouns and more of other parts of speech (verbs, prepositions), and had more complex grammatical structure (deeper parse trees, more parse trees with an "S-root", and fewer with an "NP-root"). These results are consistent with previous comparisons between voice and textual queries in the mobile search domain [6], as well as between voice and textual queries gathered in a lab setting for the same tasks [3–5].

Second, target had a significant main effect for all but three measures. Requests intended for a human intermediary were also longer, had a greater percentage of stopwords, had fewer nouns and more of other parts of speech (e.g., wh-words and verbs), and had more complex grammatical structure.

Finally, we found a significant interaction effect for three measures: number of words, percentage of pronouns, and the tree depth (grammatical complexity). Medium and target had a strong *additive* effect—the values for these three measures were much larger in the Voice-Human (V-H) condition.

Retrieval Performance Across Conditions: As previously mentioned, to analyze the retrieval effectiveness of participants' information requests, we used MTurk to gather relevance labels for the top-5 results returned from a commercial web search API. Each result was judged by five redundant workers, and we used a majority vote to derive binary relevance labels. Figure 2 shows the average $\mathcal{P}$@5 values across medium and target conditions.

We used a multiple linear regression to predict $\mathcal{P}$@5 based on the information requests' medium and target condition. For medium, text was coded as 0 and voice was coded as 1. For target, search engine was coded as 0 and human was coded as 1. A significant regression equation was found ($F(2, 558) = 21.450$, $p < .001$), with

Table 2: Mean (SD) of surface-level measures across conditions: text/search engine (T-S), voice/search engine (V-S), text/human (T-H), and voice/human (V-H)

	T-S	V-S	T-H	V-H	Sig.
words	6.74 (2.31)	8.14 (3.15)	9.76 (4.99)	14.35 (10.04)	m,t,m*t
non-stop	0.75 (0.16)	0.69 (0.15)	0.66 (0.17)	0.58 (0.16)	m,t
nouns	0.57 (0.18)	0.52 (0.18)	0.48 (0.16)	0.41 (0.15)	m,t
adj	0.14 (0.15)	0.12 (0.12)	0.13 (0.12)	0.12 (0.11)	–
verbs	0.06 (0.10)	0.07 (0.08)	0.11 (0.10)	0.12 (0.10)	t
prep	0.13 (0.12)	0.16 (0.09)	0.15 (0.10)	0.16 (0.08)	m
det	0.02 (0.05)	0.04 (0.07)	0.04 (0.06)	0.06 (0.06)	m,t
pro	0.01 (0.04)	0.01 (0.03)	0.01 (0.03)	0.02 (0.04)	–
adv	0.01 (0.04)	0.01 (0.03)	0.01 (0.03)	0.01 (0.03)	–
wh	0.01 (0.03)	0.02 (0.04)	0.02 (0.05)	0.02 (0.04)	t
depth	5.86 (1.66)	7.01 (2.30)	7.91 (3.00)	10.41 (5.19)	m,t,m*t
root_s	0.20 (0.40)	0.44 (0.50)	0.49 (0.50)	0.58 (0.50)	m,t
root_np	0.64 (0.48)	0.49 (0.50)	0.38 (0.49)	0.25 (0.44)	m,t

an R^2 of .269. Both factors had a significant effect. In terms of medium, $\mathcal{P}$@5 performance was significantly lower in the voice versus text condition ($\beta = -.106, p < .001$). In terms of target, $\mathcal{P}$@5 performance was significantly lower in the human versus search engine condition ($\beta = -.148, p < .001$).

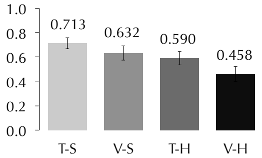


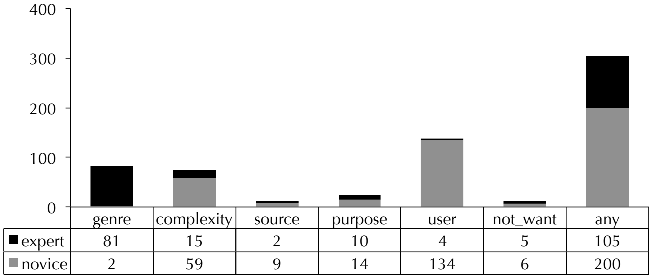
Figure 2: Mean P@5 of information requests across medium and target conditions: text/search engine (T-S), voice/search engine (V-S), text/human (T-H), and voice/human (V-H)

4.3 RQ3: Participants' Strategies

In our third research question (RQ3), we investigate the effects of our two factors (medium and target) on participants' strategies for requesting information appropriate for a domain novice or expert. We address this question from two perspectives. First, we consider the frequencies of different strategies across medium and target conditions. Second, we consider the retrieval effectiveness of different strategies. Our goal in pursuing these two perspectives was to investigate whether our participants used more effective strategies in certain medium and target conditions.

Before presenting our RQ3 results, we report on the frequencies of different strategies across all conditions. Recall that all information requests were coded by two independent coders, and that codes were treated as being mutually exclusive (i.e., an information request could be associated with zero, one, or more than one code). We considered a strategy to be present in an information request only if *both* coders agreed on its presence—the most conservative strategy. Figure 3 shows the number of information requests ($n = 560$) associated with each strategy. We show results for information requests associated with novice search tasks ($n = 280$) and expert search tasks ($n = 280$). The last column labeled 'any' shows the number of information requests that had *at least one* strategy.

Figure 3 shows four important trends. First, based on the 'any' column, in most cases (305/560) participants explicitly stated a desire for information appropriate for a novice or expert using at least one of our strategies. Second, this was mostly true for novice search tasks (200/280) versus expert search tasks (105/280). One possible explanation is that the topics surrounding our novice tasks were fairly technical, influencing our participants to *explicitly* request novice-appropriate information in order to avoid the default expert-appropriate information. Third, certain strategies were more common than others. Common strategies included asking for information of a specific genre (e.g., "scientific articles"), describing the complexity of the desired information (e.g, "simple explanation"), and describing attributes of the user (e.g., "for a middle-school child"). Rare strategies included asking for information from a specific source (e.g., "YouTube videos"), describing the purpose of the information (e.g., "for a college-level paper"), and describing *unwanted* information (e.g., "not for a layperson"). Finally, certain strategies were much more common when requesting information for a novice versus expert. For example, describing the complexity of the desired information was more common for novice tasks (e.g., "simple", "easy to understand", "high level"), while mentioning the desired genre was more common for expert tasks (e.g., "scientific articles", "research papers", "technical reports").



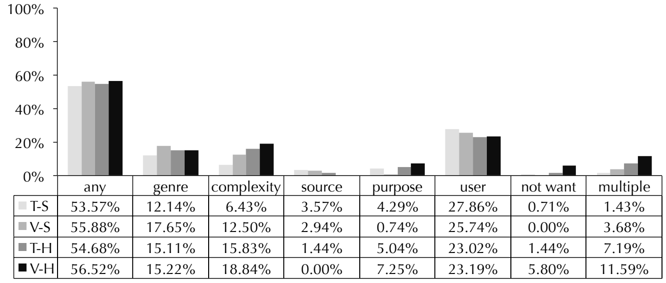
	genre	complexity	source	purpose	user	not_want	any
■ expert	81	15	2	10	4	5	105
■ novice	2	59	9	14	134	6	200

Figure 3: Distribution of strategies.

Strategies Across Medium and Target Conditions: Figure 4 shows the percentage of information requests associated with each strategy across medium and target conditions. We used a logistic regression to predict the presence or absence of each strategy given the information request's medium and target condition. We included a code for 'any' (the request had at least one strategy) and 'multiple' (the request had more than one strategy). Medium condition was not a significant predictor for any of the codes. Target condition was a significant predictor for 'complexity' (Wald $\chi^2(1) = 5.868, p < .05$) and 'multiple' (Wald $\chi^2(1) = 4.572, p < .05$). Participants were more likely to mention the complexity of the desired information and to employ multiple strategies in the human versus search engine condition.

Retrieval Performance Across Strategies: Figure 5 shows the average $\mathcal{P}$@5 values for information requests employing different strategies. Columns 'none', 'any', and 'multiple' denote information requests that did not employ any of our strategies, at least one, and more than one, respectively.

We used a multiple linear regression to predict $\mathcal{P}$@5 based the strategies present in an information request. We excluded 'multiple' and 'any' from this analysis as they are correlated with the other codes. In all cases, the absence and presence of a strategy was coded as 0 and 1, respectively. A significant regression equation was found ($F(9, 551) = 23.591, p < .001$), with an R^2 of .206. All strategies were



	any	genre	complexity	source	purpose	user	not want	multiple
▫ T-S	53.57%	12.14%	6.43%	3.57%	4.29%	27.86%	0.71%	1.43%
▪ V-S	55.88%	17.65%	12.50%	2.94%	0.74%	25.74%	0.00%	3.68%
▪ T-H	54.68%	15.11%	15.83%	1.44%	5.04%	23.02%	1.44%	7.19%
■ V-H	56.52%	15.22%	18.84%	0.00%	7.25%	23.19%	5.80%	11.59%

Figure 4: Percentage of search tasks employing different strategies across conditions: text/search engine (T-S), voice/search engine (V-S), text/human (T-H), and voice/human (V-H)

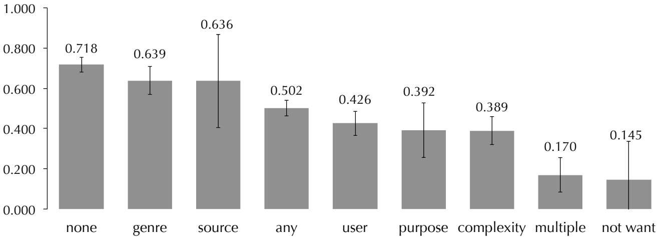


Figure 5: Mean P@5 performance across strategies.

found to *negatively* impact $\mathcal{P}$@5 performance. Those that were significant include 'complexity' ($\beta = -.256, p < .001$), 'purpose' ($\beta = -.249, p < .001$), 'user' ($\beta = -.252, p < .001$), and 'not want' ($\beta = -.437, p < .001$).

5 DISCUSSION AND CONCLUSION

In terms of RQ1, our results suggest two main trends. First, target had a greater effect than medium on participants' experience producing the request and their perceptions (Figure 1). Participants reported less difficulty and greater satisfaction and confidence in the human versus search engine condition (i.e., when they had greater expectations about the system's ability to respond to requests). Second, while the interaction effect was not significant, participants reported being the least confident and satisfied in the voice and search engine (V-S) condition.

In terms of RQ2, our results suggest three main trends. First, consistent with previous comparisons between voice and textual queries [3–6, 14, 19], our spoken requests were longer, had more complex grammatical structure (resembling natural language), and yielded worse retrieval results when issued *unmodified* to a commercial web search API. Second, the same trend was true for requests intended for a human versus search engine. Finally, both factors had a strong *additive* effect—information requests were the longest, most complex, and yielded the worst retrieval results in the voice and human condition (V-H).

In terms of RQ3, our results suggest four main trends. First, our qualitative analysis found six strategies adopted by participants where requesting information for a novice or expert. Second, target had a greater effect than medium in influencing participants to use certain strategies. When requesting information from a human versus a search engine, participants were more likely to mention the complexity of the information desired, more likely to describe the purpose of the information (not significant), more likely to describe *undesired* information (not significant), and more likely to use a combination of strategies (Figure 4). Third, all strategies hurt $\mathcal{P}$@5 performance when the request was submitted *unmodified* to a commercial web search API (Figure 5). Finally, the strategies associated with the *worst* retrieval performance (i.e., 'purpose', 'complexity', and 'not want') were *more common* in the human versus search engine condition.

To conclude, our results suggest challenges and opportunities for conversational search interfaces that enable more naturalistic interaction. Our results show that voice-input and users' raised expectations about the system can result in worst-performing queries. Both factors resulted in queries that were longer, had more natural language structure, and had fewer content terms (non-stopwords). Additionally, the target of the request (human vs. search engine) also influenced participants to use more elaborate ways of conveying a preference for novice- or expert-appropriate information, which also hurt performance. Recent work in IR has focused on automatically reducing voice queries in order to improve retrieval performance [1]. As queries evolved based on users' expectations, some of these "problematic" query terms may convey extra-topical dimensions of the task that should not be ignored, but rather used in some special way to favor certain types of content.

REFERENCES

[1] Jaime Arguello, Sandeep Avula, and Fernando Diaz. 2017. Using Query Performance Predictors to Reduce Spoken Queries. In *ECIR*. Springer.

[2] Carol L. Barry and Linda Schamber. 1998. Users' Criteria for Relevance Evaluation: A Cross-situational Comparison. *IP&M* 34, 2-3 (1998), 219–236.

[3] Fabio Crestani and Heather Du. 2006. Written Versus Spoken Queries: A Qualitative and Quantitative Comparative Analysis. *JASIST* 57, 7 (2006), 881–890.

[4] Heather Du and Fabio Crestani. 2003. Spoken Versus Written Queries for Mobile Information Access. In *MobileHCI*.

[5] Heather Du and Fabio Crestani. 2005. Spoken Versus Written Queries for Mobile Information Access: An Experiment on Mandarin Chinese. In *IJCNLP*. Springer, 745–754.

[6] Ido Guy. 2016. Searching by Talking: Analysis of Voice Queries on Mobile Web Search. In *SIGIR*. ACM, 35–44.

[7] Ahmed Hassan Awadallah, Ranjitha Gurunath Kulkarni, Umut Ozertem, and Rosie Jones. 2015. Characterizing and Predicting Voice Query Reformulation. In *CIKM*. ACM, 543–552.

[8] Jiepu Jiang, Wei Jeng, and Daqing He. 2013. How Do Users Respond to Voice Input Errors?: Lexical and Phonetic Query Reformulation in Voice Search. In *SIGIR*. ACM, 143–152.

[9] J. R. Landis and G. G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174.

[10] Kelly L. Maglaughlin and Diane H. Sonnenwald. 2002. User perspectives on relevance criteria: A comparison among relevant, partially relevant, and not-relevant judgments. *JASIST* 53, 5 (2002), 327–342.

[11] Taemin Kim Park. 1993. The Nature of Relevance in Information Retrieval: An Empirical Study. *The Library Quarterly* 63, 3 (1993), 318–351.

[12] Tefko Saracevic. 2007. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *JASIST* 58, 13 (2007), 1915–1933.

[13] Reijo Savolainen and Jarkko Kari. 2006. User's defined relevance criteria in web searching. *Journal of Documentation* 62, 6 (2006), 685–707.

[14] Johan Schalkwyk, Doug Beeferman, Françoise Beaufays, Bill Byrne, Ciprian Chelba, Mike Cohen, Maryam Kamvar, and Brian Strope. 2010. *Your Word is my Command: Google Search by Voice: A Case Study.* Springer US, 61–90.

[15] Milad Shokouhi, Rosie Jones, Umut Ozertem, Karthik Raghunathan, and Fernando Diaz. 2014. Mobile Query Reformulations. In *SIGIR*. ACM, 1011–1014.

[16] Amanda Spink, Howard Greisdorf, and Judy Bateman. 1998. From Highly Relevant to Not Relevant: Examining Different Regions of Relevance. *IP&M* 34, 5 (1998), 599–621.

[17] Anastasios Tombros, Ian Ruthven, and Joemon M. Jose. 2005. How users assess Web pages for information seeking. *JASIST* 56, 4 (2005), 327–344.

[18] Yunjie Xu and Zhiwei Chen. 2006. Relevance judgment: What do information users consider beyond topicality? *JASIST* 57, 7 (2006), 961–973.

[19] Jeonghe Yi and Farzin Maghoul. 2011. Mobile Search Pattern Evolution: The Trend and the Impact of Voice Queries. In *WWW*. ACM, 165–166.