# Toward Voice Query Clarification

Johannes Kiesel
Bauhaus-Universität Weimar
johannes.kiesel@uni-weimar.de

Arefeh Bahrami
Bauhaus-Universität Weimar
arefeh.bahrami@uni-weimar.de

Benno Stein
Bauhaus-Universität Weimar
benno.stein@uni-weimar.de

Avishek Anand
L3S Research Center Hannover
anand@l3s.de

Matthias Hagen
Martin-Luther-Universität
Halle-Wittenberg
matthias.hagen@informatik.
uni-halle.de

## ABSTRACT

Query suggestions are a standard means to clarify the intent of underspecified queries. In a voice-based search setting, the compilation of query suggestions is not straightforward, and user-centric research targeting query underspecification is lacking so far. Our paper analyses a specific type of ambiguous voice queries and studies the impact of various kinds of *voice query clarifications* offered by the system and its impact on user satisfaction. We conduct a user study that measures the satisfaction for clarifications that are explicitly invoked and presented by seven different methods. Our findings include that (1) user experience depends on language proficiency levels, (2) users are not dissatisfied when prompted for clarifications (in fact, enjoy it sometimes), and (3) the most effective way of query clarification depends on the number and lengths of the possible answers.

## CCS CONCEPTS

• **Information systems → Query intent**; **Search interfaces**;

**ACM Reference Format:**
Johannes Kiesel, Arefeh Bahrami, Benno Stein, Avishek Anand, and Matthias Hagen. 2018. Toward Voice Query Clarification. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3209978.3210160

## 1 INTRODUCTION

In traditional IR systems, users are typically aided in formulation of their search intent by suggestions that help subsequent query modeling and query understanding tasks [2, 7]. The main goal of the suggestions is to help the user clarify their information need, to resolve query ambiguity, etc. In modern conversational search systems this is arguably an important and understudied aspect [1, 4, 5].

Although query suggestions can possibly cover a broad and sometimes ill-defined set of query underspecification—i.e., sparsity, clarity, ambiguity, vocabulary mismatch—, we focus on a specific set of voice query inputs; namely, intents that have ambiguous query formulations like entity names. We title the general task as *voice query clarification* and carry out a preliminary user-centric study towards better understanding of characteristics of voice query clarifications conditioned on ambiguous queries.

To our knowledge, this paper is a first foray in the topic of voice query clarification. Informally, it attempts to provide a first empirical study on the question of how to carry over the *did you mean. . . ?*-functionality to voice interfaces. In detail, we present a first step towards answering the following research questions, which are key for the integration of query clarification functionality into voice interfaces:

(1) **RQ I** How much does user satisfaction decrease when asked for clarification compared to just giving the correct answer?
(2) **RQ II** How does the user experience of the clarification vary for users with different background?
(3) **RQ III** How should the different clarification options be presented to maximize user satisfaction?

To this end, we conducted an inter-participant study in which the 14 participants had to solve 13 different ambiguous information needs using a self-made mock-up skill for the Amazon Alexa voice assistant. For each need, participants were presented with one of seven different clarification response types and had to rate on a 5-point Likert scale whether the assistant indeed answered their question, behaved as expected, was easy to understand, and was pleasant to use. After post-processing, we collected 708 judgements as well as participant background information and took extensive notes during the study (Section 3), which we then used for both quantitative and qualitative analyses (Section 4).

Our key findings are, among others, that (1) participants did not mind being asked for clarification—but sometimes even enjoyed it, (2) the user experience is severely different for participants with different levels of English proficiency, and (3) the best way to clarify depends on the amount and length of possible answers, where even listing several answers is preferable for very short answers. The results of our study can directly be employed in the design of voice-based interfaces for which ambiguity is an issue, and open the door for more focused research in the future.

*SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA*
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5657-2/18/07. . . $15.00
https://doi.org/10.1145/3209978.3210160

## 2 RELATED WORK

Query clarification (which includes disambiguation) has been studied extensively for text-based search interfaces (e.g., [2, 7]). One analysis related to conversations is that of Braslavski et al. [3], who focus on the dialogues between the users on a community question answering website that aim to clarify some posted question. They find that humans employ a wide variety of clarification questions which make the questions’ analysis promising for the future as information retrieval and question answering converge [1].

Query clarifications play an important role in conversational IR systems, but focussed respective research is still lacking [1, 4, 5]. Conversational systems are mixed-initiative systems between a user and an agent, where the agent responds based on a model of user needs taking into account both short- and long-term information from the user [6]. For such systems, Luger and Sellen [5] found that users have to be supported more in building their cognitive model of the conversational system while Vtyurina et al. [10] present an analysis of how future conversational search may look like.

Voice-based search is discussed as a promising supplementary to text-based search in specific situations. Several studies analyzed the conversational character of voice-based search. For example, Trippas et al. [8] examine differences between interactions in voice-based and text-based search while Trippas et al. [9] did a laboratory study where participants had to perform voice-based search by interacting with a human operator that simulates the system. Their finding that query refinement plays an important role, especially for more complex search tasks, directly motivates our research.

## 3 STUDY

In order to analyze user behavior in and preferences for voice query clarification, we created a mock-up skill for Amazon’s Alexa voice assistant and employed questionnaires as well as note-taking.

### 3.1 Setup

For the study, each participant had to (1) fill out a privacy-related consent form, (2) provide basic and study-related background information, (3) read the instructions for how to complete the tasks, (4) complete a small test-task (off the record), (5) complete the 13 main tasks, and (6) give comments and suggestions (optional).

Each task consists of a small scenario description with corresponding ambiguous query, an interaction phase between the participant and the voice assistant (called *system*), and an after-interaction questionnaire regarding how the interaction with the system was perceived. Participants were presented a sheet of paper with scenario, query (interaction start), and questionnaire (e.g., Figure 1). The participants had been instructed to read the scenario carefully, query the system as written, and then to continue—or restart in case of problems—the interaction until an answer was reached. Finally, they had to fill in the questionnaire by checking four boxes. We specified the query to make sure that it is indeed ambiguous.

In order to investigate how to best present the different clarification options, we programmed our mock-up to react to the participant’s prompt for a task in different ways. Specifically, we use 11 tasks where the ambiguity stems from one word and the following 7 response methods (here grouped into baseline, standard, and many-option methods) are used for clarification:

**Baselines** (no clarification)

- **Direct** (2 tasks) Answer the query for one meaning, either the desired one (1 task, called *hit*), or not (1 task, *miss*).
- **Concatenate** (1 task) Answer the query for three possible meanings (including the desired one) with one sentence each.

**Standard** (clarification for few options)

- **3-meanings** (2 tasks) Ask to choose from 3 meanings of the ambiguous word (described by 1–5 words). The desired meaning is in the list (1 task, *hit*) or not, in which case participants can describe the meaning themselves or ask for more (1 task, *miss*).
- **3-long-meanings** (2 tasks) Like 3-meanings, but meanings are described by 8–16 words (speaking takes about twice as long).
- **Verify** (2 tasks) Ask to verify if a specific meaning is the desired one. Either “yes” (1 task, *hit*) or “no,” in which case continue with 3 meanings to choose (including the desired, 1 task, *miss*).

**Many-options** (clarification for queries with many meanings)

- **5-meanings** (1 task) Like 3-meanings, but with 5 meanings presented at a time.
- **3-categories** (1 task) Like 3-meanings, but first ask for a category, then continue with 3-meanings within that category. This is inspired by Wikipedia disambiguation pages, where meanings are often grouped by category.

For comparison, we included 2 tasks where the ambiguity stems from an acronym and use 3-meanings response method in this case.

For the study, we avoided voice recognition errors, which occur frequently in today’s voice interfaces, by using a tightly fit interaction model. In detail, the Alexa recognition model was trained to listen specifically for the keywords and phrases in the clarification options (e.g., *cocktail* or *Irish* for the example in Figure 1), as well as list index words (e.g., *first* or *last*). Participants were able to use these phrases to specify the desired meaning, which they discovered on their own. Since our mock-up is restricted to these few phrases (less than 100 in total), recognition worked very well, with only few cases were participants spoke too quietly. Therefore, our results transfer to future interfaces with better recognition.

To avoid biases, we randomized the task-order for each participant and the position of the desired meanings in the options-lists, and switched between two scenarios for the response methods. To check whether there actually are order-biases, we used Fisher’s exact test, but found no significant effect to show their existence.

### 3.2 Participants

We recruited 14 participants from our University’s Computer Science and Civil Engineering programme for the main study. Beforehand, we tested our setup in a pilot-study with 3 participants, which are not considered in the results. We requested the participants to rate their English level, but corrected for our analysis the rating of five participants where their interaction with the system was way more/less fluent than the self-rated level would suggest.

In the main study, the participants were between 18–30 (9 participants) and 31–49 years old (5). Furthermore, 9 were male and 5 female. Participants had an intermediate (5 participants) or proficient (9) English level. Finally, 8 participants stated to never use voice assistants, whereas 5 use them rarely and 1 uses them frequently. Therefore, our participants are all adults that can be seen as novice users of voice assistants.

**Scenario:** You want to surprise your Irish partner with an Irish cocktail called B-52, but you don't know how to make it.
**Interaction start:** Alexa. Find! How to do a B-52?

| **After interaction:** | Agree | | Neutral | | Disagree | Don't know |
|---|---|---|---|---|---|---|
| The system answered my question | □ | □ | □ | □ | □ | □ |
| The system behaved as I expected | □ | □ | □ | □ | □ | □ |
| The system was easy to hear/understand | □ | □ | □ | □ | □ | □ |
| The system was pleasant to use | □ | □ | □ | □ | □ | □ |

**Figure 1: One of the 13 tasks the study participants were told to do. They were instructed to start the interaction by saying "Alexa, Find!", wait for the system to react, and then to follow up with the provided question. They should then continue the interaction until the system responds with an answer. After that, they should rate their experience using the checkboxes.**

## 3.3 Data

The 14 participants required between 15 and 25 minutes for completing the study. Each participant finished 13 tasks for a total of 182 interaction phases. From these phases, we filtered out 5 due to participants being unfamiliar with the scenario (3 times), them afterwards saying they paid no attention (1), or a mock-up bug (1). For each phase, we collected 4 ratings (Figure 1), for a total of 708.

# 4 RESULTS

Our research questions we raised in the introduction mainly focus on user satisfaction, for which the system should answer the question, behave as expected, be easy to hear/understand, and be pleasant to use. Since, as expected, participants most times somewhat agreed that the system was easy to hear/understand (97% of the ratings) and answered the question (95%, disregarding the direct (miss) response method that fails by design), we mainly show results for the other criteria. For statistical testing we always use Fisher's exact test is suited for such small sample sizes.

## 4.1 How much does user satisfaction decrease when asked for clarification?

We compare the verify response method (ask the user to verify the meaning) to the direct method (answers without clarification). As Figure 2 shows, participants rate both methods very similarly when the system correctly assumes the desired meaning (i.e., a "hit"). Therefore, user satisfaction seems to be not negatively affected by asking for clarification. However, when the system incorrectly assumes a different meaning (i.e., a "miss"), the participants' ratings drop in both cases, albeit more so for the direct response method that fails "by design" (ratings not shown in Figure). Interestingly, without being asked, five participants commented after the study that they had fun in interacting with the clarification system. In conclusion, these results suggests that voice assistants should always ask for clarification when they detect an ambiguous query as users seem generally open to answer such requests.

## 4.2 How does user experience vary for users with different background?

For this question, we analyze whether the ratings for participants with different proficiency in English and with different usage experience with voice assistants differ significantly.

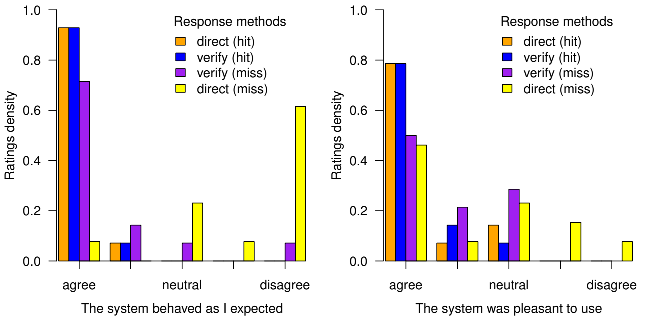

**Figure 2: Response-specific ratings for predictability and pleasantness.**

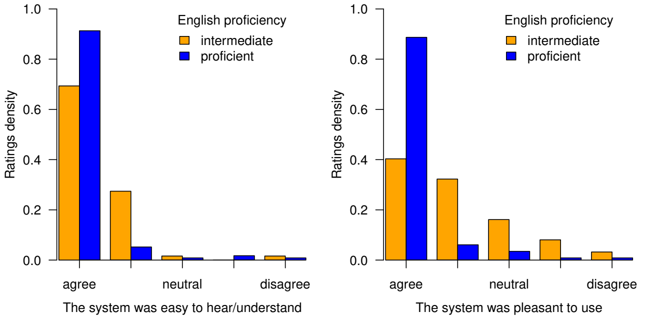

**Figure 3: Overall ratings for understandability and pleasantness by English proficiency of the participants.**

For English proficiency (Figure 3), we found that participants with a higher proficiency level rated the system both easier to hear/understand and more pleasant to use ($p < 0.001$ for both). As the figure shows, the effect is relatively strong for pleasantness (Pearson's $r = -0.44$), which suggests that voice assistants should account for the user's language proficiency when asking for clarification. We analyze this idea for different response methods below. Overall, participants rated the system as easily understandable, which shows the quality that voice assistants already reach today.

As for experience with voice assistant usage (Figure 4), we found that participants that use voice assistants rated the system both as easier to hear/understand and more pleasant to use ($p < 0.001$ for both), but also gave more extreme ratings regarding predictability ($p < 0.05$). Since only one participant uses voice assistants very frequently, we ignore them for the statistical analysis. However, as the effect is not as strong as for English proficiency ($|r| < 0.09$),

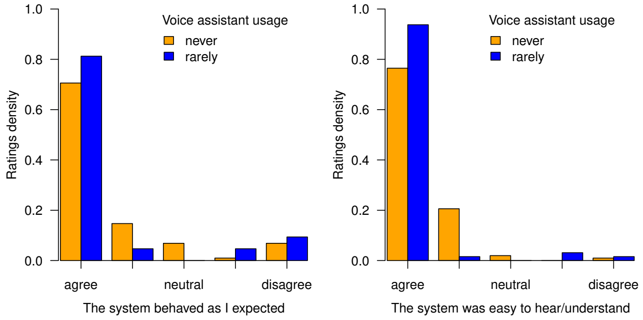

**Figure 4: Overall ratings for predictability and understandability by frequency of a participant's voice interface usage.**

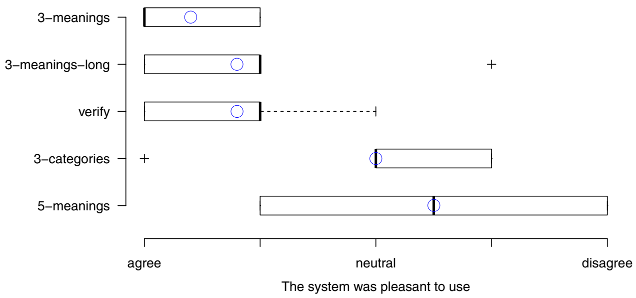

**Figure 5: Distribution and mean (○) pleasantness for participants with intermediate English level by response method.**

considering the user's familiarity with voice assistants seems not as important for the system design. We assume that the reason for the more extreme predictability ratings of more experienced users is that they entered the study with clear expectations on voice-based interactions, which then conformed or collided with the mock-up.

## 4.3 How should the different clarification options be presented?

Due to the results of Section 4.2, we approach this question separately for participants with intermediate and proficient English level. Specifically, we map the rating from agree to disagree onto the range from 1 to 5 and use the mean as a measure of satisfaction (lower is better). For this comparison, we focus on pleasantness ($\mu_p$) as the main indicator of overall satisfaction. While the mean for participants with proficient English level is similar and low ($\mu_p < 1.4$) for all response methods, the means are more spread out for participants with intermediate level. As Figure 5 shows, the 3-meanings response method is ranked as most pleasant ($\mu_p = 1.4$), but closely followed by 3-meanings-long and verify. Despite our small sample size, the difference is significant between 3-meanings and 3-categories ($p < 0.01$) and between 3-meanings-long and 3-categories ($p < 0.05$). This suggests that the 3-meanings response method should be preferred as it is the most pleasant to use for intermediate English speakers out of the methods we tested.

For queries with many clarification options, our qualitative analysis of the study protocols suggests that systems should allow users to specify the meaning themselves. In our setup, participants preferred to specify the meaning themselves over asking for more options. In fact, 10 of the 14 participants tried to interrupt Alexa in order to specify the meaning instantly. Notably, this occurred for all response methods except for 3-meanings(-long), which suggests that 3 should be the preferable list size for clarification options.

For queries with short answers, we tested whether it is feasible to give all the answers without asking for clarification. In detail, participants were asked to use a person-specific query ("Who is Heisenberg?") and received an one-sentence answer for each of three entities, separated by small pauses. Participants were very pleased by this response ($\mu_p = 1.2$), which suggests that this response method is very appropriate in case of short answers.

For completeness, we analyzed whether participants perceived a difference between queries with ambiguous words and such with ambiguous acronyms, but found no statistically significant difference in their ratings.[1] Therefore, our results are applicable for both.

## 5 CONCLUSION

We have conducted a first step towards the task of voice query clarification with a user-centric study. We identified three key research questions for voice query clarification and used our study to provide some first answers. In doing so, we found—among others—that there is no penalty in user satisfaction when the system asks for clarification, that the users' English proficiency is an important factor for designing clarification options, that three clarification options are the recommended number of choices, that users should be given the possibility to interrupt the system and clarify the query themselves, and that listing the different possible answers is preferable over asking for clarification when answers are short.

Our findings open the door for further and focussed research on the topic of voice query clarification. Specifically, promising more "algorithmic" directions could be to investigate alternative response methods that may be even more suitable and to identify strategies for choosing the best response method in different scenarios.

## REFERENCES

[1] James Allan, W. Bruce Croft, Alistair Moffat, and Mark Sanderson. 2012. Frontiers, challenges, and opportunities for information retrieval: Report from SWIRL 2012. *SIGIR Forum* 46, 1 (2012), 2–32.
[2] Sumit Bhatia, Debapriyo Majumdar, and Prasenjit Mitra. 2011. Query suggestions in the absence of query logs. In *SIGIR*. ACM, 795–804.
[3] Pavel Braslavski, Denis Savenkov, Eugene Agichtein, and Alina Dubatovka. 2017. What Do You Mean Exactly?. In *CHIIR*. ACM, 345–348.
[4] Jennifer Lai and Nicole Yankelovich. 2006. Speech Interface Design. In *Encyclopedia of Language & Linguistics*, Keith Brown (Ed.). Elsevier, Oxford, 764–770.
[5] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *CHI*. ACM, 5286–5297.
[6] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *CHIIR*. ACM, 117–126.
[7] Fabrizio Silvestri. 2010. Mining Query Logs: Turning Search Usage Data into Knowledge. *Foundations and Trends in Information Retrieval* 4, 1-2 (2010), 1–174.
[8] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the Design of Spoken Conversational Search: Perspective Paper. In *CHIIR*. ACM, 32–41.
[9] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, and Mark Sanderson. 2017. How Do People Interact in Conversational Speech-Only Search Tasks: A Preliminary Analysis. In *CHIIR*. ACM, 325–328.
[10] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles L. A. Clarke. 2017. Exploring Conversational Search With Humans, Assistants, and Wizards. In *CHI Extended Abstracts*. ACM, 2187–2193.

[1] We only tested the 3-meanings response method for acronyms