A Qualitative Analysis on Query Reformulation Types in Conversational Search Scenarios

Sophie Walboomers , Claudia Hauff
TU Delft
S.A.M.Walboomers@student.tudelft.nl, C.Hauff@tudelft.nl

Abstract

During conversational information retrieval, a user engages in a dialogue interaction with a search system in order to satisfy an information need. A profound understanding of the way in which users formulate and reformulate messages during this dialogue interaction, will aid the development and optimization of conversational search systems. This research analyses what query reformulation types are frequently used, and looks at how this differs between fact finding and information gathering search tasks. Existing research on query reformulation mainly focuses on traditional IR systems. The little research that has been conducted in a conversational context is based on interactions between humans, rather than incorporating a search engine. We are interested in conversational query reformulation in a text-based interface, using a web-based search engine. To this end, preliminary insights of an empirical user study are presented. On the basis of its results, a taxonomy of query reformulation types is defined. Additionally, significant differences are found between how fact finding and information gathering messages reformulate queries. These results contribute to a better understanding of the conversational search dialogue, which aids the further research and development of conversational search systems.

1 Introduction

Due to continuous developments in the field of natural language processing, the study on conversational information retrieval (IR) becomes increasingly relevant. During conversational search, users attempt to satisfy an information need by engaging in a dialogue interaction with the search system (Radlinski and Craswell, 2017). This dialogue interaction distinguishes conversational systems from traditional IR.

In this research, a query is defined to be reformulating when it is sent after the initial query during one search session, therefore building this cumulative information need. This research analyses the reformulation of queries during conversational search. This contributes to a more profound understanding of how users engage in the conversational search dialogue. This will aid the further development and optimization of conversational information seeking systems.

There have been an abundance of studies on query reformulation in traditional IR scenarios. Many of these studies establish taxonomies of different query reformulation types, and classify querying data accordingly (e.g. Huang and Efthimiadis (2009) and Liu and Gwizdka (2010)). While these taxonomies can partly be applied to conversational systems, they do not take into account important conversational characteristics. That is, they fail to incorporate how conversational systems should build a cumulative information need.

In conversational scenarios, the research into query reformulation is very limited. The small number of studies that have been conducted in this area are based on interacting humans, rather than incorporating a search engine (e.g. Trippas et al. (2017) and Qu et al. (2018)). These studies, too, provide some insight in conversational search and its query reformulation patterns.

Currently, no research exists that extensively analyses the cumulative process of query reformulation using a conversational IR system. This research aims to address this knowledge gap by taking an approach similar to that of Liu and Gwizdka (2010). Unlike the research by Liu and Gwizdka (2010), which was conducted in the domain of traditional IR, this research will be conducted in a conversational context. To be more precise, this research will analyse how users reformulate queries during a conversational search session. To this end, this research will establish a taxonomy for conversational query reformulation types, and analyse the differences in query reformulation between fact finding and information gathering information needs.

As outlined by Liu and Gwizdka (2010), providing the user with pre-defined search tasks sheds light on why specific types of query reformulation are used. The distinction between fact finding and information gathering information needs, was defined by Kellar et al. (2007):

> *Fact finding* is defined as a task in which you are looking for specific facts or pieces of information.
>
> *Information gathering* involves the collection of information, often from multiple sources. Unlike fact finding, you do not always know when you have completed the task, and there is no one specific answer.

Similar to the aforementioned research, this research will answer the following research questions:

RQ1: What query reformulation types are frequently used by searchers in a conversational search scenario?

RQ2: How does the frequency of each query reformulation type vary among fact finding and information gathering search tasks?

These questions are addressed by performing an empirical user study, designed to understand the types and patterns of query reformulation in a text-based, web-based, conversational IR system. To answer both questions, a taxonomy of query reformulation types is established, based on existing research combined with the data gathered from the user study. Next, the distributions of fact finding and information gathering messages as found by the user study are compared, where significant differences are found.

2 Related Work

In order to be able to classify query reformulation types, a taxonomy is established in Section 4. This taxonomy is built on existing research, as well as on the qualitative analysis of the data collected from the user study. This section elaborates on existing taxonomies and their properties.

2.1 Taxonomies for Traditional IR

Query reformulation taxonomies for traditional IR have been extensively addressed. For example, Teevan et al. (2007) describes a taxonomy of repeat queries based on traditional search web logs, with emphasis on the combinations of querying and clicking data. Jansen et al. (2007a) and Jansen et al. (2007b) defined a taxonomy and studied patterns of query modification by analysing web query logs. Furthermore, Liu and Gwizdka (2010) formally defined a taxonomy for text-based query reformulations, and analyse the influence of different search tasks. Another example is the research by Huang and Efthimiadis (2009), which formally defines a taxonomy that classifies text-based query reformulations. This taxonomy focuses on the semantic rather than syntactic query reformulations. For example, differences in query length are incorporated, rather than commenting on how the meaning of the query changes.

2.2 Taxonomies for Conversational IR

Unlike query reformulation in traditional IR scenarios, query reformulation in a conversational context has hardly been researched. Only a few taxonomies classifying conversational querying data exist.

Two observational studies by Trippas et al. (2017) and Trippas et al. (2018), qualitatively analysed query reformulation types and patterns in spoken querying data. To elaborate, Trippas et al. (2017) classified information retrieval interactions into themes. Furthermore, Trippas et al. (2018) defined a taxonomy based on query characteristics. Both studies were based on interactions between pairs of people communicating verbally, instead of on text-based human-computer interaction. A third study conducted in a conversational context was conducted by Qu et al. (2018). This research presents a query reformulation taxonomy and analysis based on the MSDialog dataset. Even though the dataset is conversational and text-based, it is based on interactions between humans on an online technical support forum. The information needs, and thus query reformulation characteristics in this context are very dissimilar to those occurring when using a web-based search engine.

2.3 Required Properties

In order to be suitable for direct re-use in this research, a taxonomy should meet several requirements. Firstly, the taxonomy should be well-defined and unambiguous. This is necessary for the classification to be reproducible. Secondly, the taxonomy should be defined with a conversational environment in mind. As previously discussed, the taxonomies for a traditional IR scenario lack some characteristics fundamental to conversational IR. For example, instead of keywords, natural language might be used more often during conversational IR than during traditional IR. Additionally, in conversational IR, queries should be interpreted in the context of previous messages, building a cumulative information need over numerous queries. Furthermore, the taxonomy should be based on a search scenario with a text-based interface, as opposed to voice-only. Finally, the taxonomy should be based on data collected from human-computer interactions, contrary to interactions between humans.

Table 1 summarizes which of these requirements are met by each taxonomy discussed in this section. None of the discussed taxonomies comply with all of the aforementioned conditions. This means none of the taxonomies are suitable for direct re-use in this research. However, the aforementioned taxonomies will be used as a basis for defining a taxonomy that is appropriate for this research specifically. This new taxonomy is defined in Section 4.

3 Methodology

A task-based user study was performed using a conversational IR system. Section 3.1 lays out the characteristics of this conversational system. Furthermore, Section 3.2 elaborates on the search tasks and Section 3.3 outlines the experimental setup. Finally, Section 3.4 comments on the data annotation process.

3.1 System Characteristics

To integrate the web-based, text-based and conversational nature of the research, the recently development Macaw was used. Macaw is an open-source framework with a modular architecture for conversational information seeking research (Zamani and Craswell, 2019). Macaw allows developers to build an IR system by connecting an interface, choosing a retrieval mode (from a dataset or using a search engine) and choosing an answer selection and results generation model. The result was a conversational search system, visualized in Figure 1B. The way in which Macaw interacted with the rest of the experimental setup is visualized in Figure 1C.

The interface that was implemented for this study was the Telegram Bot API[1]. This is an HTTP-based interface, with

[1] telegram.org

Authors	Taxonomy is well-defined	Taxonomy for a conversational context	Taxonomy for a text-based interface	Taxonomy for human-computer interaction
Teevan et al. (2007)			✓	✓
Jansen et al. (2007b), Jansen et al. (2007a)	✓		✓	✓
Huang and Efthimiadis (2009)	✓		✓	✓
Liu and Gwizdka (2010)	✓		✓	✓
Trippas et al. (2017), Trippas et al. (2018)		✓		
Qu et al. (2018)	✓	✓	✓	
This research	✓	✓	✓	✓

Table 1: Existing taxonomies and their characteristics

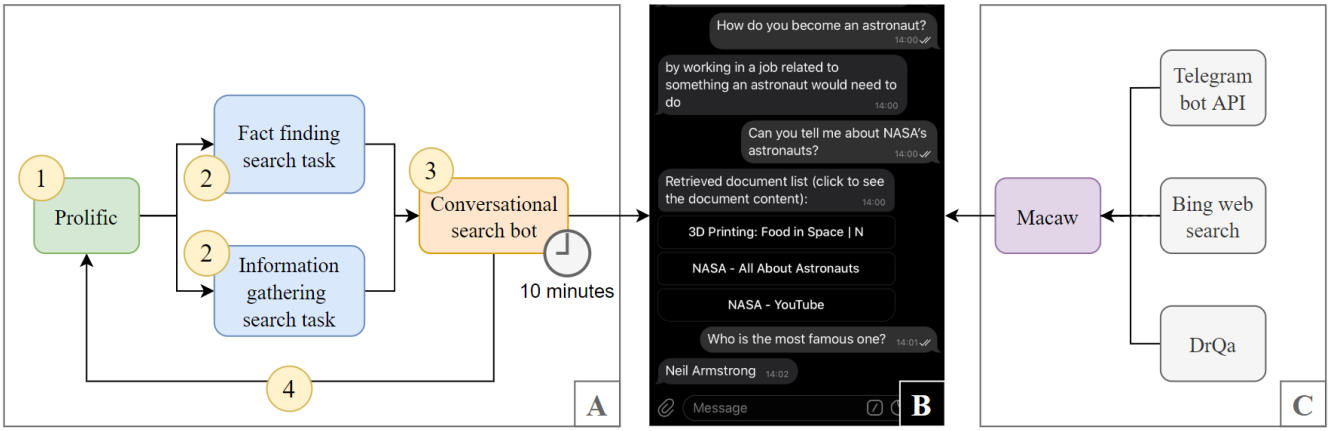

Figure 1: Overview of the workflow of participants (A), chat bot interface (B), and the implementation of the search system (C). The workflow of the participant is as follows: (1) Participants are recruited from the Prolific platform. (2) Participants are redirected to one of two Google Forms, with either a fact finding or information gathering search task. (3) Participants engage in an 10-minute information seeking chat session with the chat bot to satisfy the assigned information need. At the end of the search session the participants gets a completion code. (4) Participants return to Prolific, and claim their reward using the completion code.

which a chat bot was created. Via the Telegram application, participants engaged in a text-based chat conversation with the bot. The retrieval mode of the IR system made use of Bing web search as a search engine. The answering model used, was the conversational DrQa question answering model[2].

The reason for using DrQa and Bing web search, is that both were established to integrate well with Macaw. The reasons for using a text-based interface, as opposed to a speech-only interface, are multifold. Firstly, voice-based search systems by the major search engines do not support query reformulation by partial modification (Sa, 2016). In other words, voice-based query reformulation is fundamentally different from the text-based variant. Additionally, the interference of system recognition errors in the query formulation data are eliminated by using a text-based input. On the other hand, typing errors could be introduced. These could possibly be taken up in the taxonomy, on which will be elaborated in Section 4. The last reason, is that speech-only results presentation is still in its early stages of development (Trippas et al., 2015). The difficulties imposed by speech-only results presentation would distort the query reformulation data.

[2]https://github.com/facebookresearch/DrQA

3.2 Search Tasks

Similar to the approach of Liu and Gwizdka (2010), the research collected task-based querying data, focusing on two different types of search tasks. To elaborate, the users were presented with either a fact finding or an information gathering search task. The instances of search tasks as formulated by Liu and Gwizdka (2010) were not found suitable for direct re-use in this research, since they required a very specific audience. Therefore, based on the aforementioned definitions of fact finding and information gathering tasks, new search tasks were created. The search tasks were formulated as follows:

- Fact finding: "Ask the bot about the name of the latest NASA Interplanetary Mission. Then, find out what this mission was about."
- Information gathering: "Find out information about NASA Interplanetary Missions."

3.3 Experimental Setup

The workflow of the user study will now be outlined. A visual representation of this workflow can be seen in Figure 1A, where the steps of the participants are numbered 1 through 4.

The research was performed via the platform Prolific[3]. The platform allowed to recruit suitable participants for the research. Participants were recruited using three pre-screening conditions: participants should be native English speakers, have a Prolific acceptance rate of at least 90%, and have done at least 50 prior submissions. These criteria ensure a reliable and representative sample was recruited. Prolific users that met the pre-screening conditions were presented with some details about the study, the participant requirements and the monetary compensation that would be rewarded after taking part in the study.

After accepting the study in Prolific, participants were presented with a link for the study. This link redirected the participant randomly to one of two Google Forms. The Google Forms were designed to guide participants through the study. One of the Google Forms was used in the study on fact finding information needs, the other Google Form was used in the study on information gathering information needs. This means one of the Google Forms assigned the participant a fact finding search task, the other assigned an information gathering task. The two Google Forms differentiated only in the part where the search task was assigned, and were identical otherwise.

The identical first page of both Google Forms included a consent form with the terms and conditions of the study. This consent form will be elaborated on in Section 6. As per the guidelines of setting up a Google Form provided by Prolific, the first page also included a verification of the participant's native English language and a field to enter the participant's Prolific ID. The latter was necessary for the participant's reward post completing the study, and was strictly used for this purpose.

Following this first page, both Google Forms guided the participant through the steps to set up a Telegram account, if the participant did not already have one. Next, the participant reached a page with some instructions about the bot they were about to chat with, a link leading to this bot, and a search task. One of the Google Forms presented the participant with the fact finding search task, the other one presented the information gathering search task.

From the findings of a pilot experiment, each participant was instructed to spend at least 10 minutes trying to interact with the bot. In the pilot, some participants interpreted this instruction by asking one question to the chat bot, and then spending 10 minutes browsing the internet directly. Therefore, a guideline was given to send about 15 messages to the bot, or to stop when the participant felt the information need was satisfied. All user interaction with the system was logged.

3.4 Data Annotation

Upon completion of the study by all participants, the data annotation process started. Every message that was sent by a participant was individually assessed and annotated according to the taxonomy defined in Section 4. Because of the manual annotation of a large quantity of messaging data, a manner was sought to verify and maintain correctness of this annotation process. To this end, some verification rules were set up on the basis of the taxonomy. Next to the verification of the data annotation, these rules aid to develop a basic intuition about the different types of the taxonomy and their relationships. These rules are summarized in Section 4.4.

[3] www.prolific.co

4 Defining a Taxonomy

In order to classify query reformulation data, a taxonomy had to be defined. This newly defined taxonomy was built on existing taxonomies as discussed in Section 2, as well as on the qualitative analysis of the messaging data as obtained from the user study.

The taxonomy is visualised in Figure 2. This figure serves as a useful tool in order to understand the relationships between the taxonomy types, and it will be repeatedly referenced throughout this section and the rest of the paper. In the figure, each taxonomy type is specified fully, and a shorthand notation is given. Throughout this section and the rest of the paper, each type will be referenced using its shorthand notation. In addition to visualizing the hierarchical relationship between the categories, Figure 2 also contains information about the results of the user study. On this information will be elaborated in Section 5.

Table 2 includes examples of the taxonomy types, that were taken from the data of the user study. Some of the taxonomy types do not require an example. These self-explanatory types are omitted from Table 2. The remainder of this section will elaborate on each type included in the taxonomy of Figure 2, and refer to the relevant examples from Table 2.

4.1 Classification of all messages

As previously mentioned, reformulating queries are defined to be all queries following the initial query during one search session. As seen in orange in Figure 2, the reformulating queries make up an important part of the taxonomy. However, the taxonomy initially classifies all messages sent by the user, beyond only the reformulating queries. This is done so that the query reformulation data can be interpreted in the broader context of all messages sent by the user. The remainder of this subsection outlines the classification of all messages sent by the user, beyond just the reformulating queries. Section 4.2 will go into the classification of the reformulating queries in more detail.

Each of the messages sent by the user was classified as one of two mutually exclusive groups: either natural language message (`NatLang`) or keywords (`Keyword`). The category `NatLang` contains all messages where the user typed out the message as one would naturally verbalize it in a conversation. An example is given by query A. `Keyword` messages are all messages where the participant conceptualized their information need and messaged only a keyword query to the bot. An example is given by query B.

Furthermore, a distinction was made between non-querying messages (`NonQ`) and querying messages (`Query`). `NonQ` contains all messages that were not sent with the intent of gathering new information. Instead, these were informing messages. This can be greeting the bot (query C), thanking the bot (query D), giving positive or negative feedback to the

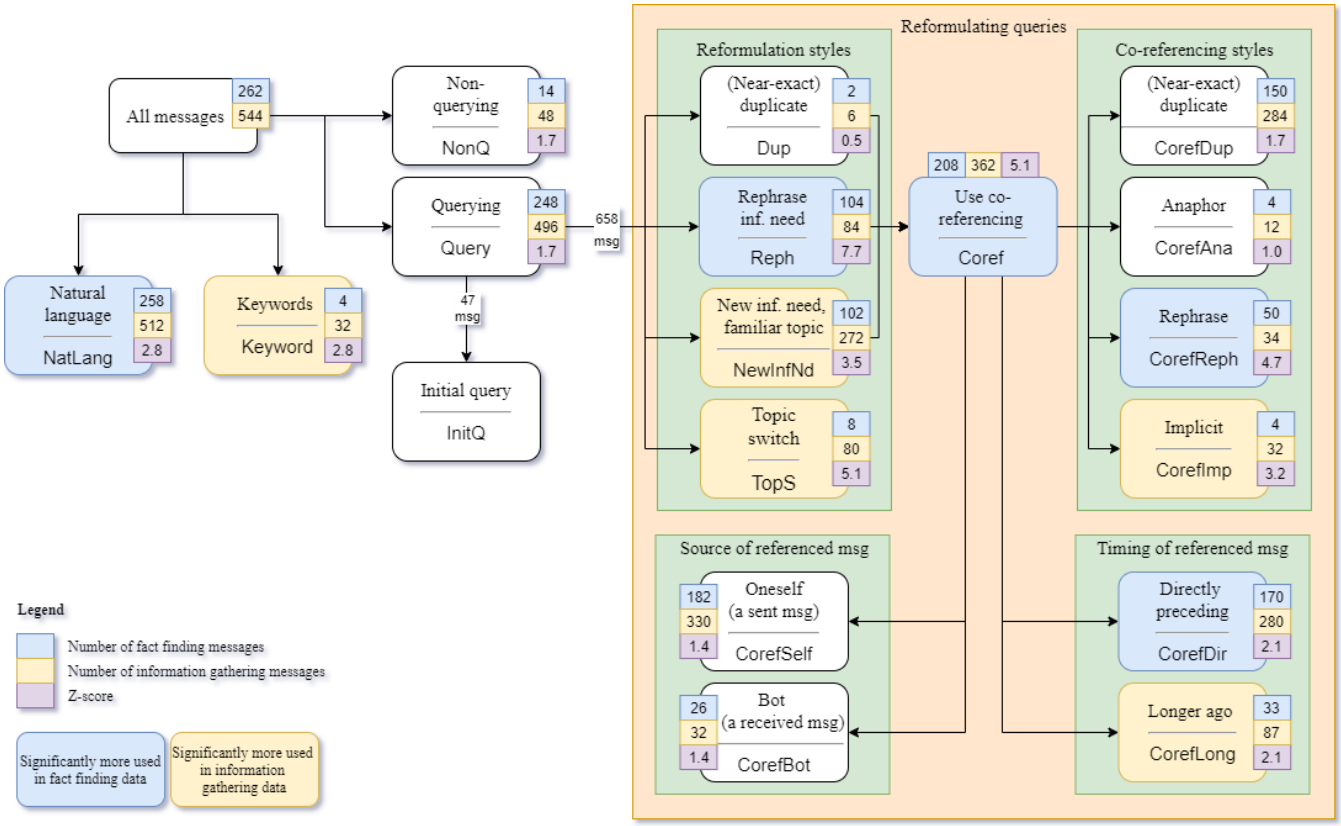

Figure 2: A graphical representation of the taxonomy as defined in Section 4. Each type is specified by a full description and a shorthand notation. Additionally, some results are included, on which is elaborated in Section 5. That is, the distribution of fact finding and information gathering messages across the taxonomy and Z-scores are included. Z-scores ≥ 1.96 are considered statistically significant. Types that occurred significantly more in fact finding data are marked blue. Types that occurred significantly more in information gathering are marked yellow.

bot, or any other messages that do not directly express an information need. In existing taxonomies, these messages were often classified as `junk` or `others`. `Query` contains all other messages. These messages are information-seeking, so they were sent to satisfy a certain information need.

The initial query (`InitQ`) is the first querying message sent by the user. It is the first message that carries an information need, similar to the definition of `Turn 1` by Trippas et al. (2017). When subtracting the `InitQ` from the `Query` messages, what remained were the reformulating queries.

4.2 Classification of reformulating queries

This subsection will elaborate on the remaining reformulating queries, depicted in orange in Figure 2. The reformulating queries were classified into four reformulation styles. These styles are (near-exact) duplicate (`Dup`), rephrase information need (`Reph`), a new information need about a familiar topic (`NewInfNd`) and topic switch (`TopS`). On each of these categories will now be briefly elaborated.

Firstly, `Dup` includes all queries that are (almost) identical to a message that has been previously sent by the user. To elaborate, this includes all queries that are exactly identical to a previous query, and queries that only change punctuation or capitalization of a previous query. Some of the existing taxonomies mention separate categories for correcting spelling/typing errors and making changes in whitespace. Interestingly, changes of these types were not seen in the messages collected from the user study. However, if this would have been present, correcting spelling/typing errors and changing whitespace would also fall under the category of `Dup`. An example is given by query F, which is a near-exact duplicate of query E.

A query of the type `Reph`, is a query that has the same information need as a previously sent query, but with a different formulation. This includes for example queries that substitute words for synonyms, change the word order, switch between the singular and the plural or change the verb tense compared to previous queries. In other words, the query is a syntactic change of a previous query, while preserving the same semantic meaning (information need). In existing taxonomies, this type was often referred to simply as `reformulation`. Sometimes, existing taxonomies subdivide this category into different subtypes. For example Huang and Efthimiadis (2009) mentions the types `word reorder`, `word substitution` and `abbreviation`. However, we choose to generalize this, so that one type captures all cases where an information need was preserved, but formulated differently. An example is

Query Type	Example
NatLang	Query A: Hi, what can you tell me about the most recent NASA mission?
Keyword	Query B: NASA mission
NonQ	Query C: Good morning
	Query D: Fantastic, thanks
Dup	Query E: What does NASA stand for?
	Query F: What does Nasa stand for?
Reph	Query G: What is NASA's latest mission?
	Query H: What is the latest mission of NASA?
NewInfNd	Query I: How many space stations are in orbit?
	Query J: What are they called?
CorefDup	Query K: What is the name of the latest NASA mission?
	Query L: When did the latest NASA mission take place?
CorefAna	Query M: What is the name of the latest NASA mission?
	Query N: When did it take place?
CorefReph	Query O: What is the name of the latest NASA mission?
	Query P: When did the mission take place?
CorefImp	Query Q: Can you tell me something about space research?
	Query R: How are conclusions made?

Table 2: Examples of the relevant taxonomy types

query H, which rephrases query G by changing the word order.

In contrast to the queries that preserve an information need, are the queries of the type NewInfNd. This type includes all queries that ask a different question compared to previous questions. However, this query revolves around a subject that has already been addressed by a previous query, or around a subject that has been previously mentioned in an answer given by the bot. In other words, this query changes the semantics of a previous query, while not changing the topic. This means that this messaging type always uses some form of co-referencing, on which will be elaborated later. An example is given by query J, which expresses a new information need about the same topic as addressed in query I, i.e., space stations that are in orbit.

The last reformulation style is TopS. This includes all queries that ask a question that has not been asked before, about a topic that has not been discussed before. Since messages of this style address an entirely new topic, they never reference a previous message. In existing taxonomies, this type was often referred to as new or change.

4.3 Classification of co-referencing queries

As previously mentioned, conversational information seeking involves a dialogue interaction between the user and the search system. As in a natural dialogue, conversational systems must be able to build a cumulative picture of the user's information need, without the user constantly having to repeating it (Radlinski and Craswell, 2017). This entails the use of co-referencing established entities from previously sent messages. As stated by Ng and Cardie (2002), "co-reference resolution refers to the problem of determining which noun phrases refer to each real-world entity mentioned in a document". The distinction between different co-referencing types is seldom mentioned in existing taxonomies. However, co-reference resolution is an important and active topic of research in the domain of conversational search (Rahman and Ng, 2009). This is the reason for incorporating the classification of co-referencing in this taxonomy. In this taxonomy, co-referencing (Coref) includes all queries that refer to an entity from an earlier message.

Each Coref message was classified according to co-referencing style, the source of the referenced message, and the timing of the referenced message. The remainder of this subsection will elaborate on each of these three distinctions.

Co-referencing styles

Firstly, a distinction is made between four different co-referencing styles. This category is based on the form of the referenced phrase. Each of four variants will now be discussed. In the category of co-referencing using a (near-exact) duplicate (CorefDup), fall all references that exactly repeat the referenced phrase, or change only capitalization or punctuation of the phrase. Note the difference between the reformulation style Dup and the co-referencing style CorefDup. The former is about repeating an entire message, whereas the latter is about repeating the referenced phrase in the co-reference. For example, query L refers to the phrase 'latest NASA mission', which was mentioned earlier by query K. Query L literally repeated the phrase, fully writing it out.

A second type of co-referencing includes are all messages that use an anaphor (CorefAna) to refer to a phrase from an earlier message. Query N is an example where the anaphor 'it' is used in order to refer to the antecedent 'the latest NASA mission' from query M.

Co-references of the type rephrase (CorefReph), are all references that rephrase or reformulate the referenced phrase in any other way. Note again the difference between the reformulation style Reph and the co-referencing style CorefReph. To elaborate, Reph is about rephrasing an earlier question. CorefReph, on the other hand, is about rephrasing only the phrase to which is referred in the co-reference. CorefReph

includes, for example, abbreviating or expanding the referred phrase, changing the word order of the referred phrase, using synonyms for the referred phrase, leaving out words of the referred phrase, or adding words to the referenced phrase. Query P presents an example, since it refers to 'the latest NASA mission' from query O, by rephrasing it to 'the mission'.

Lastly, implicit co-references (`CorefImp`) are those queries that refer to a previous message without using any signal words, phrases or repetitions. Query R is an example of this, as it refers to 'conclusions about space research' from query Q. However, it does so implicitly.

Source and timing of the referenced message

In addition to co-referencing styles, two more distinctions of co-referencing messages were made. These distinctions look at the message to which was referred. To begin with, a distinction was made between the source of the referenced message. In other words, the question was answered: 'what was the source of the message to which this co-reference is referring?' There are two possible sources of the referenced message: the user and the bot. To elaborate, a co-reference to a message of oneself (`CorefSelf`), is a reference to the user's own message. The other option is a co-reference to a message of the bot (`CorefBot`), which was received by the user.

Finally, co-referencing types were classified according to the timing of the referenced message. In other words, the question was answered: 'what was the timing of the message to which this co-reference is referring?' A distinction was made between a message that was directly preceding the co-reference, and a message that occurred longer ago. To elaborate, a co-reference to a directly preceding message (`CorefDir`), is a reference to the last message that was either sent or received. A co-reference to a message longer ago (`CorefLong`), is a reference to any earlier message that is not directly preceding the reference.

4.4 Rules

As previously mentioned, the definition of the taxonomy allowed some basic rules to be set up to verify the data annotation process. These rules are as follows:

1. Every message is in one of the following mutually exclusive types: `NatLang` or `Keyword`.
2. Every message is in one of the following mutually exclusive types: `Query` or `NonQ`.
3. Every message of the type `Query` but not of the type `InitQ`, is also classified as one of the following mutually exclusive types: `Dup`, `Reph`, `NewInfNd` or `TopS`.
4. Every message of the types `Dup`, `Reph` or `NewInfNd`, is also classified as `Coref`.
5. No message of the type `TopS` is classified as `Coref`.
6. Every message of the type `Dup` is also classified as `CorefDup`.
7. Every message of the type `Coref` is also classified as one of the following mutually exclusive types: `CorefDup`, `CorefAna`, `CorefReph` or `CorefImp`.
8. Every message of the type `Coref` is also classified as of one of the following mutually exclusive types: `CorefSelf` or `CorefBot`.
9. Every message of the type `Coref` is also classified as of one of the following mutually exclusive types: `CorefDir` or `CorefLong`.

These rules follow directly from the way the taxonomy types are defined. This means they are very intuitive, and are easily understood with the help of Figure 2.

5 Results

This section will elaborate on the results of the user study and the statistical significance thereof. In Section 5.1, general information about the collected data is outlined. Additionally, Section 5.2 and 5.3 include the results in context of the two research questions. Finally, some additional results regarding the use of co-referencing are outlined in Section 5.4. Later, in Section 7 these results will be discussed in light of caveats and limitations of the study. Additionally, in Section 7 the results will be interpreted in the context of existing research.

Throughout the remainder of this section, all statistical tests were executed using a two proportion Z-test with a significance level of $\alpha = 5\%$. This means that a result with a Z-score of ≥ 1.96 was found to be significant.

5.1 Collected Data

In total, 47 participants voluntarily agreed to participate in the study, and completed it successfully. As outlined in Section 3, participants of the user study were randomly divided over the two studies. 24 participants were redirected to the research on fact finding search tasks, and 23 participants were redirected to the research on information gathering search tasks. In total, 1520 back-and-forth messages were collected. 806 of these messages were sent by a participant, the other 714 of these messages were replies sent by the bot. An average of 17.1 message was sent per user per search session. Each search session lasted on average 9.3 minutes.

All messages sent by the user were analysed and annotated according to the taxonomy defined in Section 4. The results of this annotation process are summarized in Figure 2.

5.2 General Distribution of Messages

To address **RQ1**, a taxonomy of query reformulation types was built in Section 4. As previously outlined, not only query reformulation types, but rather a classification of all messages was taken up in this taxonomy. This subsection elaborates on the general distribution of messages over the defined taxonomy.

96% of messages sent by participants were of the type `NatLang`, and 4% were `Keyword` queries. Additionally, 8% of all messages were classified as `NonQ`, opposed to 92% `Query` messages. Of these `Query` messages, 47% were of the type `InitQ`. As aforementioned, the query reformulation data includes all messages that were not identified as `NonQ` or `InitQ`.

The reformulating queries were classified in four query reformulation styles. Only 1% of the reformulating queries were of the reformulation style `Dup`. This means almost no

queries were simply repetitions of a previous message. A majority of 57% of the queries were classified as NewInfNd. Additionally, 29% of the queries were of the type Reph and 13% were of the type TopS. This means the majority of the reformulating queries posed a new question, while sticking to the same general topic.

According to the definition of the taxonomy, all querying messages that are not of the type TopS should contain a form of co-referencing. This results in that most of the messages, as much as 87%, were classified as Coref. Among the Coref messages, 79% was of the type CorefDir, whereas the other 21% was classified as CorefLong. Furthermore, only 10% referred to a received message and was classified as CorefBot. This means that a majority of 90% was classified as CorefSelf. In other words, the majority of co-references, referenced the last message that was sent by the user.

When looking at the co-referencing style, a majority of 76% of the queries referenced by CorefDup. Additionally, 15% of the co-referencing messages were classified as CorefReph, 6% as CorefImp and only 3% as CorefAna. Therefore, the majority of referencing happened by repeating the referenced phrase. Due to this repetition, the query can still be understood without placing it in context of the other queries. On the distribution of Coref messages, will be more elaborately commented in Section 5.4.

To sum up, queries were almost never repeated without alterations. The most popular reformulation was caused by introducing a new information need, while not changing the topic of the search. Most messages contained a reference to a previous message. However, the majority of references contained an exact repetition of the referenced entity. Additionally, the majority of references referred to the last message previously sent by the user. The main task of a conversational system, is "the ability of the search system to understand a user's information need over the course of the conversation, such that he or she does not need to repeat important aspects of the information need" (Radlinski and Craswell, 2017). These results indicate that users repeat themselves anyway, making the task of co-reference resolution a trivial one.

5.3 Comparing Fact Finding and Information Gathering Messages

In order to answer **RQ2**, the messaging data for fact finding and information gathering messages was compared. The exact quantities of fact finding and information gathering messages that were identified per taxonomy type are incorporated in Figure 2. Furthermore, this figure includes the Z-scores that were used to determine the significance of differences between fact finding and information gathering data. The categories that were significantly more common in information gathering messages, are coloured yellow in Figure 2. The categories that were more present in fact finding data are coloured blue. The categories that have no colour, are those where no significant difference was demonstrable between fact finding and information gathering data. The remainder of this subsection will lay out how the fact finding messages relate to the information gathering messages in the taxonomy.

All messages

To begin with, a significant difference was found in the use of keywords versus natural language messages. To elaborate, messages sent with a fact finding information need in mind showed to have a significantly higher share of NatLang messages, than messages with an information gathering nature. In other words, messages sent with an information gathering information need were significantly more likely to be classified as Keyword. The difference between the use of NonQ messages and Query messages proved to be insignificant.

Reformulating queries

When looking at the distributions of query reformulation styles, some interesting differences occur between the fact finding and the information gathering data. Namely, the fact finding data contained a significantly bigger ratio of the type Reph compared to information gathering data. Information gathering data, on the other hand, proved to be more likely to use messages of the types NewInfNd and TopS. In other words, information gathering data was more likely to re-focus questions within one topic, and also more likely to switch the topic of questions entirely. Fact finding data, on the other hand, was more likely to stick to the same exact information need by rephrasing the same queries.

Co-referencing queries

96% of the fact finding messages and 82% of the information gathering messages were classified as Coref. This difference proved to be statistically significant. In other words, a fact finding message was more likely to refer to a previous message, than an information gathering message. This result is in line with the aforementioned result, that that fact finding data is more inclined to stick to the same topic of queries. It is intuitive that fact finding data would also use more co-referencing, which is now confirmed.

Attention was paid to the distribution of the four co-referencing styles CorefDup, CorefReph, CorefAna and CorefImp. Statistical analysis proved that whereas co-referencing in a fact finding context was more likely to be classified as CorefReph, references made during information gathering sessions were more likely to be CorefImp. The difference in the use of CorefDup and CorefAna between fact finding and information gathering messages was proven to be statistically insignificant. This means that fact finding messages were more likely to rephrase the referenced entity, whereas information gathering messages were more likely to be implicit.

Next, was looked at the difference in timing of a referenced message. Statistical analysis again showed a significant difference between the fact finding and the information gathering data. That is, fact finding messages were more likely to reference its immediate previous, and therefore be of the type CorefDir. Information gathering messages, on the other hand, were more likely to be classified as CorefLong. Finally, the difference between the use of CorefSelf and CorefBot was found to be insignificant.

To sum up, fact finding messages were more likely to stick to one information need, and try multiple rephrases of this. This entails the use of more co-referencing in fact finding data. Additionally, a fact finding message was most likely to

refer to the directly preceding message, by rephrasing the referenced entity. Information gathering messages, on the other hand, were most likely to switch between information needs and search topics. When referencing, this was more often done implicitly, and not necessarily to the immediate previous message.

5.4 Further Analysis of Co-referencing

As previously mentioned, understanding the use of co-referencing during the conversational search dialogue is an active topic of research. Therefore, the use of co-referencing was further analysed, beyond the differences between fact finding and information gathering data. This subsection will highlight the significant relationships and differences that were found.

Firstly, `CorefDir` messages were significantly more likely to be of the type `CorefDup`, compared to `CorefLong` messages. `CorefLong` messages, on the other hand, were more likely to be of the type `CorefImp`. This is in line with the aforementioned result, that information gathering messages are both more likely to be of the type `CorefLong` and `CorefImp`. In other words, `CorefLong` and `CorefImp` often occur together in information gathering search sessions.

Secondly, differences in the co-referencing styles were analysed between `CorefSelf` and `CorefBot` messages. Notable here was the difference in use of anaphors and rephrasing co-references. That is, a `CorefBot` message was significantly more likely to also be classified as `CorefAna`, compared to `CorefSelf`. A `CorefSelf` message, on the other hand, was more likely to be of the type `CorefReph`.

Additionally, queries of the type `Reph` made significantly more use of both `CorefReph` and `CorefAna`, compared to queries of the type `NewInfNd`. Queries classified as `NewInfNd`, however, were more likely to be of the type `CorefDup`.

These results are interesting for the development of systems using co-reference resolution. However, they are preliminary and could use further investigation. It would be interesting to conduct more research on this using different conversational systems, and a linguistic perspective.

6 Responsible Research

6.1 Ethical Considerations

Ethical guidelines. Before the deployment of the user study, the characteristics of the study were verified against the ethics checklist for human research of the TU Delft, and the Netherlands Code of Conduct for Research Integrity (Algra et al., 2018).

Participant consent. At the onset of the user study, a consent form was posed to the participant. Here was stated that participation was completely voluntary, and the participant had the right to withdraw at any point during the study. Terminating participation prematurely did not have any consequence for the participant, such as a lesser Prolific rating. The collected data consisted of all messages from the participant and all responses by the bot, together with an anonymous ID to separate different users. Additionally, the course of events and the purpose of the study was explained. This consent form ensured that the participant was aware of, and consented with the details of the study.

Societal relevance and harmless nature. This research aids the understanding of conversational query reformulation. This aids the further development and optimization of conversational information seeking systems. The development of conversational information seeking systems is of great relevance in today's society. Gathering querying data from real participants is essential in the process of query understanding, and this data is unprocurable by other methods or means of study. The experiment is completely harmless in nature, since it requires participants to perform an information seeking task. This can only result in possible knowledge gain, and has no other (harmful) consequences.

6.2 Epistemic Considerations

Reproducibility. This paper attempts to maximally take into account reproducibility of the research. To this end, the participant selection procedure, user study setup, data collection and data annotation processes are outlined in much detail. Additionally, this paper explains how the IR system was assembled using Macaw, and all software that was used is open-source and free.

Transparency. Due to privacy considerations for the participants, the raw messaging data as obtained from the user study will not be made publicly available. Apart from this, the paper attempts to be as transparent as possible about how results were obtained. To this end, some representative messaging examples are given for the data annotation process. Furthermore, the exact results of the data annotation are depicted. Additionally, the paper comments on what type of statistical tests were used and what significance level was used. This is enough information for others to check and verify the significance of the results, and conclusions that were drawn.

7 Discussion

7.1 Limitations and Caveats

Firstly, some of the system characteristics may have had influence on the messaging data. To elaborate, the utilization of the Telegram interface might have influenced the popularity of using natural language messages opposed to keyword queries. Since Telegram is also a platform on which people can chat with each other, this may have given participants the impression that natural language use was appropriate while chatting with the bot.

Furthermore, Macaw is still limited in its capability of question answering. This means users sometimes received a response that did not make sense or did not answer their question. This may have influenced the messaging data in several ways. Firstly, it have affected how users alternated between information needs and search topics. Furthermore, the results show that users repeat important phrases over numerous messages. This, too, could be influenced by flaws

of the conversational system. To elaborate, when users notice they are misunderstood, they might start repeating themselves for the remainder of the search session. Future research could benefit from the further development of Macaw into a more sophisticated conversational search system.

Furthermore, the characteristics inherent to the definition of the taxonomy had some influence on the results. For example, the type `Dup` is defined to always make use of the co-referencing style `CorefDup`. These inherent relationships should be well-understood when interpreting the data.

7.2 In Light of Existing Work

As previously mentioned, currently no other research exists that analyses conversational querying data in a similar way. The taxonomy used in this research, has also not been previously used elsewhere. However, some parallels can be drawn between the findings of this research, and that of Liu and Gwizdka (2010).

One of the findings of Liu and Gwizdka (2010), was that fact finding messages used more synonymous reformulations than information gathering messages. A synonymous reformulation is similar to the `Reph` type from this study. This is in accordance with the result of this study, that fact finding messages were more likely to be classified as `Reph`.

Furthermore, Liu and Gwizdka (2010) found that specializations were the most popular query reformulation types. Whereas this type does not comment on the change in meaning of the reformulation, intuitively it would fall under the `NewInfNd` category of this research. The second most popular reformulation type found by Liu and Gwizdka (2010), was that of a word substitution. This type would translate to the type `Reph` in this research, which was also the second most popular. The least popular reformulation type found by Liu and Gwizdka (2010) was called repeat. This type is synonymous to the `Dup` type in this research, which was also least used. Therefore, the findings of Liu and Gwizdka (2010) are in accordance with the findings of this study.

8 Conclusions and Future Work

This work investigates query reformulation patterns of a conversational search dialogue, and how this differs for fact finding and information gathering search tasks. To this end, a taxonomy of query reformulation types was established. Messages collected from a task-based user study were classified, differentiating between fact finding and information gathering search tasks. The results suggests significant differences between query reformulation of fact finding and information gathering messages. These findings will aid the development of conversational information retrieval systems.

Future research might compare the results from this study to the results from a traditional context. Additionally, future work might use the established taxonomy to assess query reformulation types using different conversational search software. Furthermore, future research could elaborate on the preliminary findings on the use of co-referencing.

References

KA Algra, Lex M Bouter, AM Hol, and Jan van Kreveld. Nederlandse gedragscode wetenschappelijke integriteit, 2018.

Jeff Huang and Efthimis N Efthimiadis. Analyzing and evaluating query reformulation strategies in web search logs. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 77–86, 2009.

Bernard J Jansen, Amanda Spink, and Bhuva Narayan. Query modifications patterns during web searching. In *Fourth International Conference on Information Technology (ITNG'07)*, pages 439–444. IEEE, 2007a.

Bernard J Jansen, Mimi Zhang, and Amanda Spink. Patterns and transitions of query reformulation during web searching. *IJWIS*, 3(4):328–340, 2007b.

Melanie Kellar, Carolyn Watters, and Michael Shepherd. A field study characterizing web-based information-seeking tasks. *Journal of the American Society for Information Science and Technology*, 58(7):999–1018, 2007.

Chang Liu and Jacek Gwizdka. Analysis of query reformulation types on different search tasks. In *Proceedings of iConference 2010, Urbana-Champaign, IL.* iConference, 2010. URL http://hdl.handle.net/2142/15049.

Vincent Ng and Claire Cardie. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 104–111. Association for Computational Linguistics, 2002.

Chen Qu, Liu Yang, W Bruce Croft, Johanne R Trippas, Yongfeng Zhang, and Minghui Qiu. Analyzing and characterizing user intent in information-seeking conversations. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 989–992, 2018.

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.

Altaf Rahman and Vincent Ng. Supervised models for coreference resolution. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 968–977, 2009.

Ning Sa. Improving query reformulation in voice search system. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, pages 365–367, 2016.

Jaime Teevan, Eytan Adar, Rosie Jones, and Michael AS Potts. Information re-retrieval: repeat queries in yahoo's logs. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 151–158, 2007.

Johanne R Trippas, Damiano Spina, Mark Sanderson, and Lawrence Cavedon. Results presentation methods for a spoken conversational search system. In *Proceedings of*

the First International Workshop on Novel Web Search Interfaces and Systems, pages 13–15, 2015.

Johanne R Trippas, Damiano Spina, Lawrence Cavedon, and Mark Sanderson. How do people interact in conversational speech-only search tasks: A preliminary analysis. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, pages 325–328, 2017.

Johanne R Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. Informing the design of spoken conversational search: Perspective paper. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, pages 32–41, 2018.

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