# Toward Conversational Query Reformulation

Johannes Kiesel[1], Xiaoni Cai[1], Roxanne El Baff[2], Benno Stein[1] and Matthias Hagen[3]

[1]Bauhaus-Universität Weimar, Bauhausstraße 11, 99423 Weimar, Germany
[2]German Aerospace Center (DLR), Germany
[3]Martin-Luther-Universität Halle-Wittenberg, Von-Seckendorff-Platz 1, 06120 Halle (Saale), Germany

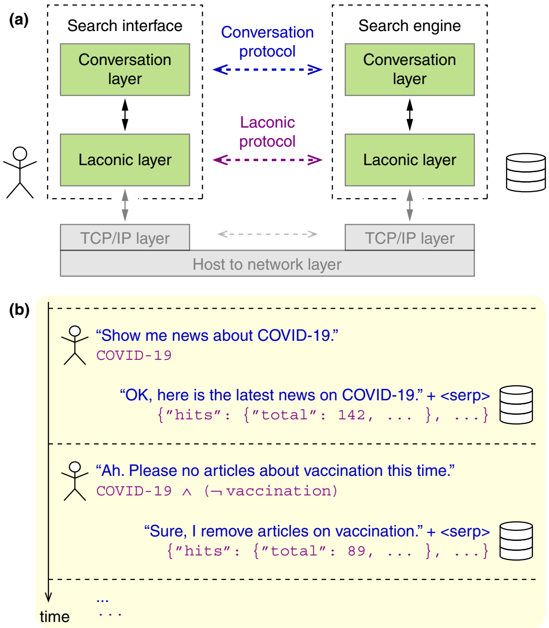
Abstract
In traditional web search interfaces, information seekers reformulate their queries by editing the terms in the search box in order to guide the retrieval process. Such kind of editing is at odds with the natural language interaction paradigm in conversational interfaces, and for purely voice-based interfaces it is impossible. Conversational search studies reveal that participants instead *describe* their changes to a query; however, the principles of such "editing conversations" have not been analyzed in depth. The paper in hand formalizes the problem of conversational query reformulation. We cast reformulations as meta-queries that imply operations on the original query and categorize the operations following the standard CRUD terminology (create, read, update, delete). Based on this formalization we crowdsource a dataset with 2694 human reformulations across four search domains. Our analysis of the meta-queries reveals a large variety in word usage and indicates ambiguous reformulations as an important research topic of its own.

**Keywords**
Conversational search, Information seeking, Query reformulation, Query refinement, CRUD, Crowdsourcing

## 1. Introduction

During web search, information seekers frequently find a search engine's results either too specific, too generic, or containing results relevant only to an unintended interpretation of their query. In such cases seekers may want to reformulate their queries [1, 2]. In a traditional search interface, the seeker would directly edit the previous query in the search field, creating, updating, or deleting terms. Such reformulations account for about half of all queries [3, Sec. 6.3]. However, conversational search interfaces—be they chat-like or voice-based—usually do not allow modifying the previous query. Though some chat interfaces not used for search allow to edit previous messages, such a functionality breaks temporal continuity, making the interaction significantly less conversational. Still, reformulations are also frequent in conversational search lab studies [4] and can be seen as one user-facing service of the search interface's conversational layer, as illustrated in Figure 1 (a).

As the example in Figure 1 (b) illustrates, reformulations allow to specify information in small steps. As the main advantage of incremental formulation, seekers do not have to formulate the complete query in advance,



"Show me news about COVID-19."
`COVID-19`

"OK, here is the latest news on COVID-19." + <serp>
`{"hits": {"total": 142, ... }, ...}`

"Ah. Please no articles about vaccination this time."
`COVID-19 ∧ (¬ vaccination)`

"Sure, I remove articles on vaccination." + <serp>
`{"hits": {"total": 89, ... }, ...}`

...
. . .

**Figure 1:** (a) Web application protocol stack, emphasizing the two top layers. A conversational search interface implements a "conversation layer" (an API) to operationalize the respective NLP translation functionality on top of a "laconic" layer (an API) that implements the functionality to interact with a traditional web search interface. (b) Example conversation about COVID-19, showing the messages of a fictitious dialog at the level of a conversational and a laconic protocol. Recall that when going downwards (upwards) the stack, the messages of the protocol at layer $n-1$ (layer $n$) are generated from those at layer $n$ (layer $n-1$).

*DESIRES 2021 – 2nd International Conference on Design of Experimental Search & Information REtrieval Systems, September 15–18, 2021, Padua, Italy*

johannes.kiesel@uni-weimar.de (J. Kiesel); xiaonicaiweimar@gmail.com (X. Cai); roxanne.elbaff@dlr.de (R. El Baff); benno.stein@uni-weimar.de (B. Stein); matthias.hagen@informatik.uni-halle.de (M. Hagen)
0000-0002-1617-6508 (J. Kiesel); 0000-0003-2114-8162 (X. Cai); 0000-0001-9033-2217 (B. Stein); 0000-0002-9733-2890 (M. Hagen)

CEUR Workshop Proceedings (CEUR-WS.org)

substantially reducing the required mental effort.[1] Moreover, incremental formulation might simplify and thus increase the use of search operators for seekers, like the exclusion in Figure 1 (b), allowing seekers to formulate even complex needs more intuitively.[2]

Though the seeker may not consciously ask to create, update, or delete query terms, conversational reformulations are essentially such meta-queries that request changes to the previous query. While these operations are not implemented in standard retrieval engines, one can imagine a "conversation" layer on top of such engines (Figure 1 (a)) that—among others—resolves reformulations similar to co-reference resolution in conversational question answering systems (e.g., [6]).

However, critical issues for conversational systems concerning reformulations have barely been analyzed in the literature, especially the reformulations' inherent ambiguity. For example, consider "OK, how about vaccinations?" in place of the seeker's second message in Figure 1 (b). Is the intent to create a "vaccination" term or to replace the query entirely? If aware of the ambiguities, a system could ask for clarification or use heuristics to resolve the ambiguity. It could also stress and thereby teach unambiguous language in its replies ("I reduced the list to those on vaccinations.").

To foster research on conversational query reformulations, we contribute the following: (1) a conceptualization that casts conversational reformulations as meta-queries following CRUD terminology (cf. Section 3); (2) the first dataset on conversational query reformulations,[3] containing 2694 messages and associated meta-queries, crowdsourced from 284 study participants from 5 countries in 4 different search domains (cf. Section 4); and (3) an in-depth analysis of the reformulations' word patterns, emphasizing ambiguous word patterns as important research direction and suggesting the general feasibility of a domain-independent rewriting system (cf. Section 5).

# 2. Related Work

Though conversational search is an active research area, conversational query reformulation has attracted little attention so far. In contrast, several recent publications target co-reference resolution for follow-up questions in conversational question answering [7]. Reformulations and follow-up questions are similar in that both ask for information connected to information just retrieved. However, whereas query reformulations change the criteria that identify relevant information, follow-up questions request a completely new answer. This difference in intent causes linguistic differences between query reformulations and follow-up questions that warrant separate investigations. Moreover, conversational query reformulation relates to much of the available research on queries.

## 2.1. Queries in Conversational Search

A central promise of the conversational search paradigm is to bring search closer to real-world assistance of a reference librarian [8]. In this regard, conversational reformulations are one piece for allowing the seeker to specify their need on a more natural level [9]. Still, unlike "query by babbling" [10], reformulations require some formalized description in the seeker's mind. Instead, conversational query reformulations are one instance of "user revealment" [11] where the seeker incrementally specifies their need. The advantages of small steps, as in orienteering [12], are that seekers have to specify less and obtain context information on the way.

In their history of IR research, Sanderson and Croft [13] divided interactions against some first text-based conversational search systems into natural language or keyword-based and into non-querying and querying. In a later fine-grained study of conversational seeker messages for passage retrieval, Lin et al. [14] categorize query ambiguity, though they focus on ambiguity regarding (not) referenced entities, whereas the paper at hand focuses on ambiguity regarding the desired operation. Trippas et al. [4] present a model for spoken conversational search that also covers information requests beyond queries, for example, within a result document. In their lab study, they observe that both seekers conversationally reformulate queries ("query embellishments") and that participants in the system's role conversationally offer reformulations based on what they see.

## 2.2. Query Reformulation

Query reformulations are queries based on the previous one with a similar information need [1]. Boldi et al. [1] classified query reformulations on two axes from generalization to specification and from being the same query to mission change. Sanderson and Croft [13] proposed a categorization based on the latter axis. Jiang et al. [15] analyzed voice query repeats after voice input errors. They found that seekers tended to stress words that the system misunderstood.

Though studies took note of conversational query reformulations (e.g., [16]), they have rarely been analyzed. As an exception, Sa and Yuan [17] asked 32 participants in a Wizard of Oz study to perform one generalization and one specialization of a displayed query, speaking

[1] This effect likely also holds for more instruction-like (but still conversational) interactions, like in Adobe's "phonic filters" image search demo, https://blog.adobe.com/en/2019/05/29/preview-technology-gives-your-voice-the-power-of-a-creative-director.html.

[2] Some years ago, only about 1% of web search queries contained operators [5].

[3] Publicly available at https://doi.org/10.5281/zenodo.5031960

to the system like to a human. The participants preferred conversational query reformulations (which Sa an Yuan call “partial query modification”) over repeating the query with changes (“complete query modification”), even though several participants reported employing the latter for being used to it. We build upon this work, asking for more complex reformulations in longer query sessions and focusing on analyzing the language employed and ambiguities therein.

### 2.3. Query Rewriting

In contrast to query reformulation, query rewriting refers to processing the query before the retrieval. This task currently attracts much attention in conversational question answering, mostly concerning co-reference resolution. Available datasets for this task include CSQA [18], CoQA [19], QReCC [6], QuAC [20], and TREC CAsT [21]. The availability of datasets has already led to several approaches, often employing sequence-to-sequence learning [22, 23, 19] and previous interactions [24, 25, 26].

### 2.4. Natural Language Queries

Several studies analyzed natural language queries even before conversational interfaces. Belkin et al. [27] found that, in a text-based search interface, the average query length increased by nearly 50% when the search box label encouraged to write a problem description. Still, the automatic “translation” of long queries to shorter keyword queries later also gained attention with systems reducing natural language queries to some key concepts more compatible with keyword-based interfaces [28, 29]. Moreover, also the translation of natural language to database or knowledge graph queries attracts much attention (e.g., [30, 31, 32]). As example of spoken reformulations in a different setting, researchers have for decades investigated ways to edit text by voice [33, 34], using commands like “Capitalize the first letter in each word in each title” [35].

## 3. Conceptualizing Conversational Reformulations

Conversational reformulations are query reformulations using natural language. While query reformulations in web search usually are stand-alone queries that can directly be submitted to retrieve results, formulating queries from conversational reformulations requires an additional step that “adds” the conversational context. This section discusses the implications on three levels: (1) a model of conversational reformulations as meta-queries; (2) the problem of algorithmically understanding reformulations; and (3) the process of creating the “laconic” queries from the reformulations.

**Table 1**
Conversational examples for each basic operation in the standard CRUD terminology [36].

| Operation | Target | | |
|---|---|---|---|
| | Query | Expression | Literal |
| **C** reate | *Show me news about COVID-19* | *Remove all without NCD, NI, or WHO in the headline* | *Any news for its treatment?* |
| **R** ead | *What do I have so far?* | *What did I say for the headline?* | *What was the last filter?* |
| **U** pdate | *Start a new search on the flu* | *Remove my criteria for the headline but search only in economical news* | *No, NI means National Insurance* |
| **D** elete | *No, let’s start again* | *Remove the headline criteria* | *Remove the filter for treatments* |

### 3.1. Casting Reformulations as Meta-Queries

From the information system’s perspective, the information seeker uses a meta-query language when expressing conversational reformulations: they tell the system to perform specific operations on the previous query. On a syntactic level, the basic reformulation operations in traditional search interfaces are adding, changing, and removing a term. These correspond to the basic operations *create*, *update*, and *delete* of data systems [36]. The fourth basic operation of data systems, *read*, may also be useful if the previous query is not visible, like in some conversational interfaces. For illustration, Table 1 shows conversational examples for each basic operation. One query reformulation can contain several basic operations.

### 3.2. Algorithmically Understanding Natural Language Reformulations

In the past years, impressive advancements have been achieved in natural language understanding. Still, when a message can be interpreted as different meta-queries, the problem is far from solved. For example, what if the seeker would have asked “OK, how about vaccinations?” as their second message in Figure 1 (b)? Is the intent to specify the previous query or to start a new one? Hints on the true intent might be found in previous messages (relation between ‘vaccination’ and the previous query) or previous results (maybe the seeker read something that caused the question). Other conversations may suggest quite different interpretations. For example, if asked after "Can you show me articles about its treatments?", one could interpret 'vaccination' as a replacement for 'treatments.' To resolve such ambiguities, search systems may ask the seeker for clarification [37, 38] or they may try heuristic disambiguation. Such heuristics could, for example, employ word or entity relationships (e.g., using WordNet or knowledge graphs) or query performance predictors like term specificity and result coherence [39].

### 3.3. Formulating Laconic Queries for Retrieval

The example in Figure 1 (b) shows that conversational reformulations (like the seeker's second message) have to be converted to context-independent queries to submit them to standard retrieval systems. This is similar to "query rewriting," a process that resolves co-references in conversational question answering (e.g., [6]). In fact, similar methods may be effective for conversational reformulations. In the protocol stack of Figure 1 (a), conversational reformulations can thus be seen as a service of the conversation layer that builds upon the retrieval service of the laconic layer.

## 4. Crowdsourcing Reformulations

To foster research on conversational query reformulations, we publish a respective crowdsourced dataset[4] that accounts for diversity in seeker location (five English-speaking countries) and search domain (four different ones). The goal is to analyze the diversity and ambiguity in the language of conversational reformulations. However, the dataset also allows to bootstrap the natural language understanding component of conversational systems [40].

Figure 2 shows the interface used in Amazon's Mechanical Turk marketplace to collect "natural" reformulations. We iteratively refined the interface in eight pilot studies with 80 participants to minimize the interface's influence on the participant's choice of words. Based on insights from the pilot studies, we formulated the tasks as bullet points with a sentence structure clearly different from the reformulations we asked for. Moreover, automatic checking routines help the participants to stick to the task (e.g., alerts for undesired repetitions or missing terms). The interface resembles a WhatsApp chat to prime participants on chat messages [41].

After an initial "ready"-interaction (cf. top of Figure 2), each participant completed twelve assignments from one domain as a single search session. To analyze reformulation diversity, we changed the task domain and topic between participants: either finding *arguments* on banning plastic bags, finding *books* on (Sci-Fi) viruses, finding *news* on COVID-19, or finding *trips* to San Jose. However, the search tasks for each participant had the same structure of abstract operations (e.g., create one term) with only keywords being replaced between the domains.[5] To ensure a variety of reformulations, we formulated the instructions to cover all four CRUD operations, to vary the targets from a single literal to the whole query, to cover conjunctions and disjunctions, and to include some special cases like a filter attribute, an unspecified literal, or a negation. Participants completed a session in about 12 minutes (observed in the pilot studies and the final study) and we adjusted the payment to cover the minimum wage of the respective country.[6] Unfortunately, we had to stop our study in India and Australia after the first domain (news). Only 22% of the Indian participants provided reasonable messages for the tasks (61% in other countries) while getting answers from 20 Australian participants alone exhausted our time constraints. In total, we accepted the work of 284 participants.

**Table 2**
Key statistics of the dataset by the participants location (country). Messages have been manually categorized as being either a command (Co.), question (Qu.), or statement (St.).

| **Participant location** | **Participants by domain** | | | | | **Messages** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\sum$ | **Arg.** | **Book** | **News** | **Trip** | $\sum$ | **Co.** | **Qu.** | **St.** |
| Australia | 20 | 0 | 0 | 20 | 0 | 192 | 0.76 | 0.08 | 0.16 |
| Canada | 83 | 20 | 20 | 23 | 20 | 781 | 0.58 | 0.19 | 0.22 |
| Great Britain | 80 | 20 | 20 | 20 | 20 | 774 | 0.60 | 0.25 | 0.15 |
| India | 21 | 0 | 0 | 21 | 0 | 181 | 0.78 | 0.10 | 0.12 |
| United States | 80 | 20 | 20 | 20 | 20 | 766 | 0.54 | 0.20 | 0.26 |
| Total | 284 | 60 | 60 | 104 | 60 | 2694 | 0.60 | 0.20 | 0.20 |

To ensure the dataset's quality and ease processing, we manually checked each message. Of the initially 3408 messages, 2917 are grammatically and semantically meaningful in the respective context. Of these, 2694 (79% of all) can be interpreted as the respective intended meta-query and form the final dataset of 558 messages for the argument domain, 573 messages for book, 961 messages for news, and 602 for trip (cf. Table 2 for other key statistics).

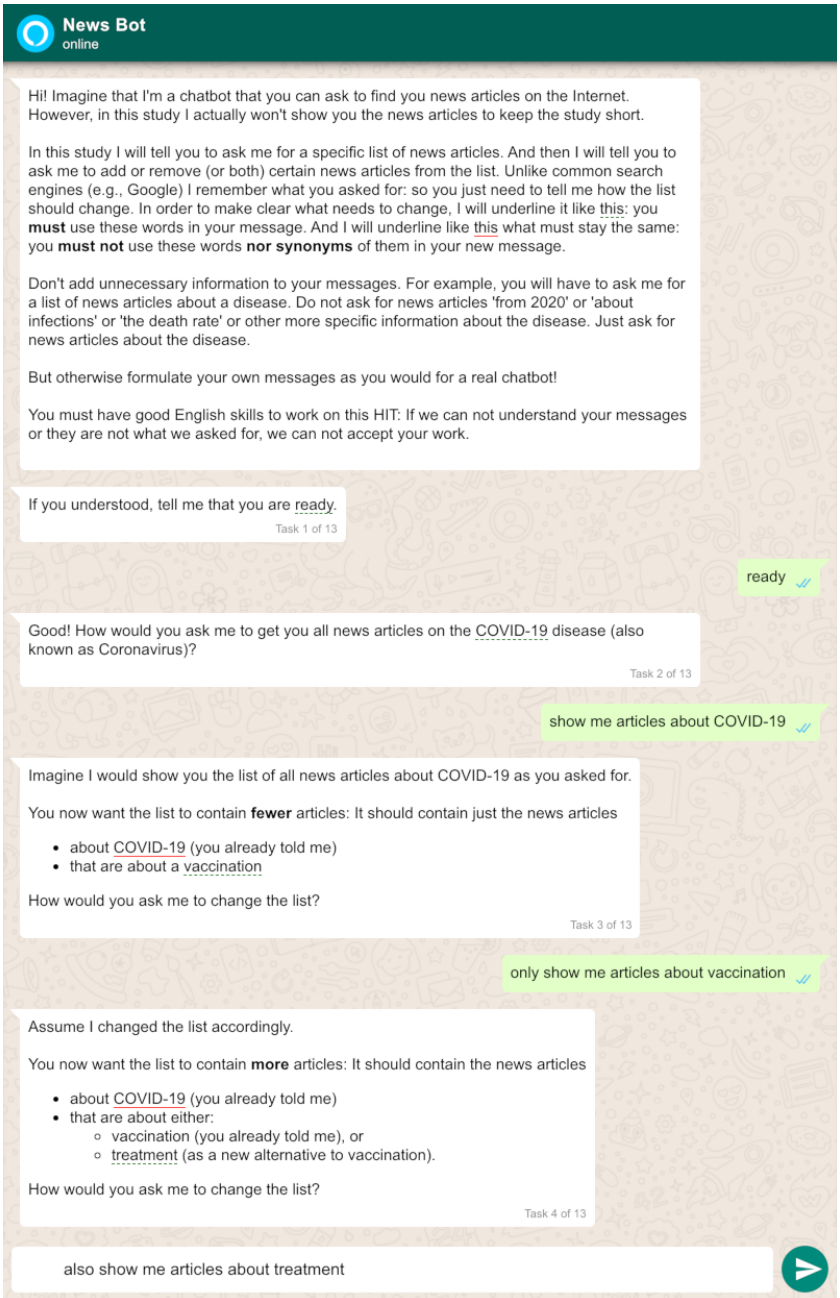
## 5. Analyzing Reformulations

Like for all natural language systems, also developing systems that allow for conversational query reformu-

[4] Publicly available at https://doi.org/10.5281/zenodo.5031960.

[5] The keywords are contained in the README file of the dataset, whereas the annotation interface for each domains in the respective `<domain>-interface.html` file

[6] https://medium.com/ai2-blog/crowdsourcing-pricing-ethics-and-best-practices-8487fd5c9872



Hi! Imagine that I'm a chatbot that you can ask to find you news articles on the Internet. However, in this study I actually won't show you the news articles to keep the study short.

In this study I will tell you to ask me for a specific list of news articles. And then I will tell you to ask me to add or remove (or both) certain news articles from the list. Unlike common search engines (e.g., Google) I remember what you asked for: so you just need to tell me how the list should change. In order to make clear what needs to change, I will underline it like this: you **must** use these words in your message. And I will underline like this what must stay the same: you **must not** use these words **nor synonyms** of them in your new message.

Don't add unnecessary information to your messages. For example, you will have to ask me for a list of news articles about a disease. Do not ask for news articles 'from 2020' or 'about infections' or 'the death rate' or other more specific information about the disease. Just ask for news articles about the disease.

But otherwise formulate your own messages as you would for a real chatbot!

You must have good English skills to work on this HIT: If we can not understand your messages or they are not what we asked for, we can not accept your work.

If you understood, tell me that you are ready.

Task 1 of 13

ready

Good! How would you ask me to get you all news articles on the COVID-19 disease (also known as Coronavirus)?

Task 2 of 13

show me articles about COVID-19

Imagine I would show you the list of all news articles about COVID-19 as you asked for.

You now want the list to contain **fewer** articles: It should contain just the news articles

- about COVID-19 (you already told me)
- that are about a vaccination

How would you ask me to change the list?

Task 3 of 13

only show me articles about vaccination

Assume I changed the list accordingly.

You now want the list to contain **more** articles: It should contain the news articles

- about COVID-19 (you already told me)
- that are about either:
  - vaccination (you already told me), or
  - treatment (as a new alternative to vaccination).

How would you ask me to change the list?

Task 4 of 13

also show me articles about treatment

**Figure 2:** Dataset collection interface (excerpt). After submitting a message (bottom box), the interface alerted participants of potential missing or forbidden terms. Valid messages appear on the right and a next assignment on the left.

lations demands investigating the peculiarities of the respective language. To this end, Section 5.1 provides a general overview of the messages collected in our dataset, highlighting differences in language use between search domains and countries. Though the exact patterns occur with different frequencies, they, in general, are similar across both search domains and countries, indicating that a generic reformulation resolution system might be feasible. In Section 5.2, we report on our detailed analysis of the ambiguities in the messages, showcasing both the

**Table 3**

The abstracted sequence of operations (Create, Read, Update, Delete) that each participant performed in one of four domains, together with a statistical overview of messages in the dataset, including their absolute number and relative frequency of each type (command (Co.), question (Qu.), or statement (St.)) as well as unambiguous and ambiguous ones. For clarity, queries $q_i$ are provided here in Boolean form. An item is relevant for a query $q_i$ if $q_i$'s expression evaluates to true for the item, where $l_i$ denotes a literal that evaluates to true if the corresponding term occurs in the item, $?_i$ a literal with no corresponding term, and $f_i(e)$ an expression that evaluates to true if $e$ evaluates to true for attribute $f_i$ of the item. See the dataset's README file for the corresponding terms and attributes for each domain and the interface HTML files for the respective task descriptions.

| # | Operation (CRUD) | Result query | Messages | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | $\sum$ | Co. | Qu. | St. | Unam. | Alt. interpretations [messages] |
| 1 | C: $l_1$ | $q_1 = l_1$ | 275 | 0.61 | 0.32 | 0.07 | 1.00 | - |
| 2 | C: $l_2$ | $q_2 = l_1 \wedge l_2$ | 226 | 0.68 | 0.23 | 0.09 | 0.36 | U: $q_1 \to l_1 \vee l_2$ [0.35]<br>U: $q_1 \to l_2$ [0.64] |
| 3 | U: $l_2 \to l_2 \vee l_3$ | $q_3 = l_1 \wedge (l_2 \vee l_3)$ | 212 | 0.69 | 0.24 | 0.07 | 0.27 | U: $q_2 \to l_3$ [0.16]<br>U: $q_2 \to (l_1 \wedge l_2) \vee l_3$ [0.72] |
| 4 | D: $l_2 \vee l_3$ | $q_4 = l_1$ | 70 | 0.60 | 0.34 | 0.06 | 1.00 | - |
| 5 | C: $f_1 : (l_4 \vee l_5 \vee l_6)$ | $q_5 = l_1 \wedge f_1 : (l_4 \vee l_5 \vee l_6)$ | 262 | 0.71 | 0.18 | 0.10 | 0.48 | U: $q_4 \to f_1 : (l_4 \vee l_5 \vee l_6)$ [0.52] |
| 6 | U: $l_5 \to ?_1$ | $q_6 = l_1 \wedge f_1 : (l_4 \vee ?_1 \vee l_6)$ | 258 | 0.42 | 0.03 | 0.54 | 0.61 | U: $l_5 \to \neg l_5$ [0.39] |
| 7 | U: $?_1 \to l_7$ | $q_7 = l_1 \wedge f_1 : (l_4 \vee l_7 \vee l_6)$ | 276 | 0.32 | 0.01 | 0.66 | 1.00 | - |
| 8 | U: $l_1 \to l_8$ | $q_8 = l_8 \wedge f_1 : (l_4 \vee l_7 \vee l_6)$ | 259 | 0.67 | 0.17 | 0.15 | 1.00 | - |
| 9 | R: $q_8$ | $q_9 = l_8 \wedge f_1 : (l_4 \vee l_7 \vee l_6)$ | 189 | 0.38 | 0.59 | 0.03 | 1.00 | - |
| 10 | U: $f_1 : (l_4 \vee l_7 \vee l_6) \to l_9$ | $q_{10} = l_8 \wedge l_9$ | 134 | 0.66 | 0.23 | 0.11 | 1.00 | - |
| 11 | U: $q_{10} \to l_{10}$ | $q_{11} = l_{10}$ | 269 | 0.72 | 0.14 | 0.14 | 0.52 | C: $l_{10}$ [0.48] |
| 12 | C: $\neg l_1$ | $q_{12} = l_{10} \wedge \neg l_1$ | 264 | 0.76 | 0.11 | 0.13 | 1.00 | - |

ambiguities and possible ways to avoid them.

To investigate on the word patterns of conversational query reformulations and differences between search domains and countries, our study design employs the same sequence of twelve abstract tasks for each participant, only exchanging a few keywords to specify the different search domains. Table 3 shows the formal tasks and provides general characteristics of the collected messages. The $\sum$-column shows the total number of valid messages per task. Though this number is close to the maximum of 284 (the number of participants) for most tasks, it is relatively low for Tasks 4, 9, and 10, indicating a misunderstanding of the participants. We see such misunderstandings as an artifact of our study setup and filter out the affected messages from our below analyses.[7]

## 5.1. Comparing Messages across Tasks, Domains, and Countries

We have systematically analyzed the 2694 messages of our dataset. Besides the more general analyses of message types, we also focus on word frequencies and patterns. Interestingly, apart from a small difference in preposition frequencies for trips ('to' compared to 'about' in the other domains), argument search has one difference to the other domains in that a few participants formulated their requests not as a query but asking for the system's opinion (e.g., "What do you think about banning plastic bags?").

**Message types** For a first general overview, we manually annotated each message as being a command, a question, or a statement. The left part of Table 2 shows the message type usage per country. While participants from Australia and India used more commands, participants from Great Britain used more questions, and participants from the United States used more statements on average. Overall, we found this observation to show the main difference between countries, with no notable difference in word choice for the relatively small amount of data we gathered.

Table 3 shows the message type usage per task. The most frequent are commands (e.g., "Please remove arguments about banning plastic bags." for Task 12) that often make up at least 60% of all messages per task. The one task where the majority of messages are questions is Task 9, which is the only task that involves a read operation (e.g., "Can you please remind me of my previous commands?" and "What did I search for?"). Though statements (e.g., "I would like to see news that are not about COVID-19." for Task 12) are relatively rare in general, they are the dominant type for the two tasks that deal with correcting a misunderstanding: Task 6 is to tell the system that it misinterpreted an acronym (but leaving the acronym unspecified, e.g., "I did not mean NI as North Ireland."), whereas Task 7 is to specify the intended meaning (e.g., "I meant National Insurance.").

[7] For completeness, we provide these messages in a separate file along with the dataset.

Participants switching to statement messages might thus be an indicator that they are pointing out problems.

**References to current list** As a potential signal to identify reformulations, participants sometimes, but unfortunately rarely, refer to the items in the (imagined) result list when reformulating the query (e.g., "Which of these include vaccination?"). Specifically, in the rare cases for the respective tasks (all but Tasks 1, 6, and 7), participants use *those* (3%), *ones* (3%), *these* (1%), or *them* (0.4%). Somewhat frequently, 16% refer to the *list* and 2% to *results* (e.g., "only show me results that include vaccination"). As a difference between domains, 1% of messages in the respective tasks of the trip domain use *there* (e.g., "I want to have a travel to there by ship."). More common than references to the current list is the use of the domain-specific item (*argument*, *book*, *article*, *trip*), with the special case of the phrases 'pros and cons' and 'for and against' to specify that both sides should be considered in argument search (e.g., "Can you give me arguments for and against banning plastic bags please"). However, these do not indicate reformulations.

**Growing and shrinking** As a somewhat strong signal for reformulations, many participants explicitly expressed whether the result list should grow or shrink. Verbal expressions for shrinking the current list (Tasks 2, 5, 12) are *remove* (9% of messages in these tasks), *filter* (5%, e.g., "Filter list to just about vaccination."), *exclude* (4%), *narrow down* (3%), *reduce* (1%), *limit* (0.7%), *filter out* (0.5%, e.g., "Please filter out all articles that are not about vaccination.") and *shorten* (0.5%). Note that some of these verbs indicate which items to remove, whereas others indicate which items to keep. Overall, 26% of the messages in these tasks contain a verb that explicitly requests to shrink the list. Other signals for shrinking are the use of *only* (18%) and *just* (6%), though these percentages may be inflated as the descriptions of these tasks also contain *only* and *just*.

Frequently used verbal expressions for growing the current list (Tasks 3, 4) are *add* (25%), *add back* (2%, e.g., "add back the trips by car."), *expand* (3%, e.g., "Expand the list to include books about plants too."). Other signals for growing are the use of *also* (23%), *as well* (4%), *too* (3%, e.g., "can you also add those including treatment too?"), and *as well as* (1%). Interestingly, participants used the verb *keep* both to shrink the list in Task 2 (2% of messages for Task 2, e.g., "Keep articles related to vaccination") and in a lexically indistinguishable way to partially undo such shrinking in Task 3 (1%, e.g., "Please keep the arguments about renewable resources").

**Summary** The observed differences between countries and domains are relatively small. A change of the seeker from asking questions to expressing statements often indicates specific unusual requests, though differences between countries need to be considered. Finally, many participants directly requested the growing and shrinking of the result list in their reformulations.

## 5.2. Analyzing Operation Ambiguities

A common problem for natural language interfaces is the ambiguity of natural language. For reformulations, this means that the same message can be interpreted as different operations. In our study, we found the below three main ambiguities.

**Specializing a query or starting a new one** When asked to specialize the query by adding a new term (Task 2), the majority of our participants (66%, cf. Table 3) used a message that one could also interpret as starting a new query with that one term (e.g., "Just show me arguments about CO2 emissions"). We observe the same ambiguity in other specialization tasks (Task 3, 16% of messages, e.g., "Can you show me arguments that are about renewable resources?"; Task 5, 52%, e.g., "May I please see the articles that have NCD, NI, or WHO in the headline?") and in tasks that ask to start a new query (Task 11, 48%, e.g., "Find a list of books about evolution.").[8] Still, some participants directly used unambiguous messages, either by explicitly referencing the current list (e.g., "Great can you refine that to articles with NCD, NI or WHO in the headline?") or indicating a new list or search (45% of the messages for Task 11, e.g., "Find a list of books about evolution," "New search on evolution," or "Disregard all previous instructions and now only find me books about evolution"). Moreover, 8% of the messages for Task 11 are unambiguous due to explicitly expressing a replacement, for example, "Show me trips to Santiago instead."

**Unclear precedence** Though there are precise rules for operator precedence in logics, no such rules exist for natural language. Indeed, 72% of the messages for Task 3 do not clearly express whether the new term should be an alternative just to the last term (as asked for) or to the entire query (e.g., "Could you please also include arguments about renewable resources?"). About 11% of the participants' messages are unambiguous by explicitly stating the relation (e.g., "Show me trips by ship or by car," where 'ship' is the previously added query term). Though not asked to do so, a few of these participants also hinted at a reason for asking for the alternative (e.g., "Show me trips by ship, if ship trips are not available then I would

[8] Taken literally, also several messages for Task 1 and 12 would be ambiguous. However, the alternative interpretations make no sense in the respective contexts. We ignore these strictly lexical ambiguities in our considerations.

like to select trips by car."). A few participants (1%) made use of an explicitly stated *filter* term from the previous query, which then allowed them to refer back to it (e.g., "Filter list for infected animals" and then "Add plants to filter as alternative").

**Ambiguous negation** Surprisingly, several messages submitted to tell the system that it misinterpreted an acronym are lexically indistinguishable from filtering by the acronym. While the majority of messages for Task 6 are unambiguous as expected (61%, e.g., "I'm not asking for North Ireland"), 39% of the messages are ambiguous and could easily be misunderstood (e.g., "Do not include articles about North Ireland"). Indeed, only the fact that the user had just added 'NI' as an acronym hints at the intended meaning.

## 6. Conclusion

We have formalized the problem of supporting reformulations in conversational search systems. By casting reformulations as meta-queries that imply standard CRUD operations on the "actual" query, we demonstrate that such functionality could be implemented in a conversation layer on top of standard retrieval architectures. An analysis of a new dataset of 2694 crowdsourced human reformulations across four search domains shows that a generic reformulation component is feasible when considering the peculiarities of the respective search domains. However, we also find that ambiguities in the reformulations will likely be a major challenge for conversational systems that merit further investigations.

### Future Work

We see several opportunities to extend the analysis of this paper.

**Other languages.** We considered only English messages so far but expect at least some of the results to be language-dependent. Further analyses will need to be conducted for other languages.

**Generalized read operation.** We considered only the most simple read-operation: reading the current query. However, seekers may also want to fetch a query they used some time ago, maybe to continue or refresh a previous search.

**More search operators.** We considered only the standard logical operators (∨, ∧, ¬) and attribute-specific filters, while most retrieval systems support several more. How would seekers highlight phrases (words to be retrieved in that sequence), initiate boosting (a term or attribute being especially important), or fuzzy / strict term matching? As hypothesized in Section 1, step-wise query formulation might increase the use of diverse search operators.

**Clarifications.** We considered only messages from the seeker, but studying possible system reactions is equally essential to account for implicit feedback (repeating what was understood) or asking clarification questions. Both methods are likely helpful to explain and resolve ambiguities, and could at the same time allow the system to showcase unambiguous formulations in an attempt to teach the seeker how to prevent the ambiguities in the future.[9]

### Implications for Conversational Search

We can only hypothesize how a seeker's interactions differed if a search system supported conversational query reformulations. As mentioned above and in Section 1, one possibility is that the reduced cognitive effort due to step-wise and natural language query formulation encourages more complex queries that contain more search operators. Extending on these considerations, we expect that some seekers will desire to regularly use the same query like a feed (e.g., for news, but also for professional activities like scholarly search [43]) and may see query formulation as an act of personalization. Therefore, some systems may even aim to support reformulations like "A bit more on soccer." At the same time, we believe that supporting reformulations will be essential for having a conversation between seeker and search system, as reformulations implement a straightforward way for the seeker to ground the conversation [44], complementing clarification questions from the system. At such a stage, the conversations will be more natural. And the users will be "relieved" from the below laconic layer—just like today's users of the laconic layer do not need to know any details about the underlying TCP/IP layer.

## Acknowledgments

This work has been partially funded by the Google Digital News Initiative as part of the "Conversational News" project.

## References

[1] P. Boldi, F. Bonchi, C. Castillo, S. Vigna, Query reformulation mining: models, patterns, and applications, Information Retrieval 14 (2011) 257–289. doi:10.1007/s10791-010-9155-3.

[9] Methods to explain ambiguities in reformulations may build upon early work on explaining ambiguities in formal query languages [42].

[2] J. Chen, J. Mao, Y. Liu, F. Zhang, M. Zhang, S. Ma, Towards a better understanding of query reformulation behavior in web search, in: J. Leskovec, M. Grobelnik, M. Najork, J. Tang, L. Zia (Eds.), Proc. of WWW, ACM / IW3C2, 2021, pp. 743–755. doi:10.1145/3442381.3450127.

[3] W. B. Croft, D. Metzler, T. Strohman, Search Engines - Information Retrieval in Practice, Pearson Education, 2009.

[4] J. R. Trippas, D. Spina, P. Thomas, M. Sanderson, H. Joho, L. Cavedon, Towards a model for spoken conversational search, Information Processing Management 57 (2020) 102162. doi:10.1016/j.ipm.2019.102162.

[5] R. W. White, D. Morris, Investigating the querying and browsing behavior of advanced search engine users, in: W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, N. Kando (Eds.), Proc. of SIGIR, ACM, 2007, pp. 255–262. URL: https://doi.org/10.1145/1277741.1277787. doi:10.1145/1277741.1277787.

[6] R. Anantha, S. Vakulenko, Z. Tu, S. Longpre, S. Pulman, S. Chappidi, Open-domain question answering goes conversational via question rewriting, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proc. of NAACL-HLT, ACL, 2021, pp. 520–534. URL: https://www.aclweb.org/anthology/2021.naacl-main.44/.

[7] M. Zaib, W. E. Zhang, Q. Z. Sheng, A. Mahmood, Y. Zhang, Conversational question answering: A survey, CoRR abs/2106.00874 (2021). URL: https://arxiv.org/abs/2106.00874.

[8] J. Culpepper, F. Diaz, M. Smucker, Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in lorne (SWIRL), SIGIR Forum 52 (2018) 34–90.

[9] R. S. Taylor, The process of asking questions, American Documentation 13 (1962) 391–396. doi:10.1002/asi.5090130405.

[10] D. W. Oard, Query by babbling: A research agenda, in: Proc. of IKM4DR, IKM4DR'12, ACM, New York, NY, USA, 2012, pp. 17–22. doi:10.1145/2389776.2389781.

[11] F. Radlinski, N. Craswell, A theoretical framework for conversational search, in: Proc. of CHIIR, CHIIR '17, ACM, New York, 2017, p. 117–126. doi:10.1145/3020165.3020183.

[12] J. Teevan, C. Alvarado, M. S. Ackerman, D. R. Karger, The perfect search engine is not enough: a study of orienteering behavior in directed search, in: E. Dykstra-Erickson, M. Tscheligi (Eds.), Proc. of CHI, ACM, 2004, pp. 415–422. doi:10.1145/985692.985745.

[13] M. Sanderson, W. B. Croft, The history of information retrieval research, Proc. of IEEE 100 (2012) 1444–1451. doi:10.1109/JPROC.2012.2189916.

[14] S.-C. Lin, J.-H. Yang, R. Nogueira, M.-F. Tsai, C.-J. Wang, J. Lin, Multi-stage conversational passage retrieval: An approach to fusing term importance estimation and neural query rewriting, 2020. arXiv:2005.02230.

[15] J. Jiang, W. Jeng, D. He, How do users respond to voice input errors?: lexical and phonetic query reformulation in voice search, in: G. J. F. Jones, P. Sheridan, D. Kelly, M. de Rijke, T. Sakai (Eds.), Proc. of SIGIR, ACM, 2013, pp. 143–152. doi:10.1145/2484028.2484092.

[16] J. R. Trippas, D. Spina, L. Cavedon, H. Joho, M. Sanderson, How do people interact in conversational speech-only search tasks: A preliminary analysis, in: Proc. of CHIIR, ACM, 2017, pp. 325–328. doi:10.1145/3020165.3022144.

[17] N. Sa, X. Yuan, Examining users' partial query modification patterns in voice search, Journal of the Association for Information Science and Technology 71 (2020) 251–263. doi:10.1002/asi.24238.

[18] A. Saha, V. Pahuja, M. M. Khapra, K. Sankaranarayanan, S. Chandar, Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph, 2018. arXiv:1801.10314.

[19] S. Reddy, D. Chen, C. D. Manning, Coqa: A conversational question answering challenge, 2018. arXiv:1808.07042.

[20] E. Choi, H. He, M. Iyyer, M. Yatskar, W. tau Yih, Y. Choi, P. Liang, L. Zettlemoyer, Quac : Question answering in context, 2018. arXiv:1808.07036.

[21] J. Dalton, C. Xiong, J. Callan, Cast 2020: The conversational assistance track overview, in: E. M. Voorhees, A. Ellis (Eds.), Proc. of TREC, volume 1266 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2020. URL: https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.C.pdf.

[22] A. Elgohary, D. Peskov, J. Boyd-Graber, Can you unpack that? learning to rewrite questions-in-context, in: Proc. of EMNLP-IJCNLP, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5918–5924. URL: https://www.aclweb.org/anthology/D19-1605. doi:10.18653/v1/D19-1605.

[23] S. Yu, J. Liu, J. Yang, C. Xiong, P. Bennett, J. Gao, Z. Liu, Few-shot generative conversational query rewriting, 2020. arXiv:2006.05009.

[24] Z. Chen, X. Fan, Y. Ling, L. Mathias, C. Guo, Pre-training for query rewriting in A spoken language understanding system, CoRR abs/2002.05607 (2020). URL: https://arxiv.org/abs/2002.05607. arXiv:2002.05607.

[25] S.-C. Lin, J.-H. Yang, J. Lin, Contextualized

query embeddings for conversational search, 2021. arXiv:2104.08707.

[26] S. Yuan, S. Gupta, X. Fan, D. Liu, Y. Liu, C. Guo, Graph enhanced query rewriting for spoken language understanding system, in: Proc. of ICASSP, IEEE, 2021, pp. 7997–8001. doi:10.1109/ICASSP39728.2021.9413840.

[27] N. J. Belkin, D. Kelly, G. Kim, J. Kim, H. Lee, G. Muresan, M. M. Tang, X. Yuan, C. Cool, Query length in interactive information retrieval, in: C. L. A. Clarke, G. V. Cormack, J. Callan, D. Hawking, A. F. Smeaton (Eds.), Proc. of SIGIR, ACM, 2003, pp. 205–212. doi:10.1145/860435.860474.

[28] N. Balasubramanian, G. Kumaran, V. R. Carvalho, Exploring reductions for long web queries, in: F. Crestani, S. Marchand-Maillet, H. Chen, E. N. Efthimiadis, J. Savoy (Eds.), Proc. of SIGIR, ACM, 2010, pp. 571–578. URL: https://doi.org/10.1145/1835449.1835545. doi:10.1145/1835449.1835545.

[29] M. Bendersky, W. B. Croft, Discovering key concepts in verbose queries, in: S. Myaeng, D. W. Oard, F. Sebastiani, T. Chua, M. Leong (Eds.), Proc. of SIGIR, ACM, 2008, pp. 491–498. URL: https://doi.org/10.1145/1390334.1390419. doi:10.1145/1390334.1390419.

[30] F. Li, H. V. Jagadish, Constructing an interactive natural language interface for relational databases, Proc. VLDB Endowment 8 (2014) 73–84. doi:10.14778/2735461.2735468.

[31] K. Affolter, K. Stockinger, A. Bernstein, A comparative survey of recent natural language interfaces for databases, VLDB Journal 28 (2019) 793–819. doi:10.1007/s00778-019-00567-8.

[32] E. Kuric, J. D. Fernández, O. Drozd, Knowledge graph exploration: A usability evaluation of query builders for laypeople, in: M. Acosta, P. Cudré-Mauroux, M. Maleshkova, T. Pellegrini, H. Sack, Y. Sure-Vetter (Eds.), Proc. of SEMANTiCS, volume 11702 of *LNCS*, Springer, 2019, pp. 326–342. doi:10.1007/978-3-030-33220-4_24.

[33] J. J. Leggett, G. Williams, An empirical investigation of voice as an input modality for computer programming, International Journal of Man-Machine Studies 21 (1984) 493–520. doi:10.1016/S0020-7373(84)80057-7.

[34] L. Rosenblatt, Vocalide: An ide for programming via speech recognition, in: Proc. of SIGACCESS, ASSETS'17, ACM, New York, NY, USA, 2017, pp. 417–418. doi:10.1145/3132525.3134824.

[35] A. W. Biermann, L. Fineman, J. F. Heidlage, A voice- and touch-driven natural language editor and its performance, International Journal of Man-Machine Studies 37 (1992) 1–21. doi:10.1016/0020-7373(92)90089-4.

[36] J. Martin, Managing the Data Base Environment, 1 ed., Prentice Hall PTR, USA, 1983.

[37] J. Kiesel, A. Bahrami, B. Stein, A. Anand, M. Hagen, Toward Voice Query Clarification, in: Proc. of SIGIR, ACM, 2018, pp. 1257–1260. doi:10.1145/3209978.3210160.

[38] H. Zamani, S. T. Dumais, N. Craswell, P. N. Bennett, G. Lueck, Generating clarifying questions for information retrieval, in: Y. Huang, I. King, T. Liu, M. van Steen (Eds.), Proc. of WWW 2020, ACM / IW3C2, 2020, pp. 418–428. doi:10.1145/3366423.3380126.

[39] J. Arguello, S. Avula, F. Diaz, Using query performance predictors to improve spoken queries, in: N. Ferro, F. Crestani, M. Moens, J. Mothe, F. Silvestri, G. M. D. Nunzio, C. Hauff, G. Silvello (Eds.), Proc. of ECIR, volume 9626 of *LNCS*, Springer, 2016, pp. 309–321. doi:10.1007/978-3-319-30671-1_23.

[40] C. Pearl, Designing Voice User Interfaces: Principles of Conversational Experiences, 1st ed., O'Reilly Media, Inc., 2016.

[41] A. Papenmeier, D. Kern, D. Hienert, A. Sliwa, A. Aker, N. Fuhr, Starting conversations with search engines - interfaces that elicit natural language queries, in: F. Scholer, P. Thomas, D. Elsweiler, H. Joho, N. Kando, C. Smith (Eds.), Proc. of CHIIR, ACM, 2021, pp. 261–265. doi:10.1145/3406522.3446035.

[42] J. A. Wald, P. G. Sorenson, Explaining ambiguity in a formal query language, ACM Transactions on Database Systems 15 (1990) 125–161. doi:10.1145/78922.78923.

[43] K. Balog, L. Flekova, M. Hagen, R. Jones, M. Potthast, F. Radlinski, M. Sanderson, S. Vakulenko, H. Zamani, Common Conversational Community Prototype: Scholarly Conversational Assistant, CoRR abs/2001.06910 (2020). URL: https://arxiv.org/abs/2001.06910.

[44] H. H. Clark, S. E. Brennan, Grounding in communication, in: Perspectives on Socially Shared Cognition, APA, 1991, pp. 127–149.