Comparing User Approach to Complex Information Needs in Traditional Web Search and Conversational Search

Louise Leibbrandt , Claudia Hauff
TU Delft
L.G.Leibbrandt@student.tudelft.nl, C.Hauff@tudelft.nl

Abstract

Conversational search systems have recently gained popularity due to their effectiveness in performing small tasks and answering factoid questions. However, under complex search scenarios, these systems fail and fall back to techniques used by traditional search engines. As tasks requiring user learning are inherently complex, user knowledge is a powerful indicator of system performance under complex search and user ability to successfully interact with the information content. We present a study in which user knowledge gain and query formulation is explored in both the traditional web search and conversational search formats. Through crowd sourcing, 50 participants were recruited and conducted complex search sessions on either the traditional or conversational search medium. Through the use of a knowledge test given to the participant both before and after the search session, knowledge gain was determined. Furthermore, session queries and timestamps were tracked. It is found that participants have a significantly higher knowledge gain in traditional search, while conversational search sessions tended to have a higher number of queries and average query length.

1 Introduction

Conversational search is a relatively new field aiming to provide users a means of information retrieval through a dialogue based system. This system should mimic the behaviour of human driven communication processes and effectively satisfy an information need through a series of small exchanges between user and system [1].

Existing automated personal assistants have gained popularity due to their effectiveness in performing small tasks and answering factoid questions. However, when faced with a user wishing to satisfy a complex information need, current conversational systems fail and fall back to techniques used by traditional search engines [2]. These complex information needs can be defined by search topics that lead to a "a multi-step and time consuming process that requires multiple queries, scanning through many documents, and extracting and compiling information from multiple sources" [3]. In a traditional search engine these steps are performed by the user; the user types in a query, scans through the returned documents, retains bits of information and then reformulates their query. In a conversational search system, the user is still responsible for providing information seeking queries based on the information need and results provided by the system. However, tasks such as document scanning and excerpt extraction are automated and performed by the system. It is this distinction between the two forms of search that explains the failings of current conversational search systems and reveals its potential when implemented properly.

Complex search tasks require people to extensively interact with information. The user goal in these types of searches is to learn about a certain topic or to discover new information [4]. *Search as learning* (SAL) is an emerging field in information retrieval aimed to support these learning outcomes by considering user knowledge throughout a search session. Recent research in SAL for traditional web search systems has realized the importance of learning behaviour on system design. Research has primarily focused on measuring user knowledge gain throughout information search sessions [5; 6] and on improving user learning experience [7]. SAL has not yet been a topic of interest in conversational search. In its current state, most systems fail to successfully handle complex information needs, ultimately meaning that a lot of existing research is done on theoretical frameworks [8] rather than existing implementations. Studies relying on existing state-of-the-art systems such as Google Assistant discuss at length the "drawbacks and situations where it failed to respond properly" [9]. Current automatic agents are prone to providing irrelevant or incorrect responses. Co-referencing frameworks used to track context are also not yet advanced and are prone to failure. This causes frustration in users and removes credibility of answers provided by the system. Furthermore, as many of the existing conversational search systems are not open-source it is often difficult to say why and where these systems fail.

This research is aimed at filling the aforementioned gap by providing insight on knowledge gain and user behaviour in complex conversational search as apposed to traditional web search. Furthermore, it takes the approach of recreating an existing user study in traditional web search, applying it to both the traditional web and conversational mediums. The re-

search question explored within this study is as follows; *How do users approach complex information needs in a conversational search system compared to traditional web search?*. In order to explore this broad question, the user study aims to answer the following two sub questions.

RQ1: *How does a user's knowledge evolve when satisfying a complex information need?*
To further understand the current state of conversational systems, knowledge gain and evolution of queries within search sessions are explored. This was expected to be lower for conversational search as the task of document scanning and excerpt extraction performed by these systems is not yet advanced.

RQ2: *How do complex topics impact user query formulation?*
As query formulation is done by the user in both search systems, exploring user approach within each type can indicate the benefits each medium provides to query formulation. In traditional web search, the average query length is around 2.3 words [10] and queries often lack context. In conversational search users are inclined to provide more information through the dialogue based system. This research aims to reveal the extent to which conversational search can benefit complex query formulation.

To explore these questions, a user study was conducted on 50 participants recruited from a crowd sourcing platform. The study presented the users with a well defined complex information need and a search tool in which to conduct their search. Users were randomly assigned to either the traditional web or the conversational search format. The study employed a *knowledge test* corresponding to their given information need. This was given to the participant both before and after their search in order to quantify knowledge gain. Furthermore, queries and corresponding time stamps were logged in both search systems. It was found that users partaking in a traditional web search session have a 14.36% higher knowledge gain than those using the conversational tool. Furthermore, users in conversational search tended to send on average 7.33 more queries per session with an average length of 2.12 terms longer than in traditional web search.

2 Related Work

This research stems from two domains of related work; studies focused on (i) knowledge gain and (ii) query formulation. Furthermore, these two realms are explored for both the traditional web search and conversational search mediums. An extensive search did not yield any existing user studies within the realm of (i) or (ii) that primarily focuses on comparing the two search mediums.

2.1 Knowledge Gain

The emerging field of *search as learning* (SAL) has recently gained traction in *Information Retrieval* (IR). The Second Strategic Workshop on IR [11] recognized SAL as being a key research area that could lead to better system design. This sparked a number of studies focused on quantifying and understanding knowledge gain in current IR systems.

Traditional Search. Existing studies in traditional web search have focused on the impact of information needs on the knowledge gain of users [5], predicting knowledge gain [12] and studying within-session learning [13; 6]. Studies pertaining to the measure of user learning have the similar goal of reshaping systems to support learning outcomes. We now provide the different approaches of obtaining a numerical value for knowledge gain in order to support our study design choices. [5; 12; 6] all make use of a pre and post test setup to measure user learning. In [5; 12] this test is composed of a 10 to 20 TRUE / FALSE / I DON'T KNOW questions quiz. The test used in [6] differs in content and chooses 10 vocabulary knowledge questions to test the participants. [6] makes the decision of testing participants at intervals throughout the study. [5; 12] also make use of query length and term evolution to determine user knowledge gain. We adopt the methods used in [5; 12] and apply them in our study design.

Conversational Search. Studies pertaining to SAL in the conversational domain are lacking. In this subsection we attempt to explain the current gap in research within the realm of conversational search. Due to the novelty of conversational search systems, studies in this field mainly focus on user satisfaction and behaviour rather than user learning [14; 15; 9]. Current research on existing systems is still in the phase of system development for trivial tasks and queries. [16] gives an example of a recent study in 2020 where Dubiel et al. investigate the effect of different conversational strategies in goal-oriented tasks. Furthermore, user studies oriented around information seeking tasks often rely on Wizard of Oz systems rather than using existing chat bot technology [14; 15; 9; 16]. In these studies, participants are told that they are chatting with a bot when in fact the system is backed by a human.

2.2 Query Formulation

Within IR systems, *query formulation* is the action of combining terms into a question or statement that expresses an information need [17]. Query formulation remains a topic of interest as the "quality of queries submitted to IR systems directly affects the quality of search results generated" [10].

Traditional Search. Within the realm of traditional web search, there have been a wide range of studies focusing on query formulation with the intent of improving system design. These studies have focused on the effect of context on query formulation [18], different methods of query suggestion [19; 20] and deriving user intent from queries [21; 22]. Current search engines are based on decades of research and techniques such as query suggestion and result ranking have become complex and highly effective. However, in this user study we intend to compare the two search mediums in the essence of their design and we wish to eliminate this head start by keeping our search engine traditional. As query suggestion can have a significant impact on user behaviour [19], we do not incorporate it into our tool. We also make the decision of using the result ranking provided by the Bing Search API.

Conversational Search. Within this section, we motivate our decision for keeping the search tool typed rather than in-

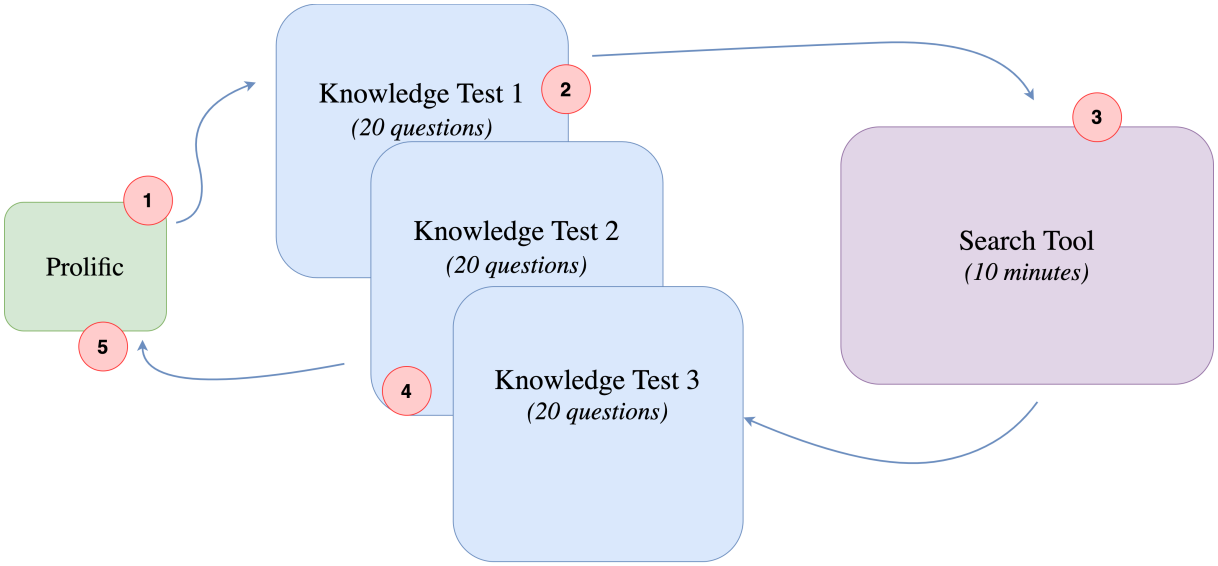

Figure 1: Workflow of participants in the experimental setup for the user study. (1) Participants are recruited from Prolific and are redirected to either the traditional or conversational Google Form. (2) Participants fill out a knowledge test on 1 of 3 topics that they are least familiar with. (3) Participants engage in a search session with either the traditional or conversational search tool. (4) Participants are again asked to fill out the knowledge test on the same topic. (5) Participants are redirected to Prolific to confirm participation.

corporating spoken queries. There are two main distinctions in query formulation for conversational search studies; those that focus on textual input and those on voice input. As explored in [23], the audio channel used in spoken conversational search introduces a magnitude of factors and complications. As the goal of this research is to compare traditional and conversational search in their core differences, only textual conversational systems are considered to minimize influence of speech factors.

3 Methodology

This chapter starts with an overview of the user study, followed by an in depth explanation on topic choice and the search tools used. Section 3.1 and 3.3 have a corresponding traditional web search and conversational search sub section to explain the specifications pertaining to each condition.

3.1 Study Design

Design choices for this user study are inspired by [5] in which Gadiraju et al. examine the knowledge gain of users given varying complex search topics in traditional search. This recent user study in SAL is chosen as it successfully quantifies knowledge gain and provides a clear defined and justified methodology. The study was conducted on 500 participant and 10 complex search topics using *SearchWell*, a search engine built on top of the Bing Web Search API. It primarily focused on the effect of search topic on user learning. We are only interested in the performance of complex search in the two mediums, therefore the user study in [5] is slightly altered to fit our needs. As opposed to 10 complex search tasks, the study is scaled down to just one chosen topic from these existing 10. In order to make this viable, an additional constraint is placed on topic selection; the user is given three topics of which he/she must choose the one which he/she is least familiar with. This decision follows the design choices of [12] as we are not interested in receiving responses from participants who are already expert on one of our topics and may not experience a significant knowledge gain. The number of participants are reduced from 500 to 50. Furthermore half conduct their search on our traditional web search tool and the other half is redirected to our conversational search bot. These tools are thoroughly explained in section 3.3. The study methodology in [5] did not specify a minimum search time but instead incentivized participants through the use of an extra bonus payment depending on the final test score. As we did not ethically agree with the method of basing reward on performance, this was altered in our workflow. We provide no bonus payment, and instead incentivized users by implementing a 10 minute search requirement. This amount was chosen based on the search times obtained in [5].

Workflow. Figure 1 captures the workflow of the participants taking part in the user study. In the first phase, workers are recruited from Prolific[1]. The user study is posted on Prolific with the title *A Study about Searching for Information*. Custom prescreening is applied with the following filters: Native English Speaker, Prolific Acceptance Rate $> 90\%$, Minimum Number of submissions 50+. Furthermore, the study is predicted to take 20 minutes and the workers are told that the study can be conducted on any device. Workers are given a redirect link that randomly assigns them to one of two Google Forms corresponding to either the conversational or traditional web search condition. Participants are then presented three search topics; NASA Interplanetary Missions, Altitude Sickness and Tornadoes. Further detail pertaining to these topics is given in section 3.2. Participants are then asked to

[1] prolific.co

Topic	Information Need
1. Tornado	In this task, you are required to acquire knowledge about the weather phenomenon that is called 'tornado'.
2. Altitude Sickness	In this task you are required to acquire knowledge about the symptoms, causes and prevention of altitude sickness.
3. NASA Interplanetary Missions	In this task, you are required to acquire knowledge about the past, present, and possible future of interplanetary missions that are planned by the NASA.

Table 1: Topics and corresponding information needs as given to the participants.

choose the topic that they are least familiar with. Based on this choice they are redirected to the corresponding knowledge test. In the second phase, participants fill out a 20 question *knowledge test* on their chosen topic without consulting the web for help. The questions take the form of topical statements. The participant is asked whether this statement is 'TRUE' or 'FALSE'. If the participant is not sure, he/she is asked to fill out the option 'I DON'T KNOW'. In phase 3 the participant conducts their search on either the traditional or conversational search tool, this is explained in further detail below. Once the participant feels that their information need is satisfied and they have spent at least 10 minutes searching, they are given the same *knowledge test* as provided in the second phase. If the user has completed all steps in the Google Form, they are provided with the redirect link to prolific.

Traditional Web Search. For participants redirected to the traditional Google Form, phase 3 in figure 1 employs a traditional search engine. Specifics on the tool designed for this can be found in section 3.3. Participants are required to spend at least 10 minutes searching. They are informed that if this requirement is not met, no reward will be received. This rule is further enforced by a built in timer into the search engine. The user is given a well defined search task as defined in section 3.2 and is provided with a link to the search engine.

Conversational Search. For participants redirected to the conversational Google Form, phase 3 in figure 1 employs a conversational bot. The conversational bot requires users to have a Telegram[2] account. Users are informed of this requirement at the start of the user study. As in the traditional search condition, participant are provided with an information need relating to their chosen topic. Participants are required to spend at least 10 minutes chatting with the bot. Details associated with the search bot are provided to the user and they are given the corresponding bot link.

3.2 Topics

Topics and corresponding knowledge tests[3] were chosen from the topics used by [5]. The study conducted by [5] differentiates between topics based on number of items in the knowledge test. This is to "attempt to feature varying scopes of information needs; relatively narrow ... as well as broad" [5]. The 20 question topics correspond to their definition of a broad search topic. As this study focuses on complex search, 3 broad search topics are chosen from this collection. These topics and their corresponding information needs can be found in Table 1.

[2]telegram.org

[3]sites.google.com/view/knowledge-gain

3.3 Search Tools and Data Collection

Here we provide details on the tools created for the two search mediums. Both the traditional tool and conversational tool are built on the same search engine, the Bing Web Search API. Figure 2 shows the traditional web search tool and Figure 3 the conversational search tool.

Traditional Web Search. The traditional web search tool is built using the Azure cognitive services REST api samples[4]. The bare bone Microsoft framework for Bing Web Search is used as starting point. Added features include logging user queries and timestamps. The engine also has a built in timer that informs participants of minimum remaining search time. Furthermore, the static website is published allowing the website to be accessed through a public url.

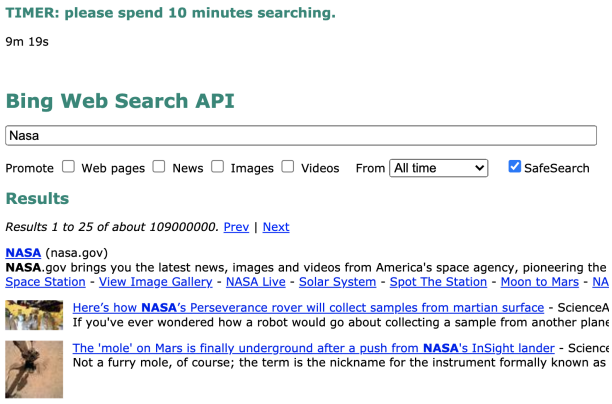

Figure 2: The traditional web search tool.

Conversational Search. Macaw[5] is the chosen tool for the conversational search system. Macaw is "an open-source framework with a modular architecture for Conversational Information Seeking research" [24]. It supports question answering and can integrate with a variety of interfaces and data sets. For this study, Macaw is integrated with the Telegram interface and a corresponding bot exists that can be accessed by participants. It is setup for question answering using DrQA and is capable of standard document retrieval. Questions are answered through the 'qa' mode and statements through the 'retrieval' mode. Queries and timestamps are logged using MongoDB.

[4]github.com/Azure-Samples/cognitive-services-REST-api-samples.git

[5]github.com/microsoft/macaw

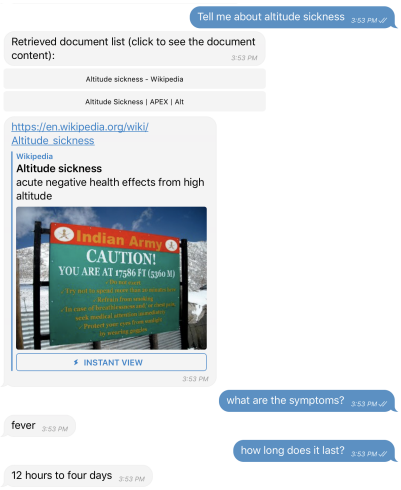

Figure 3: The conversational search tool.

4 Results

The results of the user study are presented here. The chapter starts with an analysis on knowledge gain in which the knowledge test data and query length evolution are examined. Query formulation is then explored and the corresponding metrics are presented.

The user study was conducted on 50 participants, 5 of which took part in a pilot version of the study and 4 of which failed to submit the Google Form. The data of these 9 participants was omitted from the analysis provided here. Of the 41 valid responses, 23 took part in the traditional search condition and 18 in the conversational condition. Of the 406 queries formulated by participants, 245 were sent to the conversational bot and 161 to the search engine. On average, 5.83 queries with an average length of 3.62 terms were sent per traditional search session and 13.16 queries with an average length of 5.72 terms per conversational search session. Participants spent on average 8 to 9 minutes searching in both mediums.

4.1 RQ1: Knowledge Gain

In this subsection we provide relevant data pertaining to the first research question: *How does a user's knowledge evolve when satisfying a complex information need?*

Following the techniques used in [5], knowledge gain is measured as the difference between a participants pre and post search test score. Responses of the form "I DON'T KNOW" are considered to be incorrect. Table 2 gives an overview of test scores for the 3 topics in the traditional web search condition and we present the average pre test score, post test score and estimated knowledge gain. Similarly, the results of the conversational search condition can be found in Table 3. 83% of users experienced some form of knowledge gain in the traditional search condition and 72% in the conversational condition. On average users experienced a 27.23% knowledge gain after performing their search on the traditional search engine, and a 12.87% knowledge gain after chatting with the conversational bot.

Topic	N	**Avg. Pre Score** *(%)*	**Avg. Post Score** *(%)*	**Knowledge Gain** *(%)*
Tornado	7	39.10±17.42	58.65±20.25	19.55±19.20
Alt. Sickness	11	43.54±29.12	77.03±12.95	33.49±22.84
NASA	5	30.53±13.62	54.74±10.26	24.21±15.91
Overall	23	39.36±23.20	66.59±15.10	27.23±20.42

Table 2: Average knowledge gain for users partaking in the traditional web search condition of the user study.

Topic	N	**Avg. Pre Score** *(%)*	**Avg. Post Score** *(%)*	**Knowledge Gain** *(%)*
Tornado	7	24.06 ± 9.54	41.35±17.83	17.29±14.92
Alt. Sickness	5	51.58 ± 6.86	65.26±17.69	13.68±20.81
NASA	6	19.30 ± 7.19	26.32 ± 5.77	7.02 ± 9.70
Overall	18	30.12 ± 8.11	42.98±14.88	12.87±15.43

Table 3: Average knowledge gain for users partaking in the conversational search condition of the user study.

Figure 4 visualizes the distribution of the data obtained in both search conditions. The data for the traditional search medium is distributed slightly higher than in the conversational search medium. It is interesting to note that the traditional search study contained a much wider distribution than that of the conversational search. This research suggests that users partaking in a traditional web search session have a knowledge gain that is 14.36% higher than those conducting their search in the conversational format.

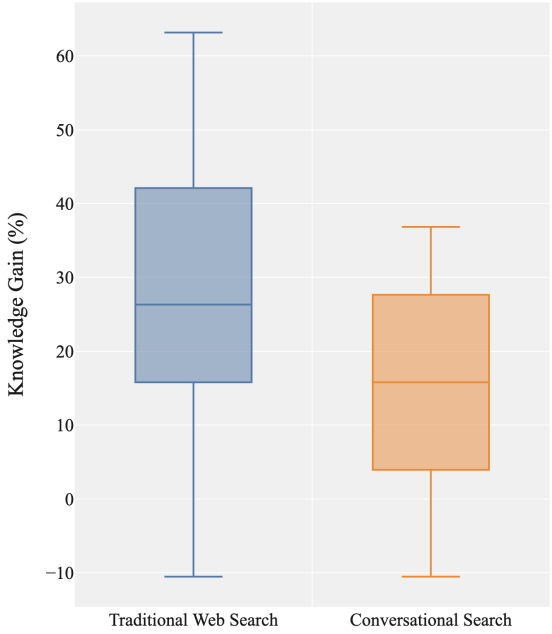

Figure 4: The distributions of the traditional and conversational search data obtained in the user study.

In order to validate that there is a significant difference between the two conditions, an unpaired t-test is conducted on the two sets of data resulting in a p-value of 0.01. As this is substantially smaller than the accepted value of 0.05, it can be concluded that the search medium has a statistical significance on knowledge gain.

As done in the analysis of [5], query evolution is examined. This is done by measuring query term length for both the first and last queries provided by the participant in their corresponding search session. This is done for all search sessions consisting of 2 or more queries within the two search mediums. It is important to note that some participants in the conversational track started their session with a greeting such as *Hello* or *How are you?*, these messages were not taken into account for the query evolution analysis. The corresponding results for the two search mediums can be found in table 4. In both cases, it is found that participants tend to increase the length of their search query as the session evolves. It is also found that this increase was 1.38 terms higher for participants in the conversational search track.

Search Medium	**Avg. First QL** *(terms)*	**Avg. Last QL** *(terms)*	**Difference** *(terms)*
Traditional Search	2.57 ± 1.41	3.96 ± 2.16	1.39 ± 2.11
Conversational Search	3.78 ± 2.07	6.56 ± 2.97	2.78 ± 3.39

Table 4: Average first and last Query Lengths (QL) for the two search mediums.

These results are presented here as query evolution is an indicator of users becoming more informed on a given topic throughout a search session. Most users partaking in a complex search session gain the ability to precisely formulate their information need. These results have been shown to hold in traditional search [5], and it is interesting to see that they translate into conversational search.

4.2 RQ2: Query Formulation

In this subsection we provide relevant data pertaining to the second research question: *How do complex topics impact user query formulation?* First a few examples are presented and discussed. This is followed by an overview of query related metrics obtained in the user study.

Table 5 illustrates an example of a typical search session in both mediums. It shows two users partaking in a complex search task corresponding to the topic of *Altitude Sickness*. This particular example highlights the general theme of user approach to queries in traditional search vs. conversational search. 79.1% of queries given to the traditional search engine were of the following format; *topic*[6] followed by *subtopic*. Users tended to provide the search engine with short precise queries containing only key terms related to their search. Furthermore, the majority of the participants provided their overarching search topic at the start of each query sent to the search engine. This translates into conversational search as 83.8% of the messages sent to the conversation bot contained the broad search topic somewhere within the message. The conversational search session provided in Table 5 gives an example of this. Although most messages are formulated as questions, users are not yet familiar with the *conversational* aspect of the bot and that it is not necessary to specify the topic within each new message sent.

[6]Here topic is defined to be any variation on the words tornado, altitude sickness and NASA.

Traditional Search	**Conversational Search**
Q1 *altitude sickness*	Q1 *What temperature does altitude sickness start?*
Q2 *altitude sickness symptoms*	Q2 *What are the symptoms of altitude sickness?*
Q3 *altitude sickness bleeding gums*	Q3 *Can your gums bleed with altitude sickness?*
⋮	⋮
Q8 *altitude sickness exercise*	Q10 *Does altitude sickness happen in high temperatures?*

Table 5: Example search sessions for the two mediums pertaining to the topic of Altitude Sickness.

The conversational search condition of the user study yielded some interesting messages which are presented below. These highlight some of the potential benefits the medium provides in contrast to traditional search.

Ex. 1 *anything else?*

Ex. 2 *what exactly?*

Ex. 3 *A yes would have sufficed! :)*

Ex. 4 *Tell me more please*

All four examples highlight a common behavioral theme among users in the conversational setting. Participants were quick to provide feedback to the system, whether this be asking for more information, as done in example 1 and 4, or asking for more specific information as in example 2. Example 3 is interesting as the participants gives direct feedback on the answer provided by the bot. These types of queries did not occur in the traditional search medium.

To further examine user approach to query formulation, three metrics are analyzed. These are the number of queries per search session, the average query length per session and the average session length. The results for the traditional search sessions can be found in Table 6 and the corresponding results for the conversational search sessions in Table 7. The results suggest that users performing a search session send on average 7.33 more queries with an average length that is 2.12 terms longer in conversational search as apposed to a traditional search. Conversational search sessions were on average 0.15 minutes longer than those in traditional search.

Topic	Avg. Q per session	Avg. QL per session *(terms)*	Avg. SL *(minutes)*
Tornado	7.00 ± 1.54	3.75 ± 1.28	8.64 ± 3.34
Alt. Sickness	7.73 ± 3.80	3.62 ± 1.38	8.53 ± 0.80
NASA	5.40 ± 2.88	2.52 ± 0.97	7.90 ± 2.18
Overall	5.83 ± 3.02	3.62 ± 1.38	8.43 ± 2.18

Table 6: Average number of queries per session, query length per session and Session Length (SL) for the traditional search.

Topic	Avg. Q per session	Avg. QL per session *(terms)*	Avg. SL *(minutes)*
Tornado	16.86 ± 3.89	5.5 ± 1.97	9.20 ± 1.10
Alt. Sickness	9.40 ± 1.82	7.00 ± 2.81	8.46 ± 1.68
NASA	13.33 ± 1.86	4.98 ± 2.17	7.96 ± 0.95
Overall	13.16 ± 2.82	5.74 ± 2.30	8.58 ± 1.25

Table 7: Average number of queries per session, query length per session and Session Length (SL) for conversational search.

The performance of a two tailed unpaired t-test reveals that the search medium has a statistical significance on both number of queries per session and average query length. However, it has no significant effect on session length.

5 Responsible Research

In order to reflect on the ethical aspects of this research, it is important to discuss both the integrity of the collected data and the consideration taken in the design of the human research. Lastly reproducibility is considered.

Data. The data collected was in no way manipulated, fabricated or falsified. In order to discuss the integrity of the data, omitted submissions are discussed and justified. From the 50 participant submissions, only 41 were used in our evaluation. 5 participants took part in a *pilot* study, however this revealed certain limitations of the study design and alterations were made. These 5 participants were rewarded but their data was omitted from further evaluations. 4 of the remaining 45 participants failed to submit the Google Form resulting in the loss of their knowledge test results. As this was a clear requirement of the user study, their query submissions were omitted and no reward was received. Lastly, any data that was omitted for certain calculations is discussed within the results section and will not be repeated here.

Human Research. In designing this user study, the *Netherlands Code of Conduct for Research Integrity* was consulted. Before the start of the user study, the participant is required to read and consent to their participation. He/she is told what the study will entail, the data that will be collected and the requirements that must be met in order to receive reward. Participants are ensured of their anonymity and instructed on how to withdraw their consent and data at any point throughout the user study. Within each step of the workflow, attention is placed on being transparent to the participant. There is a fixed reward that is not altered by their performance in the user study. The requirements are also repeated at the top of each section in the Google Form and timers are integrated into the search tools. No sensitive questions are asked and no identifiable data is collected.

Reproducibility. Consideration was taken in the design of the user study to allow reproducibility. All tools used are open source and accessible to the general public. Specifics pertaining to the tools designed are given in section 3.3. Furthermore, we provide a detailed methodology and provide the necessary links to the topics and knowledge tests used.

6 Discussion

In this section we provide some of the limitations of the users study. This is followed by a discussion on how to best interpret our results and what they mean for current conversational search systems.

6.1 Limitations

Here we discuss the main caveats and limitations of our user study design.

In order to simulate complex search, the users were provided with complex search topics. However, it was observed that the pre knowledge test had an impact on the types of queries provided to the system. Rather than doing exploratory search as prompted by the information need, participants formed their search around answering the test questions. Most queries contained the terms provided by the pre knowledge test. In this sense, the complex search was not organic and was impacted by the workflow of the study.

The decision was made to use Macaw as the foundation for the conversational search tool. This is as it is an open-source framework specifically designed for conversational information retrieval research. However, Macaw is relatively new and still has many limitations. Compared to current advanced conversational systems, it is slow and provides less than optimal answers. These answers are often only around 1-6 terms in length and can be incorrect. The framework is also prone to crashes and is not able to handle a large load of users. The connection with Telegram introduces a few problems, the main one being that URLs retrieved through the Bing API are often too long for the 64 bit limit placed on Telegram messages.

Participants were asked to spend at least 10 minutes searching in both search mediums. This is reflected by the results presented in Table 6 and 7, users in both studies spent on average 8-9 minutes searching for information. This restriction was placed in order to motivate participants to spend time with the system and to not rush through their search. However, we had not foreseen that the majority of users would stick to this exact search time. We acknowledge that this will have had significant impact on the session query data.

6.2 Interpreting Results

Here we provide insight into our results and when applicable, comparison is made to prior works.

The users partaking in the traditional search track experienced on average a 27% knowledge gain. This is significantly higher than the results obtained for these 3 topics in

[5] in which users experienced on average a 16% knowledge gain. However, the study in [5] did not implement a minimum search time and it was found that users spent on average 5.01 minutes searching. This is lower than the average session length of 8.43 minutes obtained in our results.

As predicted, users had a significantly higher knowledge gain in the traditional search medium than conversation search. However, the majority of users in the conversational search medium also experienced some form of knowledge gain. This is considerable as users are much less familiar with the conversational medium, and more notably because of the current limitations of the Macaw framework. It suggests that given more sophisticated software and user experience, conversational tools could successfully be used to carry out complex search.

The results presented in Table 6 and 7 show that users send more queries with a longer term length in a conversational search session compared to a web search session. However, this does not necessarily mean that users provide more information. While a user in a traditional search session may have provided the query "*altitude sickness symptoms*", a user in conversational search would formulate this as "*What are the symptoms of altitude sickness?*". It is also important to note that the conversational search queries contained many query reformulations as participants were not receiving satisfactory answers from the search bot.

7 Conclusions and Future Work

How do users approach complex information needs in a conversational search system compared to traditional web search? This was the research question proposed at the start of this project. To answer this, two sub question for complex search were formulated; *Q1: How does a user's knowledge evolve?* and *Q2: How is user query formulation impacted?*. These questions were investigated through a user study on 50 participants engaging in a complex search task. Roughly half of these participants conducted their search on a traditional search engine and the other through engaging with a conversational bot. In order to examine *Q1*, a method for quantifying knowledge gain was found. This is measured through the use of a knowledge test given to the participants both before and after engaging in their complex search. It is found that on average all participants experienced some form of knowledge gain, however, that participants using the search engine had a significantly higher percentage of knowledge gain than those using the conversational bot. In order to tackle *Q2*, query sessions were logged for both search mediums. Participants entered a significantly higher amount of queries with a higher average term length in the conversational search sessions compared to the traditional search. It is interesting to note that around 80% of all queries sent in both search mediums contained the overarching complex search topic within the query. This suggests that users are not yet aware of the "conversational" aspect of conversational search and that traditional search techniques are carried over into the conversational medium. Lastly, participants engaging in the conversational search were often inclined to provide feedback to the system through queries such as *anything else?* and *tell me more*. These types of queries did not occur in the traditional search medium.

The results presented in this study are promising for complex conversational search. Further research should focus on the effect of different conversational search techniques and systems on user learning and search behavior. This research also suggests that due to the conversational nature of these systems, users are willing to provide direct feedback to the search tool. This is an interesting byproduct of the research conducted here and if understood and utilized could have significant impact on conversational system design.

References

[1] Krisztian Balog, Lucie Flekova, Matthias Hagen, Rosie Jones, Martin Potthast, Filip Radlinski, Mark Sanderson, Svitlana Vakulenko, and Hamed Zamani. Common conversational community prototype: Scholarly conversational assistant. *arXiv preprint arXiv:2001.06910*, 2020.

[2] Alexandra Vtyurina. Towards non-visual web search. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pages 429–432, 2019.

[3] Georg Singer, Dmitri Danilov, and Ulrich Norbisrath. Complex search: Aggregation, discovery, and synthesis. *Proceedings of the Estonian Academy of Sciences*, 61(2):89, 2012.

[4] Ahmed Hassan Awadallah, Ryen W White, Patrick Pantel, Susan T Dumais, and Yi-Min Wang. Supporting complex search tasks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 829–838, 2014.

[5] Ujwal Gadiraju, Ran Yu, Stefan Dietze, and Peter Holtz. Analyzing knowledge gain of users in informational search sessions on the web. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, pages 2–11, 2018.

[6] Nirmal Roy, Felipe Moraes, and Claudia Hauff. Exploring users' learning gains within search sessions. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, pages 432–436, 2020.

[7] Wan-Ching Wu, Diane Kelly, Ashlee Edwards, and Jaime Arguello. Grannies, tanning beds, tattoos and nascar: Evaluation of search tasks with varying levels of cognitive complexity. In *Proceedings of the 4th Information Interaction in Context Symposium*, pages 254–257, 2012.

[8] Filip Radlinski and Nick Craswell. A theoretical framework for conversational search. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*, pages 117–126, 2017.

[9] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles LA Clarke. Exploring conversational search with humans, assistants, and wizards. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2187–2193, 2017.

[10] Giridhar Kumaran and James Allan. Adapting information retrieval systems to user queries. *Information Processing & Management*, 44(6):1838–1862, 2008.

[11] James Allan, Bruce Croft, Alistair Moffat, and Mark Sanderson. Frontiers, challenges, and opportunities for information retrieval: Report from swirl 2012 the second strategic workshop on information retrieval in lorne. In *ACM SIGIR Forum*, volume 46, pages 2–32. ACM New York, NY, USA, 2012.

[12] Ran Yu, Ujwal Gadiraju, Peter Holtz, Markus Rokicki, Philipp Kemkes, and Stefan Dietze. Predicting user knowledge gain in informational search sessions. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 75–84, 2018.

[13] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. Lessons from the journey: a query log analysis of within-session learning. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 223–232, 2014.

[14] Mateusz Dubiel, Martin Halvey, Leif Azzopardi, and Sylvain Daronnat. Investigating how conversational search agents affect user's behaviour, performance and search experience. In *The Second International Workshop on Conversational Approaches to Information Retrieval*, 2018.

[15] Sandeep Avula, Gordon Chadwick, Jaime Arguello, and Robert Capra. Searchbots: User engagement with chatbots during collaborative search. In *Proceedings of the 2018 conference on human information interaction & retrieval*, pages 52–61, 2018.

[16] Mateusz Dubiel, Martin Halvey, Leif Azzopardi, Damien Anderson, and Sylvain Daronnat. Conversational strategies: impact on search performance in a goal-oriented task. In *The Third International Workshop on Conversational Approaches to Information Retrieval*, 2020.

[17] Gondy Leroy, Jennifer Xu, Wingyan Chung, Shauna Eggers, and Hsinchun Chen. An end user evaluation of query formulation and results review tools in three medical meta-search engines. *International journal of medical informatics*, 76(11-12):780–789, 2007.

[18] Carla Teixeira Lopes and Cristina Ribeiro. Context effect on query formulation and subjective relevance in health searches. In *Proceedings of the third symposium on Information interaction in context*, pages 205–214, 2010.

[19] Sumit Bhatia, Debapriyo Majumdar, and Prasenjit Mitra. Query suggestions in the absence of query logs. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 795–804, 2011.

[20] Ryen W White and Gary Marchionini. Examining the effectiveness of real-time query expansion. *Information Processing & Management*, 43(3):685–704, 2007.

[21] Makoto P Kato, Takehiro Yamamoto, Hiroaki Ohshima, and Katsumi Tanaka. Cognitive search intents hidden behind queries: a user study on query formulations. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 313–314, 2014.

[22] Eldar Sadikov, Jayant Madhavan, Lu Wang, and Alon Halevy. Clustering query refinements by user intent. In *Proceedings of the 19th international conference on World wide web*, pages 841–850, 2010.

[23] J Trippas. Spoken conversational search: audio-only interactive information retrieval. 2019.

[24] Hamed Zamani and Nick Craswell. Macaw: An extensible conversational information seeking platform. *arXiv preprint arXiv:1912.08904*, 2019.