# The Role of Conversational Grounding in Supporting Symbiosis Between People and Digital Assistants

JANGHEE CHO, Department of Information Science, USA University of Colorado Boulder, USA
EMILEE RADER, Department of Media and Information Michigan State University, USA

In "smart speaker" digital assistant systems such as Google Home, there is no visual user interface, so people must learn about the system's capabilities and limitations by experimenting with different questions and commands. However, many new users give up quickly and limit their use to a few simple tasks. This is a problem for both the user and the system. Users who stop trying out new things cannot learn about new features and functionality, and the system receives less data upon which to base future improvements. Symbiosis—a mutually beneficial relationship—between AI systems like digital assistants and people is an important aspect of developing systems that are partners to humans and not just tools. In order to better understand requirements for symbiosis, we investigated the relationship between the types of digital assistant responses and users' subsequent questions, focusing on identifying interactions that were discouraging to users when speaking with a digital assistant. We conducted a user study with 20 participants who completed a series of information seeking tasks using the Google Home, and analyzed transcripts using a method based on applied conversation analysis. We found that the most common response from the Google Home, a version of "Sorry, I'm not sure how to help", provided no feedback for participants to build on when forming their next question. However, responses that provided somewhat strange but tangentially related answers were actually more helpful for conversational grounding, which extended the interaction. We discuss the connection between grounding and symbiosis, and present recommendations for requirements for forming partnerships with digital assistants.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: digital assistant; conversational system; qualitative analysis; conversational analysis; common ground; voice interface; symbiosis; Google Home; Human-AI interaction

**ACM Reference Format:**
Janghee Cho and Emilee Rader. 2020. The Role of Conversational Grounding in Supporting Symbiosis Between People and Digital Assistants. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 33 (May 2020), 28 pages. https://doi.org/10.1145/3392838

## 1 INTRODUCTION

Digital assistants integrated into commercial "smart speaker" systems, like those sold by Google (Google Assistant), Amazon (Alexa), and Apple (Siri), use automatic speech recognition and natural language processing to appear to be able to converse naturally with people [31, 39, 46]. They allow people to speak like they would to another person, and receive an answer back in a conversational

Authors' addresses: Janghee Cho, Department of Information Science, USA, University of Colorado Boulder, USA, janghee.cho@colorado.edu; Emilee Rader, Department of Media and Information, Michigan State University, USA, emilee@msu.edu.


2573-0142/2020/5-ART33 $15.00
https://doi.org/10.1145/3392838

tone and sentence structure. Digital assistant systems are advertised as tools that can be used effectively without formal training (e.g., "Your Google Assistant is ready to help, anytime, anywhere.")[1]. In other words, these systems are intended to be 'walk up and use': self-explanatory even for people who have never used them before [43].

Because smart speaker systems commonly do not have a visual interface—only an indicator signaling when the system is awake and listening for a command—people cannot navigate graphical menus to discover commands that the system can respond to. Instead, they must encounter the system's features through trial and error. They do this by asking questions and hearing how the system responds [8]. However, once people encounter examples of the system's limitations, they can become reluctant to continue exploring the system's features [10, 50]. For example, Bentley et al. [5] found that after a very short period of initial exploratory use lasting only a few days, Google Home users seemed to only use the system for a small number of different tasks. Also, once they stopped exploring the system's functionality, they rarely changed their usage patterns over a period of several months.

Users who stop trying to discover new or unfamiliar functionality present a challenge for the design of so-called "artificial intelligence" (AI) systems which, like digital assistants, rely on large datasets and machine learning algorithms to shape the user experience. This is because the smart speaker device in a user's home is merely a means of accessing a cloud-based system. The functionality of the system is not limited by the physical hardware of the speaker device. Instead, the speaker device is essentially a client front-end for a computational infrastructure that exists on networked servers, and the system's ability to respond to questions can be improved without having to alter the device itself [42]. This means that the system can be improved over time without any visible indication towards the user. Therefore, users must continue experimenting—trying out new questions and commands—in order to learn about updated functionality.

Continued interaction with digital assistants is also important from a system perspective, as more frequent and longer conversations with users generate data that the system tracks and uses to improve its performance [24]. In this way, people and digital assistants can be thought of as being in a mutually beneficial, symbiotic relationship, as described by Licklider [29] in his seminal paper, *Man-Computer Symbiosis*. People benefit from the convenience afforded by digital assistants via a 'walk up and use' voice interface which provides access to information and the ability to execute hands-free commands. Meanwhile, the system benefits because as people use it, data are collected about their interactions with it which the system uses to become more effective. Many AI systems, not just digital assistants, have the potential to be symbiotic in this way. However, Grudin and Jacques [21] argue that while many software systems currently work "autonomously around the clock on our behalf", these systems have not yet achieved the equal partnership promised by symbiosis.

Digital assistants that do not have screens, like the Google Home, are an interesting case for investigating the potential of symbiotic partnerships with AI systems. This is because in order for digital assistants and other chatbot systems to become true partners to humans, Grudin [20] argues that advances in AI technologies on their own are not enough. Behavioral research focusing on the interactions between people and these systems is also needed. Conversation is the primary medium for interacting with voice user interfaces [7], and therefore people's attempts to interact with smart speakers are shaped by the rules and processes that govern human conversation. This means that designers must combine AI advances with behavioral science findings about the nature of human conversation in order for voice-based user interfaces to become effective conversational partners.

[1] https://assistant.google.com/learn/, retrieved October 2019

According to Clark's theory of *common ground* [11], in order for two human interlocutors (conversation partners) to successfully have a conversation, they must make spoken contributions to the conversation which signal that they understand each other well enough to build on what the other person has said. Through the conversation, they each build up common ground, which is an individual mental representation of the shared knowledge and beliefs held by the participants in the conversation that develops over time. When the conversation partner is a digital assistant, not another person, the responses and answers it provides become part of the human user's mental representation of how well (or poorly) the system is able to respond to them, and what kinds of questions it is capable of answering. This means that the process by which common ground is formed—called grounding—is a key part of supporting continued experimentation and learning about the system. We argue that studying the difficulties people have with interacting with digital assistants can help human-AI interaction researchers better understand the requirements for truly mutually beneficial symbiotic relationships with AI systems.

In this study, we focus on the initial interactions between people with no previous experience using a smart speaker system, and the Google Home, a commercially available smart speaker product that is a front end for the Google Assistant. The goal of this research was to identify ways that the Google Home's responses might fail to provide opportunities for users to extend their interactions with the system, thereby preventing them from learning about its capabilities. A symbiotic relationship between the user and the system only works if the system provides opportunities for the user to build on what the system just said, even when it might not be clear how the system should respond. We suspected that problems with the grounding process may be related to why users struggle in their early interactions with digital assistants. This insight led us to focus on grounding-related barriers to developing a symbiotic relationship between people and AI-based digital assistant systems.

We conducted a user study in a lab setting with 20 participants in which they completed several information seeking tasks using the Google Home, and used an approach guided by common ground theory [11] and conversation analysis [55] to analyze pairs of back-and-forth utterances between the participant and the Google Home. We found that the Google Home exhibits usability problems related to conversational grounding that discourage user experimentation, exploration and learning about new functionality. However, we also found that some of the Google Home's responses that were only tangentially related to the question that was asked showed participants that the Google Home was knowledgeable about a topic and able to engage in a conversation with them about it. This feedback allowed participants to speculate about how to ask new questions that helped them make progress on tasks. This paper contributes new findings to the research literature about general purpose digital assistant systems through an investigation of a commercial smart speaker product that has been widely adopted. Based on our findings, we discuss the importance of conversational grounding for symbiosis between humans and AI systems, and present recommendations for ways to better support a mutually beneficial partnership.

## 2 RELATED WORK

### 2.1 Digital assistants in Everyday Settings

Previous research has investigated ways to improve voice user interface performance on specific tasks, such as robot navigation [33] or conference room scheduling [6]. However, as commercially-available smart speaker devices have become increasingly popular, researchers have begun to study the use of similar systems in everyday settings and for many different kinds of tasks. Unlike single-task voice user interface systems, digital assistants such as Google Assistant and Amazon Alexa are marketed as general-purpose information seeking and management appliances useful in users' kitchens, living rooms and bedrooms, and as interfaces to smart home controls and entertainment systems[2]. When people first start using these devices, they try out different kinds of questions requests in an effort to figure out what the system is capable of [50], experimenting with the system's functionality so they can learn how to better use it. This initial experimentation period typically lasts only a few days, during which time the variety and frequency of requests is higher than later on. Stable long-term levels of usage include only a small subset of the initial commands attempted by users [2, 5].

These results indicate that early use is important for how people come to understand the limitations of these devices. When people are unsuccessful at using their digital assistants for accomplishing tasks, they assume that this is because the system's functionality does not support what they are trying to do. Disappointing performance in the initial experimentation phase of use, as described by participants in one study, led to lowered expectations for future use of the system [17]. For instance, participants in multiple studies expected that a digital assistant or conversational agent system would be able to learn from the context of their preceding conversations with it and use that information to infer their intentions on the current task [17, 31], as human interlocutors do. Unmet expectations like this were one of the main reasons people ultimately limited their use of digital assistants, or abandoned them completely [10]. Moreover, users had trouble knowing what their digital assistant was supposed to be capable of doing [31], and found it difficult to discover features through trial and error that they were not already familiar with [50].

Several studies presented analyses of conversations illustrating how interacting with digital assistants is integrated into regular family conversation. In these situations, a back-and-forth interaction to arrive at a successful question or command can become a collaborative effort among multiple people and the device [4, 45]. Asking a question of a digital assistant in a group setting makes both the question and the success or failure of the response a part of the larger group's interaction [46]. In addition, children develop communication habits based on interactions with digital assistant devices which then spill over into their conversations with other family members [18]. Our study complements this work by examining use of a commercial, general purpose digital assistant in a controlled lab setting to investigate what might lead users to stop experimenting with new questions and commands.

## 2.2 Conversation and Common Ground

Understanding human-to-human conversation is essential for analyzing the interaction between people and digital assistants. Conversations are collaborative activities that two or more people engage in together, in which each person must "coordinate both the content and process of what they are doing" [12, p. 127] in order to meet the goals they have for the conversation. Clark and Schaefer [13] described a conversation as being organized in pairs of adjacent utterances where one conversation partner makes a contribution, and the other person responds. In order to successfully have a conversation, each partner must present evidence that they think they understood what the other person meant or wants to know. To establish common ground—an individual mental representation of others' knowledge, beliefs, and motivations—each conversation partner acts in response to the other, and coordinates with them by aligning their knowledge and experience with their partner [11]. This often looks like tailoring what they say based on the common ground they share with the other person [16]. For instance, people are willing to provide more context and put more effort into grounding if they don't share much common ground with the interlocutor [16, 27].

During the grounding process, interlocutors look for positive evidence from their conversation partner that they have been understood. This can look like continued attention, backchannel

[2] https://assistant.google.com/platforms/speakers/, retrieved October 2019

responses, or contributions to the conversation that are relevant to what's already been said. They also look for negative evidence that indicates they have been misunderstood, or their conversation partner is not paying attention, such as an unrelated next conversational turn or eye gaze that signals inattention. This evidence serves a coordination function, allowing the conversation to progress (or not) toward meeting the goals of the people involved.

Mutual belief that all parties in a conversation understand each other and are talking about the same thing is important for interlocutors to feel able to continue the conversation [12]. This means that the human user needs to receive some kind of evidence from the smart speaker device about how it has understood them, in order to make progress toward finding answers to their questions. Of course, the Google Home does not actually believe or understand anything, because it is a device, not a sentient being. But these conversation processes are automatic for people who are speaking with one another, and this means that they come into play when speaking with a digital assistant, too [3, 46]. For example, when a digital assistant answers a question from a user in an unexpected way, people respond in ways similar to when they're talking to a person who doesn't understand them: they try to speak more clearly and loudly or rephrase their question by adding, removing, or changing the order of words [4, 25]; they repeat the question verbatim or with small modifications to vocabulary or grammar [41, 46]; and they attempt to identify what might be causing the problem so that they can try again with a new approach [41, 45]. This also means that a human user needs to receive some kind of feedback from the digital assistant in order to make progress toward finding answers to his or her questions. Fischer et al. [15], referencing Schegloff [49] wrote about this as "progressivity" and emphasized the importance of continuous progress in conversation with a digital assistant.

### 2.3 Correcting Errors and Communication Breakdowns

The difficulty that voice user interfaces have with handling communication breakdowns is well recognized, and researchers have been working on the problem of how to better support conversational grounding in these systems for decades [34, 36]. Most works focus on reducing the occurrence of breakdowns by detecting potential errors accurately and recovering from miscommunication efficiently [6, 47, 52]. For instance, Bohus and Rudnicky [6] examined which type of recovery strategies are more likely to correct an error. They conducted a Wizard-of-Oz experiment to determine whether users correctly understood the system's utterances and were able to obtain the requested information. They also measured the amount of time spent by the system to recover from the communication breakdown, to assess the level of recovery efficiency. However, this study and other similar research focused on constrained problem spaces (e.g., single-task systems and/or controlled experiments [23, 26, 33]) where defining and measuring correct task performance is more straightforward, and where the boundaries of the system's abilities are more clearly articulated. These systems and how they are used are different from general purpose commercial digital assistants like Amazon Alexa and the Google Home.

Several recent studies have presented evidence of the different kinds of communication breakdowns that occur when people are using commercial digital assistants. While remarkable progress has been made in voice user interfaces [35], most of the burden of recovering from communication breakdowns still rests on the people using them [3, 4]. For example, speech recognition remains a problem for users, which manifests in the system being unable to provide the expected response, or providing responses that seem unrelated to the question or request [25, 41, 45]. Beneteau et al [4] also mentioned that current digital assistants still have limited ability to detect contextual details. We build on this work by investigating how inexperienced users interacting with a digital assistant for the first time adapt to responses the system provides. We view breakdowns as opportunities for learning—as common ground theory does. While communication breakdowns are inevitable, a better understanding of how people adapt to and learn from these breakdowns will help to clarify the kinds of responses that encourage continuous experimentation and provide opportunities for a symbiotic relationship to develop between people and digital assistants.

## 3 METHOD

### 3.1 The Google Home

The Google Home is a "smart speaker" device with an integrated microphone that enables people to talk to it and interact with services provided by the Google Assistant, a digital assistant developed by Google, Inc. To use the Google Home, a person must first speak the "wake word" ("Hey Google" or "OK, Google"), which signals to the device that the next thing that is said is a question or request directed towards the device. The Google Home then speaks a response. This process is operated by its dialog system module (i.e., Dialogflow) designed to match voice data to the best intent in the system and to generate a response. The Google Home generates a response by a combination of web search results and a predetermined dialog process (i.e., Action)[3]. By leveraging machine learning techniques, each step in the dialog system module becomes improved over time[4]. The speaker device has no display, but there are LED lights on its top that provide visual feedback about whether it is "awake" and listening after the wake word has been spoken. Our focus in this research is on digital assistants that are a part of standalone devices that do not have screens, referred to as "smart speakers" in the commercial market. Because these devices do not have screens, all of the feedback to the user must be spoken by the digital assistant itself.

We selected the Google Home for this study because it uses digital assistant technology, and is becoming increasingly popular: 43 million Google Homes devices were sold as of December 2018 in the U.S., installed in 23% of households [44]. Ammari et al. [2] reported that "search or informational queries" is the most common use of the Google Home, and it is also reported to have the highest accuracy when answering the greatest variety of questions [14]. In addition, a comparative study of digital assistants published in 2017 found that the Google Assistant outperformed competitors Siri (Apple), Cortana (Microsoft) and Alexa (Amazon) in terms of having the most natural responses [30].

### 3.2 Participants

Twenty participants were recruited through a paid subject pool consisting of members of the community surrounding a large public university located in the Midwest region of the United States. Eligible participants were at least 18 years old, native speakers of American English, had no experience with voice-enabled smart speaker devices used in the home, and had little or no experience with mobile-driven digital assistants. We recruited only native speakers of American English because digital assistants, including the Google Home, are notorious for large variations in speech recognition accuracy by accent group [22]. Our study did not focus on evaluating digital assistants' automatic speech recognition, so we excluded non-native speakers of American English to avoid variation in system performance due to participants' accents. In addition, initial use of these systems has been shown to differ from longer-term use [5, 31, 50]. A sampling frame focused on a more experienced population would have prevented us from studying the grounding process during the initial phase of use. Therefore, we restricted eligibility to only those without prior experience using digital assistants to simulate the important early experimentation phase of use. We also excluded participants who were technology experts or worked at any company related to digital assistants, because their expert knowledge about how the technology works could affect their interactions with it. Eligible participants were selected to participate on a "first-come-first-served"

[3]https://cloud.google.com/dialogflow/docs/basics, retrieved January 2020
[4]https://dialogflow.com/, retrieved January 2020

basis, prioritizing gender balance (10 women and 10 men) and age diversity. Participants' average age was 40, and ranged from 19 to 71. All identified as "White", and one participant also identified as "Asian".

### 3.3 Task Design

We designed five information seeking tasks for participants to complete using the Google Home [32]. The tasks instructed participants to engage in a dialogue with the Google Home, but did not provide explicit directions for how to do so. The tasks were based on our pilot testing and available documentation about the Google Home at the time we conducted the study during the Fall of 2017. Three tasks were designed to be *In Scope* with respect to the kinds of information-seeking questions the Google Home should have been able to provide reasonable answers for at that time. Therefore, we expected that it would be possible for the participants to find the information the In Scope tasks requested. For example, one of the explicit features of the Google Home was providing help with finding information for planning an upcoming trip[5] (Task 1, below). Our testing also showed that the Google Home was usually successful at providing answers to simple factual questions.

Two tasks were deliberately designed to be *Out of Scope*, meaning that we did not restrict these tasks to the advertised capabilities of the Google Home[6]. For these tasks, we selected topics which might come up in a human-only conversation where the desire might arise to search for more information on the Internet, which is a use case others have identified for digital assistants [4, 46]. In other words, these tasks approximate the kinds of things a voice-activated smart speaker device in one's home might be asked to do in real life. We expected that finding the requested information on the Out of Scope tasks would be difficult for participants. We included both types of tasks because we believed participants would receive different kinds of responses and feedback to the different types of tasks, and we wanted to generate a wide range of experiences from which participants could learn about the Google Home's capabilities. There were only two Out of Scope tasks, compared with three In Scope tasks, because, based on evidence from our pilot testing, we expected these to take longer for participants to complete. Below are the task descriptions for the five tasks:

(1) *In Scope*: Assume that you are planning to take a trip to Japan in the near future. You would like to find useful information about flights, accommodations, and features of interest in Japan. Please ask Google Home about it. When you finally decide where to go, how to get there, what to do, and so on, please say 'Thanks, Google.' (Please select any time period you might want to go.)
(2) *In Scope*: Assume that you want to purchase a new television. You want to find a place where you can buy one near your home. Please ask Google Home about it. When you decide where you want to buy one, please say 'Thanks, Google.' (Please assume that this place is your home.)
(3) *In Scope*: Assume that you need a passport to travel overseas, but you do not know how to issue it and where to go for it. Please talk with Google Home to find solutions.
(4) *Out of Scope*: Assume that you just read news about Obamacare and you are not sure if you want to sign up for it. So, you want to check Google Home's opinion on Obamacare in order to make a decision on whether you should enroll. Ask for Google Home's opinion.
(5) *Out of Scope*: Assume that you are a big fan of the Detroit Pistons basketball team and you hope they will advance in the Playoffs this season. You may want to know Google Home's prediction about it. Please ask Google Home about it. (NBA Playoffs are a tournament among the best 16 teams in the league.)

[5] https://support.google.com/googlehome/answer/7128171?hl=en, retrieved January 2020
[6] located at https://support.google.com/googlenest/topic/7195017 as of January 2020

### 3.4 Procedure

We conducted an in-person user study from Nov. 28, 2017 to Jan. 11, 2018. We purchased a new Google Home for this study so that the device would not have any history associated with it. Before each study session, we created a new Google account for each participant and associated it with our device, because we did not want previous participants' task history to influence the Google Home's answers. However, it is possible that the system tracked the device's usage history using an identifier that we could not change. Also, the Google Home's responses were personalized for the location where the study took place.

We obtained Institutional Review Board (IRB) approval for the research, and informed consent was given by all participants prior to the study. Participants were given brief instructions on how to use the "wake" word to activate the Google Home. While all participants completed the tasks in the same order, we varied which task participants started on to partially control for learning effects. Note that there was little variation in the number of questions participants asked in each task, on average, based on which task they completed first[7]. There was no time limit, and participants stopped each task on their own when they were satisfied with the answer or could not think of anything else to ask. After each participant had finished all 5 tasks, they had an opportunity to ask the Google Home anything they wanted. Finally, the first author asked some general follow-up questions like, "Tell me what you thought about the interaction with the Google Home during the task." The questions were designed to be open-ended and non-leading, and sought to understand the participants' impressions and reactions after using the Google Home [40]. Study sessions lasted 50–60 minutes, and participants received $20 as a thank you for participating. The supplementary file provides the screening questionnaire, participant instructions, and post-task questions.

### 3.5 Limitations

This study used a small convenience sample, which means our findings are not generalizable to all Google Home users. Also, the tasks we chose constrained what participants asked the Google Home. By recruiting users who had no experience with the Google Home, it allowed us to observe the interactions of users who did not have well-formed expectations about what it could do. However, our results may have been different if we had recruited more experienced users, who might have given up more quickly or asked questions in ways they expected to be effective based on their past experience. We also would have observed different interactions with the Google Home if we had conducted a longitudinal field study. This would have enabled analysis of behavior change in situ over time; however, we opted for a lab-based study so that we could give all participants the same tasks to complete and thereby compare task progress across participants. Another limitation is that although we showed the participants how to use the wake word to speak to the Google Home and had them practice it, two participants did not do this consistently, which hampered their ability to interact with the device. (Fourteen out of 20 participants forgot to use the wake word at least once during the study.) Finally, we conducted the study sessions with a live system connected to our campus network, and as such we were not able to hold the functionality or performance of the Google Home constant across all study sessions. This more closely approximates real-world use of the device, but it also means that the study sessions could have been affected by fluctuations in the performance of the Google Home service or the network in ways that were out of our control. We can't guarantee that the functionality did not change during the data collection period, but we did

[7] All participants completed the tasks in the same order, but a different entry point to conduct tasks. P01, P06, P11, P16 started with Task 1; P02, P07, P12, and P17 started with Task 2; P03, P08, P13, and P18 started with Task 3; P04, P09, P14, and P19 started with Task 4; and P05, P10, P15, and P20 started with Task 5. See the supplementary file for details.

not observe differences in the system's ability to respond to participant questions within the tasks that they all completed.

# 4 ANALYSIS

## 4.1 Theoretical Background for Analysis

The theory of common ground describes conversation as "a joint action projected by one of its participants and taken up by the others" [11]. In this study, the *joint action* takes place between the participant and the Google Home, and is bounded by the task description. The task instructs the participant to ask for information they must find out from the Google Home. This means that the goal of the conversation is also the goal of the task: to find out the requested information. In addition, the human interlocutor in this study also has another, secondary task: to learn about how to coordinate the content and process of the conversation with the digital assistant in order to have a successful conversation and complete the task. We focused our analysis on the types of questions the participants asked of the Google Home, and the categories of responses the Google Home provided, to find out whether and in what ways the Google Home was able to provide information that participants interpreted as progress toward completing the task.

If participants interpreted the Google Home's responses as providing feedback that continuing the conversation would help them move closer to completing the task, then we would expect based on common ground theory that their subsequent contributions to the conversation—the questions they asked of the Google Home—would reflect this by building on what the Google Home has just said. For example, this would look like following up on a specific piece of information in the Google Home's response, or asking a more narrowly focused version of the previous question. These types of responses would indicate that the participants assume that they and their interlocutor (the Google Home) have established some common ground and are working toward the same conversational goal, which in our study was finding the information to complete the task. If, however, participants interpreted the Google Home's responses as an indication that it would not be able to help them complete the task, this provides feedback about the Google Home's capabilities—or lack thereof—that could discourage future interaction. In this case, we would expect that participants may ask the same question in a slightly different way, as a more general question, or attempt to ask a totally different question altogether.

In our analysis, which is based on conversation analysis [55], we use the theory of common ground to inform our coding scheme for participants' questions and the Google Home's answers. Conversation analysis has long been used in HCI and CSCW as a method for understanding and designing support for human-machine interaction [19]. This approach enables us to identify the types of responses the Google Home provides that help participants to form a mental representation of the ways it is able to participate in a productive conversation with them and help them accomplish their goals for using the device. This mental representation that each human interlocutor forms about their conversation partner consists of mutual knowledge, beliefs, and assumptions that are verbally demonstrated through conversational grounding. The early interactions with the Google Home are especially important formative experiences for new users who don't have any experience with the system, because their understanding of its capabilities is based on these early conversations.

## 4.2 Data Preparation and Segmentation

The study sessions were recorded and transcribed. We also downloaded the logs of each participant's interactions with the Google Home from the Google account created for that participant, using

Google Takeout[8], the interface for downloading an archive of one's data from Google products. This means we had data about both what the participant said (from the transcript) and what the Google Home "heard" (from the logs). We combined the log data with the transcripts to create text files for coding which indicated places where the words spoken by the participant disagreed with the representation of those words in the log data. This is similar to the approach taken by Jiang et al. [25] in a 2013 study of the "Google voice search app" to determine speech recognition accuracy. We did this so that we could distinguish speech recognition failure from other types of challenges participants experienced when interacting with the Google Home. In 55% of the 100 tasks completed across all participants, the Google Home correctly recognized all of the participant's questions. On the remaining 45% of tasks, the Google Home was not able to correctly recognize the participant's speech on at least one question. The mean number of times per participant and task that their speech was not correctly recognized by the Google Home was 1.93 (*SD*=1.64, *Max*=7).

The second author then segmented the transcripts into "attempts." We define an attempt as a dyadic conversational exchange between the participant and the Google Home, consisting of the participant's question and the Google Home's response. Any thoughts the participant spoke before they asked a question were included in the attempt as well. If the participant followed the Google Home's response by saying "thanks" or "stop", that was not considered to be a new attempt. Also, if the Google Home asked a follow-up question, any answer by the participant was considered to be part of that same attempt. The first author then inspected the segmentation for all participants and tasks, and the two authors discussed other ways a small number of specific attempts could be segmented (less than one attempt per participant) in order come to an agreement on those particular instances. There were 752 attempts total across all participants and tasks. Participants used an average of 7.52 attempts per task (*SD*=4.95, *Min*=1, *Max*=31).

### 4.3 Qualitative Coding Process

We then engaged in three rounds of iterative, inductive qualitative coding. The authors met regularly to discuss the coding during the entire analysis process, and revised the codes together over a period of several months. In the first round of coding, the first author used a process coding approach to code each utterance made by the participant, within attempts (question-response pairs). Process coding is a first-cycle qualitative coding method which consists of using verb gerunds as codes to indicate actions in the data that change over time [48]. The goal of this round of coding was to identify what the participant appeared to be trying to accomplish with each question, based on the specific phrasing and words that were used as well as the context of the previous questions and responses within each task. In addition, we categorized the types of responses the Google Home provided to the participant's questions according to the relationship we perceived when analyzing the transcripts between the information the Google Home provided and the specifics of what the participant had been asking. We provide a description and examples of these categories in Section 5.1.

In the second round of coding, we focused on the relationship between the response the Google Home provided, and characteristics of the participant's subsequent question in the next attempt, which is directly influenced by Google Home's previous utterance. The second author took the lead in this coding round, and focused on coding for patterns in the Google Home response codes and participant process codes across attempts. This approach enabled us to connect the participant's reaction to what the Google Home had just said. These codes reflected whether the participant had asked their next question as though they believed the Google Home had understood their previous question or not.

[8] https://takeout.google.com/

In the final round of coding, both authors worked together to group the participant codes into two higher level categories: *Advancing* and *Backtracking*. The terms advancing and backtracking refer to whether the question the participant asked seemed to be moving the conversation closer to completing the task (advancing), or not (backtracking). *Advancing* questions indicate that the participant acted as though they believed they were making forward progress toward completing the task. *Backtracking* questions indicated that it seemed like the participant had interpreted the Google Home's response as indicating their question had been unsuccessful, and they needed to try something else to be able to find the information requested by the task. A small number of participant questions in each task did not fit into these two groups, such as the first question participants asked when starting the task which was typically rephrasing (and sometimes reading verbatim) the task description. We provide more detail and examples of Advancing and Backtracking questions in Section 5.2.

Our Advancing and Backtracking codes resemble concepts that are also present in structured annotation schemes for dialogue acts, such as the semantic annotation framework in ISO 24617-2[9]. However, our analysis was informed by common ground theory, not based on other frameworks. In addition, our purpose for conducting this analysis was not to produce an annotated corpus using standardized codes, but rather to connect the types of Google Home utterances we identified with evidence of how our participants responded to them.

## 5 CHARACTERISTICS OF THE CONVERSATIONS

### 5.1 Types of Google Home Responses

The most common response type the Google Home gave was a version of the statement, "Sorry, I'm not sure how to help." We labeled these responses *Cannot Help*. There were many variants: e.g., "Sorry, I don't know how to help with that yet, but I'm still learning"; "Sorry, I can't help with that yet"; etc., all similar in that the Google Home said it could not do something or didn't know something. This response type occurred in 304 out of 752 attempts (40%) across all tasks. Participants often blamed themselves when the Google Home said it could not respond to the question. For example, in this exchange between P01 (Task 1) and the Google Home, the participant explicitly stated that she felt like she wasn't asking in the correct way to get a helpful response:

> P01: Hey, Google. Next flight out from Lansing, Michigan to Japan.
> Google Home (GH)[10]: Sorry, I'm not sure how to help at that, but I'm still learning.
> P01: I feel like I am not asking the question correctly. Hey, Google. What is the next flight from Lansing, Michigan to Osaka, Japan?

The second type of response from the Google Home we labeled *Related*. These occurred in 175 out of 752 attempts (23%). What these responses have in common was that the Google Home both correctly recognized the participant's speech when they asked the question, and also provided a response that was related to what the participant asked. The relatedness was determined both by examining the content of the Google Home's response, and how the participant reacted to it. For example, when P06 (Task 2) asked about where to buy a new television, the Google Home responded exactly how he expected, and he signaled this by ending the task. This was a best-case example of a *Related* response:

> P06: What are the local stores that sell brand new televisions?

[9]https://www.iso.org/standard/51967.html, retrieved January 2020
[10]Some Google Home utterances, particularly where it recites a long list of options, have been shortened. This is noted with an ellipsis.

GH: I found a couple places within 3.4 miles. The first one is Cedar Street Electronics at 1630 South Cedar Street in Lansing. The second one is Best Buy at 2020 West Grand River Avenue in Okemos.
P06: Thanks, Google. (end of task)

The third type of Google Home response we saw in our transcripts we labeled *Unrelated*, which occurred in 178 out of 752 attempts (24%). In these responses, the Google Home has recognized the participant's speech and responded with an actual answer, but that answer was not perceived by the participant as being helpful in providing information that would allow them to complete the task. In some cases, the response was about the same topic as the task asked about, but at too high a level to be useful to the participants. In others, it was difficult for them to see a conceptual connection between their question and the response.

For instance, here are examples of both of these types of *Unrelated* responses. In the first, from P14, Task 5, the Google Home has clearly recognized the topic the participant was asking about (the U.S. National Basketball Association, or NBA). But, after it had responded, the participant said something about how the response was not what they were expecting, so she would try to ask another question:

P14: Hey, Google, do you know what the NBA is?
GH: Today, the Clippers will play the Magic at 7:00 PM. The Grizzlies will play the Wizards at 7:00 PM. The Thunder will play the Pacers at 7:00 PM. For other games, try asking about a particular team.
P14: Okay, so Google Home knows what the NBA is, but they didn't give me the definition I was hoping they would. They gave me the schedule. So, since the task said to ask about a specific team, I will ask about the Detroit Pistons. Hey, Google, can you tell me about the Detroit Pistons?

In the next excerpt, from P05 (Task 4), the Google Home interrupted the participant and gave a bizarre response to his question:

P05: Google. Hi, Google. Hey, Google, can you tell me how much–
GH: Hey, I'm having the craziest day. I just learned that bananas are curved because they grow towards the sun.
P05: Let me think. Google, can you tell me how much Obamacare costs?

We suspect that what happened here was that rather than saying the wake word, "Hey Google" or "OK Google", the participant started with "Hi Google" which triggered it to respond with an interesting fact about bananas. This exact situation happened to two other participants in our study as well, both of whom received this same response, word for word, from the Google Home. What the Google Home's responses in the previous two excerpts have in common is that they are *Unrelated* to the question the participant has asked.

The remaining 13% of responses (95 attempts) were instances where the Google Home incorrectly recognized the participant's speech, which we determined by comparing the transcript to the Google Home logs, and it therefore gave an answer that did not make sense given the question that was just asked. This group of responses also includes attempts where the Google Home gave a non-response that did not answer the participant's question, such as "I like to look up scores and stats," or asked the participant to repeat what they had said, e.g., "Sorry, what are you asking?"

## 5.2 Types of Participant Questions

*5.2.1 Advancing Toward the Goal.* The most common way that participants' questions indicated they thought they were advancing toward the goal of the task was by asking a question that was more specific and narrowly focused than the previous question, as if they were closing in on the answer they were looking for. For example, in this excerpt from P09, Task 1, the participant began by first asking a general question about hotels in Tokyo, and then after the Google Home listed specific hotels, followed up by asking for information about a hotel with specific characteristics. This indicates that the participant had received feedback from the Google Home that it understood the question and could provide information about hotels in Tokyo, and the participant should be able to ask a more specific question on that topic:

> P09: Hey, Google. Can you provide me a list of hotels in Tokyo?
> Google Home (GH): I found a few places. The first one is Sheraton Miyako Hotel Tokyo...
> P09: Okay, Google. Which hotel in Tokyo has the highest rating?
> GH: I've found a few top-rated places. The first one is Park Hotel Tokyo at Higashi Shinbashi...

Another way that participants' questions showed that they felt they were advancing toward the goal was by directly following up on something the Google Home had said, and using the same words, verbatim, as part of the subsequent question. Here, P01 (Task 1) asked a general question about attractions in Osaka, Japan. She then asked about a particular destination, Osaka Castle, that had been mentioned by the Google Home in its response:

> P01: Hey, Google. What other places of interest are there in Osaka Japan?
> GH: Here's a summary from the website Japan-top.com. Amongst Osaka's many attractions, these stand out: Dotonbori in Namba, and K-Comedy Theatre, Osaka Castle...
> P01: Hey, Google. How much does it cost to visit Osaka Castle?

Participants also indicated that they felt they were advancing towards completing the task by asking a new question that further developed a topic or concept that the Google Home had introduced. In this excerpt, P16 began Task 1 by asking a high-level question about places to visit in Japan. The Google Home provided information about cities in Japan, which confirmed to them the Google Home could provide information about the topic through a relevant next turn. So, she asked another question on the same high-level topic:

> P16: Okay, Google, what are good places to visit in Japan?
> GH: Destinations in Japan include Tokyo, Kyoto, Osaka, and other locations.
> P16: Okay Google, what is the weather going to be like in Japan for the next two weeks?

*5.2.2 Backtracking.* After some of the Google Home's responses, the questions participants asked indicated they thought they had reached a dead end and needed to start over from the task description, or change the wording of their question to try to get a different response. For example, one obvious clue that this had happened was when the participant asked a question which repeated the information from the task description. Task 5 asked participants to find out whether it was likely that a nearby NBA basketball team would make it to the playoffs this year. P09 began the task by asking, "Okay, Google. What are the Detroit Pistons' chances of making the NBA playoffs?" After asking a few more questions and receiving only general information about the NBA in response, he asked a slight rephrasing of same question that he had initially asked. Many participants had similar interaction patterns to the this one, on at least one of the tasks:

> P09: Okay, Google. Do you think the Detroit Pistons will make the playoffs?
> GH: Sorry, I don't know how to help with that.

Sometimes, participants asked a different version of the same question by rearranging the order of the clauses, or by changing their vocabulary slightly, as P11 does below on Task 3:

> P11: Okay, Google, what is the price associated with receiving a passport? *GH heard: "what is the price of associated with"*

GH: Associated Bank Corp shares are trading at $25.07 a share on NYSE as of 11:07 AM Eastern Standard Time today, up 0.5% from yesterday.
P11: Okay, Google, stop.
P11: Okay, Google, how much does it cost to issue a passport?

In the above excerpt, the Google Home responded to only part of the question the participant spoke, with an answer that was clearly not related to the question. This unrelated response was feedback from the Google Home which indicated that the P11's previous question had been unsuccessful, and she responded by asking again in a different way.

A third way that participants demonstrated that they had interpreted the Google Home's response as feedback that their question was unsuccessful was by asking a broader version of the question they had just asked. Here, P10 (Task 5) was trying to find out the betting odds for the Detroit Pistons to make the playoffs. But, the Google Home responded that it was not able to tell her this. She revised her question from asking about the odds in Las Vegas, to asking a very broad, high level question about the team in general:

P10: Hey, Google. What does Las Vegas say the odds are for the the Detroit Pistons basketball game to win the NBA playoffs?
GH: Sorry. I don't understand.
P10: Okay, thanks, Google. Well, that was frustrating because they should know what Las Vegas odds are, but what other way could I do this. Hmm. Let me try something else.
P10: Hey, Google. Can you predict Detroit... *GH heard: "Can you predict"*
GH: Sorry. I can't help with that yet.
P10: Hey, Google. What do you know about the Detroit Pistons?
GH: Yesterday, the Detroit Pistons played the Celtics. The final score was 91 to 81, Celtics.

### 5.3 In Scope versus Out of Scope Tasks

We had expected that participants would make fewer attempts on In Scope tasks than Out of Scope tasks, because In Scope tasks were consistent with the advertised features of the Google Home and therefore should be easier to complete. However, Out of Scope tasks (*Min*=2, *Median*=6, *Max*=18) had slightly fewer attempts per task than In Scope tasks (*Min*=1, *Median*=7, *Max*=31). In addition, In Scope tasks had more attempts that advanced toward the goal (276 attempts) than backtracked (169 attempts). The opposite was true for Out of Scope tasks: participants asked questions that backtracked (198 attempts) more often than advanced (75 attempts). We interpret this to mean that participants gave up sooner on Out of Scope tasks than In Scope tasks. We did not analyze the success or failure of each task, because the goal of this study was not to evaluate the usability or accuracy of how the Google Home supports information seeking.

There were also qualitative differences between how In Scope and Out of Scope tasks unfolded for participants. The two excerpts below illustrate this. In the first excerpt, from an In Scope task (P08, Task 2), the Google Home not only demonstrated that it understood the question by giving a *Related* response (coded as *Related* due both to the topical connection to the question and the participant's reaction to the suggestion), it also offered to answer a similar question instead:

P08: Hey, Google, what type of television should I buy?
GH: I'm not sure, but I can tell you the answer to a similar question. What type of TV is best? Would that help?
P08: Yes.
GH: On the website ConsumerReports.org, they say...

In contrast, in this excerpt from an Out of Scope task (P13, Task 5), the Google Home gave two *Cannot Help* responses to P13's questions:

> P13: Okay, Google, what are your predictions for the Detroit Pistons in the NBA playoffs this season?
> GH: Sorry, I'm not sure how to help.
> P13: Okay. So, ask a question not directly predicting, but, Okay, Google, what are the stats for the Detroit Pistons this season?
> GH: Sorry, I don't know how to help with that, but I'm learning more every day.

## 6 FEEDBACK AND PERCEPTIONS OF PROGRESS

We focus in this section on how participants followed up Google Home responses by asking questions which showed they believed they were Advancing toward the goal of the task, or Backtracking away from it. First, we describe differences in the patterns of interaction on In Scope versus Out of Scope tasks. Then, we describe how participants reacted to each of the Google Home response types. Finally, we show that even a response that is *Unrelated* to the participant's question provides more feedback and about what the Google Home is capable of, and a better opportunity for grounding, than a *Cannot Help* response like, "Sorry, I can't help with that."

There is little prior work focusing on conversation analysis of interactions with general purpose digital assistants, like the Google Home, upon which we might base predictions about how common different types of questions and responses might be. Therefore, this research is exploratory, not confirmatory. We present some code frequencies in this section; these are included to give the reader a higher-level characterization of what conversations between our participants and the Google Home were like, in addition to our detailed qualitative description of specific interactions. Our findings illustrate the patterns we observed, in the hope that this paper can be used as a baseline for future work.

To identify patterns in participants' reactions to the Google Home's responses, our unit of analysis is adjacent pairs of utterances that cross attempts: the response from the Google Home, and the next question asked by the participant. We assume, based on the theory of common ground [11], that the participant's question immediately following a Google Home response is a reaction to that response. The only way the Google Home is capable of signaling understanding is by responding with a "relevant next turn"—an utterance that is an appropriate response given what the participant just asked. The Google Home can't provide other kinds of evidence that are common in human-to-human conversation like facial expressions, nor was it capable of rephrasing the participant's question to check understanding at the time the study took place.

### 6.1 Cannot Help Responses are Unclear

If the participant understands the Google Home's response to be a relevant next turn, this presents feedback to the participant that they and the Google Home are in agreement about what the conversation is about, and the conversation can proceed. However, after a *Cannot Help* response from the Google Home, participants' next question most often backtracked. Here's an example from P02, Task 4, where the Google Home responded with a *Cannot Help* response, and the participant subsequently repeated and rephrased the question by using slightly different vocabulary:

> P02: Okay. Let's see. Okay, Google, who should apply for Obamacare?
> GH: Sorry, I'm not sure how to help with that.
> P02: Okay, Google. Who can apply for Obamacare?

In this example, the participant received the *Cannot Help* response, which did not provide any information that the participant could use to formulate a new question. The only information the participant had was that asking "who should apply" was not effective. So she backtracked, and slightly modified the wording of the question in the hope that it would be more successful.

In another example, below, P15 (Task 3) has just received a response from the Google Home about the forms and other information required to apply for a passport in the U.S. This, however, was not the information she was looking for. She then asked "where to go" in two very similar ways, using slightly different vocabulary, to which the Google Home answered with a *Cannot Help* response both times. This indicates that she had not received enough feedback from the Google Home to be able to ask a different question and make progress on the task:

> P15: Okay, Google, that's great, but I'd like to know where to go to apply for a passport?
> GH: Sorry, I'm not sure how to help, but I'm learning more every day.
> P15: Okay, Google, I know how to get prepared and what to bring to get the passport, but I live in [city, state][11], and I'd like to know where to go get a passport. *[pause]*
> P15: Okay, Google, where can I go to have a United States passport issued to me?
> GH: Sorry, I'm not sure how to help with that.

Figure 1a shows that for Out of Scope tasks, *Cannot Help* responses were followed by a participant question that backtracked by broadening, rephrasing, or falling back on the task description 104 times (76% of participant questions which followed a *Cannot Help* response on Out of Scope tasks), and only advanced the conversation toward the goal 32 times (24%). However, for In Scope tasks, advancing versus backtracking in the next question was about equally common: 73 forward attempts and 69 backward attempts.

*Cannot Help* responses, like "Sorry, I'm not sure how to help with that" convey objective information that an answer is not forthcoming from the Google Home. But, while this response led participants to rephrase or broaden what they asked, it didn't provide clues about how to do so effectively. With this response type, the Google Home has not signaled what aspects of the question make it difficult to answer. There's no way for the participant to know if the entire topic is causing the difficulty, or if is it the way they have asked the question that is the source of the problem. Therefore, they can't effectively guess what they should try next.

There were a few instances where participants asked a question following a *Cannot Help* response that was more specific and narrow than the previous question, and appeared to be advancing toward the goal. For example, in this excerpt from P17 (Task 5), the question following the *Cannot Help* response is more specific than the previous question about which season he is referring to:

> P17: Hey, Google, what is your prediction of the Detroit Pistons going to the playoffs?
> GH: Sorry, I'm not sure how to help.
> P17: Hey, Google, what is your prediction of the Detroit Pistons going to the playoffs this season?

In contrast to the previous example from P15, rather than asking a small variation on the same question, P17 has asked a narrower follow-up question about the specific timing of the playoffs. This increased specificity is an example of experimentation by P17, trying to work out how to communicate with the Google Home based on his own knowledge and his previous experience on the other tasks (he completed this task fourth out of five). But overall, in our study sessions, *Cannot Help* responses were too unclear to be able to help the participants to become aware of what kinds of answers the Google Home could provide. In other words, *Cannot Help* responses did not provide an opening for further grounding attempts by the participant that would help them learn about the Google Home's capabilities.

[11] Information in brackets has been redacted to protect participant identity.

(a) Next question after a *Cannot Help* Response

| | In Scope | Out of Scope |
|---|---|---|
| Advancing | 73 (51%) | 32 (24%) |
| Backtracking | 69 (49%) | 104 (76%) |

(b) Next question after a *Related* Response

| | In Scope | Out of Scope |
|---|---|---|
| Advancing | 85 (77%) | 17 (33%) |
| Backtracking | 25 (23%) | 34 (67%) |

(c) Next question after an *Unrelated* Response

| | In Scope | Out of Scope |
|---|---|---|
| Advancing | 75 (59%) | 19 (46%) |
| Backtracking | 53 (41%) | 22 (54%) |

Fig. 1. This figure shows the number and percent of times when participants' next questions advanced or backtracked following each type of Google Home response, for tasks that were In Scope versus Out of Scope in relation to the Google Home's capabilities. The unit of analysis is adjacent pairs of utterances that cross attempts: the response from the Google Home, and the next question asked by the participant. The percentages in each cell are calculated by column within each subfigure; for example, in subfigure (a) the percents for the In Scope column sum to 100%. A darker background means a higher percent.

### 6.2 Related Responses Signal Progress

After *Related* responses from the Google Home, participants asked questions that indicated they felt like the Google Home understood them more often for In Scope tasks (85 attempts, 77% of In Scope tasks) than for Out of Scope tasks (17 attempts, 33% of In Scope tasks). (See Figure 1b for a breakdown of *Related* responses by task type and participant question type.) For example, in the excerpt below, P10 (Task 3) asked two questions in a row on an In Scope task that received responses from the Google Home that she understood to be relevant next turns by the participant:

> P10: Hey, Google. What's the closest place to Okemos, Michigan to get a passport to go overseas?
> GH: I found a few places within 3.4 miles. The first one is United States Postal Service at at 2025 Central Park Dr. in Okemos...
> P10: Hey, Google. What do I need to get a passport?
> GH: On the website usa.gov, they say, "What you'll need to provide: One, form DS11. Two, proof of US citizenship... To read more, look for the link in your Google Home app.
> P10: Thanks, Google. I feel pretty good. I don't know what that form is, but she told me the link to find out about it.

In this sequence, the Google Home first provided a location that seemed to satisfy the participant as a close place where one can get a passport, which presented feedback that it could help with the task. Then, it also provided information about what documents are necessary. The participant felt confident enough that the Google Home had helped her that she trusted that if she looked at the link the Google Home referred to, it would prove to be helpful as well. This illustrates a conversational exchange in which the participant consistently asked questions and received responses that allowed the participant to feel like she had found the information the task had directed her to find.

*Related* responses even sometimes allowed participants to make progress on Out of Scope tasks. Here, participant P08 began his first attempt on Task 5 not by verbalizing the task description, but by asking a narrowly focused question about the specific local NBA team the task directed him to find information about. When he received a *Related* response that contained the factual answer he had been looking for, he asked another straightforward factual question and received another *Related* response—two attempts in a row with helpful feedback from the Google Home:

> P08: Okay, Google, when do[es] the Pistons play next?
> GH: The Pistons' next game is today at 8:00 PM, where they will be playing the Bucks.
> P08: The Bucks. That seems like a good match up. I wonder what both of their records are. Hey, Google, what's the Pistons' record this year?
> GH: They're in fourth place in the Eastern Conference with 14 wins and nine losses, six games behind the Celtics.

However, there were a few instances where the participant backtracked after receiving what looks from the transcript like a relevant next turn from the Google Home. This excerpt shows one such example (P06, Task 1). This participant started by asking about the "cost of a hotel" in Japan, on two sequential attempts, to which the Google Home provided a *Cannot Help* response both times. Note that he initially did not include the word "room" in his question:

> P06: Okay, Google. Can you give me an idea of what the cost of a hotel there in Japan might be?
> GH: Sorry. I'm not sure how to help with that, but I'm trying to learn.
> P06: Okay. Okay, Google. I'd like to get an idea of what the cost is of...I'm gonna ask the same question here again, we'll see if the answers are different. What are the cost of hotels in Japan? *GH heard: "I'd like to get an idea of what the cost is of I'm going to ask the same question"*
> GH: Sorry. I'm not sure how to help, but I'm still learning.

Then, on his third try, he rephrased his question from "cost of a hotel" to "costs of hotels". The Google Home gave a *Related* response:

> P06: Okay, Google. What are the costs of hotels in Japan?
> GH: On the website Japan-guide.com, they say rates range from less than 2,000 yen per person in a dormitory, to over 50,000 yen per person in a first class hotel or ryokan.

But, the participant's next question shows that he did not feel that the response answered his question. Our Backtracking code included instances where the participant changed the wording of the question to try to get a different response. Below, the transcript excerpt shows that the participant used both the word "apartment" and phrase "hotel room" in his fourth attempt, instead of "costs of hotels":

> P06: Okay, Google. Say I wanted to get an apartment... Not an apartment, but get a hotel room. Can you give me some idea of what the average price might be?

The audio recording of this question shows that P06 placed additional vocal emphasis on the word "room", indicating the importance of that word in his utterance and that he wanted his interlocutor to pay special attention to the clarification. This indicates that the response the Google Home provided was not what he was expecting—the average nightly cost for a hotel—and he felt it necessary to try again in a slightly different way. Despite this reaction from the participant, we coded the Google Home's utterance beginning with "On the website Japan-guide.com, they say..." as a *Related* Google Home response, because it does actually contain an answer to the question the participant had previously asked.

### 6.3 Unrelated Responses Can Actually be Helpful

The third and final category of Google Home responses, *Unrelated*, seems like it should be even more ambiguous and potentially difficult for participants to make sense of than the *Cannot Help* responses, because *Cannot Help* is a clear, direct response that the Google Home is unable to answer the question. However, our analysis found that the *Unrelated* responses actually helped participants move forward with In Scope tasks more often than backtracking. As Figure 1c shows, there were more Advancing attempts (75) than Backtracking attempts (53) for In Scope tasks, and for Out of Scope tasks participants moved forward and backward about equally (19 vs. 22 attempts). Participants most often followed an *Unrelated* response by asking a more specific question. For example, here's an exchange between P20 and the Google Home on Task 3. P20 articulated more specifically the information he was looking for, after an answer from the Google Home that was on topic but not what he wanted to know about:

> P20: Hey Google, how do I get a passport?
> GH: Here's a summary from the website, usps.com. Renew a passport by mail. One, step one, apply early. Two, step two, check your renewal eligibility and three...
> P20: Okay, Google. That wasn't what I was looking for. That was telling me how to renew. I need to find a new one. Hey Google, how do I apply for a new US passport?

In the above conversation, the Google Home's response was about passports, but the participant wanted to know how to get a new passport, not how to renew one. This response was not helpful—it answered a question that the participant felt he did not ask. The participant then made a statement clarifying what he wanted, "That wasn't what I was looking for. That was telling me how to renew. I need to find a new one." But, in order for the Google Home to answer, P20's clarifying statement must be phrased as a new question that uses the wake word. The fact that the participant thought about the response and then asked a more specific question indicates that even when a response from the Google Home signals a misunderstanding, an *Unrelated* response still supported continued conversation.

Below is another example of an unrelated response. P19 began Task 2 by verbalizing part of the task description, and the Google Home responded by asking the participant where he was located. However, P19 thought that he should be able to speak the location rather than having to use an app to change it. When he used the wake word and said the city and state where he was located, he believed he was still following up with the Google Home's previous request to specify the location for a search for a television set. But when he spoke the location, the Google Home gave a response that was out of context for what he was expecting:

> P19: Google, where can I buy a television set?
> GH: Please specify a location, so I can search stores near you. You can do this in your settings by turning on location.
> P19: East Lansing, Michigan.
> P19: Okay, Google. I live in [city, state].
> GH: I found a few places. The first one is Michigan State University at 220 Trowbridge Road in East Lansing. The second one is Independence Village of East Lansing at 2530 Marfitt Road in East Lansing.
> P19: Okay, so it's telling me places to live. That didn't quite work.

The statement at the end of the above excerpt shows that P19 has made an interpretation of the Google Home's response that connected his previous "I live in..." statement with the list of addresses provided by the Google Home. He spoke aloud that interpretation: "it's telling me places to live". If the Google Home had given a *Cannot Help* response, P19 would not have been able to make that interpretation, and would not have learned as much about the Google Home's capabilities. P19 then continued the task, and his questions indicated that he had learned that he must be specific about the location in each question. But, the Google Home again provided the same *Unrelated* response, which prompted P19 to ask a more specific question about the location of electronics stores:

> P19: Okay, Google. Where can I buy a television in East Lansing, Michigan?
> GH: Please specify a location, so I can search stores near you. You can do this in your settings by turning on location.
> P19: Okay, Google. Are there electronic stores in East Lansing, Michigan? *GH heard: "are there electronic stores"*
> GH: I found a few places. The first one is MSU Tech Store at 450 Auditorium Road number 110 in East Lansing. The second is Mac Men at 2660 East Grand River Avenue number 5 in East Lansing. The third one is Staples at 3003 East Michigan Avenue in Lansing.
> P19: Probably don't have TVs. Okay, Google, what are the hours for Best Buy in Okemos, Michigan?
> GH: Best Buy at 2020 West Grand River Avenue is open today from 8:00 AM to 11:00 PM.

Unfortunately, the Google Home failed to recognize the participant's speech correctly—it missed the location P19 specified after "electronic stores". But, in the end, the stores it listed were in the local area anyway, and the list helped the participant to feel he had completed the task. This example shows how a series of *Unrelated* responses can provide enough feedback to help the participant continue the conversation. However, in a small number of cases, *Unrelated* responses were truly unhelpful (P14, Task 2):

> P14: Hey, Google, what store should I go to, to purchase a new TV?
> GH: Sorry, What's your return date?
> P14: Okay, so somehow Google is still caught on, No, Google, no travel, TV.
> GH: I missed it again. Try giving me the date, like January 12th.
> P14: Hey, Google, I wanna buy a TV.
> GH: I'm having some trouble, please try again later.
> P14: I think I broke Google Home.
> GH: Sorry, I'm not sure how to help.
> P14: Google, stop. Okay.

These first and last *Unrelated* examples in this section, from P20 (on page 19) and P14 (above), are an interesting comparison. P20 was able to reach an answer that was satisfactory to him in just one additional attempt after the end of the excerpt above. P14, on the other hand, nearly gave up on the task after the exchange in the excerpt. But she continued, the Google Home recovered from whatever problem it was having, and P14 ultimately felt satisfied with the final answer she obtained.

That P20 and P14 both felt like they were able to complete the task is partially due to the fact that both of these tasks were In Scope tasks; participants tended to stop asking questions in Out of Scope tasks sooner and without feeling like they had found a suitable answer. But, there are many examples in our data, even on Out of Scope tasks, where receiving any response besides *Cannot Help* gave participants enough feedback that the Google Home had recognized something they said and responded to it that they kept going. This encouraged them to keep trying, and suggests that even *Unrelated* responses can allow the conversation to progress and further experimentation to take place.

## 7 DISCUSSION

Our findings showed that *Unrelated* responses affected subsequent utterances in a surprising way: in many cases, they allowed the conversation to continue. These responses provided unexpected, but beneficial feedback to support the grounding process, which we argue helped participants to formulate their next question. In other words, uncertainty about the response provided participants an opportunity to try different ways to make progress toward finding the information they were asking about. This extended the interaction, and provided more opportunities to learn about how to best use the system. Based on this insight, we discuss future directions for helping users extend their conversations with a digital assistant while learning about the system's capabilities, and offer design implications that can serve as initial guidance on the requirements for achieving symbiosis.

### 7.1 Towards a Symbiotic Relationship with Digital Assistants

As AI-based systems, including digital assistants, are becoming increasingly common and taking on more roles in people's everyday lives, a need to reconsider symbiosis between humans and AI—as partners rather than a user and a tool—has been recognized [20, 53]. To achieve a successful symbiotic relationship with a digital assistant system, "one's learning needs to have a positive influence on the other's learning" [42, p. 135]. Specifically, users' continuous attempts to learn about what digital assistants can do is important for a digital assistant to learn how to improve its functionality in the longer term. Additionally, generating more training data from users to improve the machine learning models that underlie system performance is beneficial for usability. Each utterance in the interaction between a user and a digital assistant should provide learning opportunities that support partnership. With accumulated learning experiences, users would be able to continue exploring digital assistants' capabilities beyond the use of a small number of features. To support continuous learning practices for mutually beneficial partnership between humans and digital assistants, we argue that the design objective for digital assistants should not always be providing the best answer as quickly as possible solely to achieve a system's goal, but also to optimize the models for new opportunities for grounding.

In contrast to earlier work which more heavily emphasized the theory of conversational grounding, recent research and development on voice user interface systems uses a data-driven approach to optimize performance accuracy and efficiency: minimizing errors, and providing precise answers or indicating failure quickly [9, 23, 37, 38, 56]. In this approach, conceptualizing the system's utterances as responses or answers emphasizes evaluating them according to how relevant they are to what has been requested by the user. Current general purpose digital assistants also aim to identify the conversation's objective quickly and accurately, relying on machine learning techniques to provide the most relevant response by matching the user's utterance to the best intent in the natural language understanding module (i.e., intent classification[12]).

AI systems have impressive computational skills, but poor social skills [42]. The inability of machines to fully understand situational context is the main impediment to generating appropriate responses to users' actions in human-machine communications [54]. People can easily update their mutual knowledge, beliefs, and assumptions when misunderstandings occur in conversation, rather than responding with the most accurate response to the best of their knowledge at once. Despite advanced machine learning techniques, on the other hand, digital assistants still have limited

[12] https://cloud.google.com/dialogflow/docs/basics?hl=en, retrieved January 2020

abilities to recognize whether a misunderstanding happened in previous turns [35]. Therefore, digital assistants cannot recognize whether they need to update common ground to resolve any mistakes, unless the user's utterance is classified as a particular intent designed to ask follow-up questions (i.e., slot-filling[13]). However, this functionality is only programmed to receive a limited range of predetermined utterances (i.e., a required parameter) from users in return, rather than to coordinate contexts in order to resolve misunderstandings. Due to this inability in grounding skills, repair strategies by using data-driven approach [23, 37, 56] may have limitations in supporting for long-term learning experiences of the system, although they may temporarily resolve communication breakdowns. Furthermore, most skills, including the Google Home's basic features[14] and additional ones created by external developers, are designed to be operated by following procedural instructions or using specific keywords. As digital assistants are designed to be able to respond to a predetermined set of accurately detected utterances, many answers from the Google Home may preclude learning opportunities for both users and the system.

Our findings demonstrated that the accuracy of intent classification may actually occasionally be harmful, such as when shorter interactions about what the system cannot do preclude future interaction. Essentially, a *Cannot Help* response from a digital assistant is functionally equivalent to a popup error message in a visual user interface where the only option the user has is to click "OK". It conveys the information that the system has reached a state where it is unable to proceed. This type of response conveys that the conversation has reached an impasse, which may allow the user to quickly begin another attempt [15]. However, at the same time it provides no feedback or mechanism guide the user's next attempt, and fails to build the common ground necessary to make progress in the conversation. A *Cannot Help* response from a digital assistant may be the most accurate response from the system's perspective, but it is unhelpful for the user, and makes it more difficult and unpleasant for users to learn about the Google Home's capabilities by trial and error. In providing *Cannot Help* responses, digital assistants may be teaching users about the kinds of future questions they should *not* ask and future requests they should *not* make, leading to shorter conversations and reduced usage.

Sengers and Gaver [51] suggested that a "single correct interpretation" in designing and evaluating technology is not always needed since there are various use cases of technology in everyday life. Given the fact that general purpose digital assistants are situated in everyday settings beyond circumscribed task-oriented use cases, a single utterance from a user can be subject to various interpretations in everyday use. This means that users' utterances in conversation with a digital assistant should not be understood to have a "single correct interpretation" [51]; however, this is what digital assistants do by optimizing for the accuracy of each response. The questions from users that elicit *Cannot Help* responses expose gaps or edge cases in the machine learning models. These gaps are unavoidable, since it is unlikely that a data-driven model can encompass an effective and efficient response to every possible question that might be asked. However, successful partnering with human users can supplement these weak points in AI systems [20]. Conversational grounding is a communication process that is specialized to recover from "edge cases" that arise in human conversation—grounding supports negotiation of meaning and understanding when gaps exist. To complement the current data-driven approach toward developing digital assistants, designers should incorporate conversational grounding as a fundamental principle underlying how symbiotic partnerships can be successfully achieved.

We do not argue against the importance of accuracy and efficiency in a digital assistant's responses. Rather, we argue that the goal for optimization of digital assistants needs to focus on

[13] https://cloud.google.com/dialogflow/docs/intents-actions-parameters, retrieved January 2020
[14] https://support.google.com/googlenest/topic/7195017 as of January 2020

how a mutual learning process can be fostered, for the competence of both interlocutors. The ability of the human interlocutor to do conversational grounding, even when their conversation partner is a machine, can supplement the weakness of current commercial digital assistants and accommodate contextual details and nuance in the interaction. Supporting the grounding process is not just important for successful task completion, but also for figuring out how to create mutually beneficial relationships between people and digital assistants that allows them to adapt to one another.

### 7.2 Design Implications

Conversation consists of collective actions on the parts of the people who are talking to one another. However, conversation between participants and the Google Home does not yet rise to the level of collective action, as responses from the Google Home often cut conversations short. Our findings suggest that today's digital assistants may treat the interaction as input/output data processing to accomplish the system's goals (e.g., recognizing an intent in the voice data, providing the most relevant response) rather than as a collective action to improve learning opportunities for both the user and the digital assistant. In order to not only improve usability but also create a mutually beneficial relationship between the user and the system, we present two alternative perspectives that address the importance of the grounding process. Rather than proposing specific technical implementation approaches, we reflect on the nature of human conversation and present implications for supporting continued experimentation for achieving symbiosis.

*7.2.1 From Response-as-Output to Response-as-Evidence.* Human-to-human conversation is not always perfect, and each interlocutor unconsciously expects that their conversation partner will provide evidence of whether their utterance is understood well, or needs to be fixed or clarified [13]. Considering that the responses from the Google Home are the only mechanism for providing information to the user about what kinds of interactions with the system are likely to be successful, these responses must provide clues about the underlying algorithmic processes that produce the system's responses. This is because more exposure to a system's inner workings—in this case, a representation of its 'thought process'—helps people to conjecture about the Google Home's ability to understand their utterances, resulting in them formulating a better subsequent question. Porcheron et al. suggest a similar perspective: responses from digital assistants should be considered as resources that people use in order to identify the state of the system for further interaction [45]. In order to facilitate the learnability of voice interactions, responses should be viewed as evidence to coordinate a conversational process where people can keep exploring a system's functionality and abilities for the long term.

Therefore, a response from a digital assistant should be regarded as *evidence* for grounding rather than *output* optimized for a single utterance's accuracy. A general purpose digital assistant today is not capable of responding accurately to every question a user asks, but it may be capable of promoting a more extended dialogue in which the user adjusts their objective, adapting to the digital assistant as they learn more about it. For instance, participants in our study were able to make a progress after receiving *Unrelated* responses, even though this type of response was not perfectly matched by the participants' intention. We believe the *Unrelated* responses function as *negative evidence* to help participants discern whether their utterance was understood properly, and if it was not, which part wasn't clear enough for the assistant to understand. By promoting common ground with users (e.g., mutual knowledge, beliefs, and assumptions), this type of response may not just facilitate learning about how to reformulate questions to get desired information, but also support continuous exploration of new features. On the other hand, a *Cannot Help* response can be neither positive nor negative evidence for grounding. In the case of digital assistants, conversational grounding is not just important for supporting recovery from communication breakdowns, but also for scaffolding a symbiotic relationship by facilitating a longer interaction.

To generate evidence for grounding, digital assistants should provide guidance about which aspects in the sentence (e.g., lexical items) need to be added or changed to resolve miscommunication, even if the resulting response does not cleanly match the question the user asked. Even though the 'confidence score'[15] may not be high enough to match between the intent and users' question, the digital assistant should try to provide a response to show not only its continued attention, but also evidence of how and why miscommunication occurred. A response also can help users discover what digital assistants can or cannot do. However, incapability should not be shown as merely providing a *Cannot Help* response that results in a dead-end conversation. Rather, the response should provide another possibility by which users can continue to seek out other related information and accomplish their conversational goals, while also noting the system's incapability (e.g., P14, Task 5 in Section 5.1).

Despite the importance of evidence in the interaction, prioritizing evidence over accuracy does not necessarily apply in all situations. Designers should consider the tradeoff between efficiency and learning: an *Unrelated* response can improve learnability, but it can also be a confusing distraction if the grounding process takes longer than users expect. On the other hand, shorter interactions may be more efficient but both users and systems may learn less from them. Future work should investigate how responses can be designed to moderate the need for more evidence for grounding depending on the context, purpose, or topic of conversation.

*7.2.2 The User as a Conversation Partner, not a Bystander.* In order to better support a symbiotic relationship between people and digital assistants, we also argue that users should be considered as conversation partners who coordinate the process and content of a conversation rather than bystanders who receive responses from the assistant, regardless of what those responses contain. The optimal scenario for future development of voice user interfaces is that digital assistants become better at conversational grounding. However, given the lack of grounding skill in current commercial digital assistants, simply providing more opportunities for humans themselves to update engage in grounding has the potential to supplement shortcomings in the system's performance and help the system become better in the process. Providing evidence through an *Unrelated* response does not require the system to improve its capacity for grounding. Instead, it enables the human interlocutor to do what they automatically do in conversation with a human: learn the bounds of what they can successfully communicate about.

Kuijer and Giaccardi [28] introduced the concept of co-performance as a new perspective on the relationship between AI system and humans. They call for AI systems that are able to learn something through interplay with humans, to establish meaningful future partnerships rather than only autonomously performing tasks. Extending the idea of co-performance to digital assistants, we envision the symbiosis between humans and digital assistants as a partnership that arises out of automatic communication processes on the part of the human conversational partner, and a shift in optimization on the part of the digital assistant in selecting responses that can serve as evidence for grounding. In this way, the system can elicit more data from users, unveil its capabilities to help users "understand what the AI system is capable of doing" [1, p. 3], and form appropriate expectations that support learnability [31].

[15] Confidence score: The system's confidence level that this knowledge answer is a good match for this conversational query. retrieved from https://cloud.google.com/dialogflow/docs/reference/rest/v2beta1/DetectIntentResponse, January 2020

## 8 CONCLUSION

Licklider envisioned man-machine symbiosis in 1960, which he described as working together with a computer as closely as one would with a "colleague whose competence supplements your own" [29, p. 5]. Many AI systems, including digital assistants, have the potential to build a mutually beneficial partnership with people through providing support for learning by the system and by the user. Our findings showed that when the system encounters a question it cannot answer, the process of conversational grounding can bridge the gap between the system and the human user, as long as the system provides an opening for grounding to occur in the form of feedback about the ambiguity of the system's ability to respond in that moment. This opening for grounding to happen is an opportunity to promote exploration and support learnability for both partners. Therefore, we argue that machine learning models in digital assistants should be optimized not just for accuracy, but also for new opportunities for grounding. By combining advanced data-driven techniques with support for human conversation partners to form common ground about digital assistants, we argue that the symbiotic relationship can be achieved. These findings and implications add to the growing body of knowledge on how to design meaningful human-AI interaction, and provide guidance to designers of digital assistants for exploring new ways to build symbiotic relationships between users and AI systems.

## ACKNOWLEDGMENTS

We thank the BITLab @ MSU and TMILab @ CU research groups for helpful discussions and feedback, and the Michigan State University Graduate School and the Department of Media and Information for providing a Graduate Office Fellowship to support this research.

## REFERENCES

[1] Saleema Amershi, Saleema Weld, DaAmershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[2] Tawfiq Ammari, Jofish Kaye, Janice Y. Tsai, and Frank Bentley. 2019. Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Trans. Comput.-Hum. Interact.* 26, 3 (Apr. 2019). https://doi.org/10.1145/3311956

[3] Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300484

[4] Erin Beneteau, Olivia K Richards, Mingrui Zhang, Julie A Kientz, Jason Yip, and Alexis Hiniker. 2019. Communication Breakdowns Between Families and Alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300473

[5] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the Long-Term Use of Smart Speaker Assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3 (Sep. 2018). https://doi.org/10.1145/3264901

[6] Dan Bohus and Alexander I. Rudnicky. 2005. Sorry, I didn't catch that! An investigation of non-understanding errors and recovery strategies. In *6th SIGdial Workshop on Discourse and Dialogue*. 128–143. https://www.isca-speech.org/archive_open/sigdial6/sgd6_128.html

[7] S.E. Brennan. 1991. Conversation With and Through Computers. *User Modeling and User-Adapted Interaction* 1 (1991), 67–86. https://doi.org/10.1007/BF00158952

[8] Susan E Brennan. 1998. The grounding problem in conversations with and through computers. *Social and cognitive approaches to interpersonal communication* (1998), 201–225.

[9] Yun-Nung Chen, Asli Celikyilmaz, and Dilek Hakkani-Tür. 2017. Deep Learning for Dialogue Systems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-Tutorial Abstracts,*. Association for Computational Linguistics, Vancouver, Canada, 8–14. https://www.aclweb.org/anthology/P17-5004

[10] Minji Cho, Sang-su Lee, and Kun-Pyo Lee. 2019. Once a Kind Friend is Now a Thing: Understanding How Conversational Agents at Home are Forgotten. In *Proceedings of the 2019 on Designing Interactive Systems Conference (DIS '19)*.

Association for Computing Machinery, New York, NY, USA, 1557–1569. https://doi.org/10.1145/3322276.3322332
[11] Herbert H. Clark. 1996. *Using Language*. Cambridge University Press. https://doi.org/10.1017/CBO9780511620539
[12] Herbert H Clark and Susan E Brennan. 1991. Grounding in Communication. (1991), 127–149. https://doi.org/10.1037/10096-006
[13] Herbert H Clark and Edward F Schaefer. 1989. Contributing to Discourse. *Cognitive Science* 13, 2 (1989), 259–294. https://doi.org/10.1016/0364-0213(89)90008-6
[14] Eric Enge. 2017. Rating the Smarts of the Digital Personal Assistants. https://blogs.perficientdigital.com/2017/04/27/1-rating-the-smarts-of-the-digital-personal-assistants/
[15] Joel E. Fischer, Stuart Reeves, Martin Porcheron, and Rein Ove Sikveland. 2019. Progressivity for Voice Interface Design. In *Proceedings of the 1st International Conference on Conversational User Interfaces (CUI '19)*. Association for Computing Machinery, New York, NY, USA, Article 26, 8 pages. https://doi.org/10.1145/3342775.3342788
[16] Susan R Fussell and Robert M Krauss. 1989. The effects of intended audience on message production and comprehension: Reference in a common ground framework. *Journal of experimental social psychology* 25, 3 (1989), 203–219.
[17] Radhika Garg and Christopher Moreno. 2019. Understanding Motivators, Constraints, and Practices of Sharing Internet of Things. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 2, Article 44 (2019), 21 pages. https://doi.org/10.1145/3328915
[18] Radhika Garg and Subhasree Sengupta. 2019. "When You Can Do It, Why Can't I?": Racial and Socioeconomic Differences in Family Technology Use and Non-Use. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 63 (2019), 22 pages. https://doi.org/10.1145/3359165
[19] Nigel Gilbert, Robin Wooffitt, and Norman Fraser. 1990. Organising Computer Talk. In *Computers and Conversation*, Paul Luff, Nigel Gilbert, and David Frohlich (Eds.). Academic Press, Chapter 11, 235–257. https://doi.org/10.1016/B978-0-08-050264-9.50016-6
[20] Jonathan Grudin. 2017. From Tool to Partner: The Evolution of Human-Computer Interaction. *Synthesis Lectures on Human-Centered Interaction* 10, 1 (2017), i–183. https://doi.org/10.2200/S00745ED1V01Y201612HCI035
[21] Jonathan Grudin and Richard Jacques. 2019. Chatbots, Humbots, and the Quest for Artificial General Intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article 209, 11 pages. https://doi.org/10.1145/3290605.3300439
[22] Drew Harwell. 2018. Why some accents don't work on Alexa or Google Home. https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/
[23] Ryuichiro Higashinaka, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, Yuka Kobayashi, and Masahiro Mizukami. 2015. Towards Taxonomy of Errors in Chat-Oriented Dialogue Systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 87–95. https://doi.org/10.18653/v1/W15-4611
[24] Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '99)*. Association for Computing Machinery, New York, NY, USA, 159–166. https://doi.org/10.1145/302979.303030
[25] Jiepu Jiang, Wei Jeng, and Daqing He. 2013. How Do Users Respond to Voice Input Errors? Lexical and Phonetic Query Reformulation in Voice Search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13)*. Association for Computing Machinery, New York, NY, USA, 143–152. https://doi.org/10.1145/2484028.2484092
[26] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 411. https://doi.org/10.1145/3290605.3300641
[27] Robert M Krauss and Susan R Fussell. 1991. Perspective-taking in communication: Representations of others' knowledge in reference. *Social cognition* 9, 1 (1991), 2–24. https://doi.org/10.1521/soco.1991.9.1.2
[28] Lenneke Kuijer and Elisa Giaccardi. 2018. Co-performance: Conceptualizing the role of artificial agency in the design of everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, Article 125, 13 pages. https://doi.org/10.1145/3173574.3173699
[29] J. C. R. Licklider. 1960. Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics* HFE-1, 1 (March 1960), 4–11. https://doi.org/10.1109/THFE2.1960.4503259
[30] Gustavo López, Luis Quesada, and Luis A. Guerrero. 2017. Alexa vs. Siri vs. Cortana vs. Google Assistant: A Comparison of Speech-Based Natural User Interfaces. In *Advances in Human Factors and Systems Interaction. AHFE 2017*, Isabel L. Nunes (Ed.), Vol. 592. Springer, Cham, 241–250. https://doi.org/10.1007/978-3-319-60366-7_23
[31] Ewa Luger and Abigail Sellen. 2016. Like Having a Really Bad PA: the Gulf Between User Expectation and Experience of Conversational Agents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '16)*. 5286–5297. https://doi.org/10.1145/2858036.2858288

[32] Gary Marchionini. 1995. *Information Seeking in Electronic Environments*. Cambridge University Press. https://doi.org/10.1017/CBO9780511626388

[33] Matthew Marge and Alexander I Rudnicky. 2019. Miscommunication Detection and Recovery in Situated Human–Robot Dialogue. *ACM Trans. Interact. Intell. Syst.* 9, 1, Article 3 (2019). https://doi.org/10.1145/3237189

[34] Michael McTear. 2008. Handling Miscommunication: Why Bother? In *Recent trends in Discourse and Dialogue*, Dybkjær L. and Minker W. (Eds.). Text, Speech and Language Technology, Vol. 39. Springer, Dordrecht, 101–122. https://doi.org/10.1007/978-1-4020-6821-8_5

[35] Michael McTear. 2018. Conversational Modelling for Chatbots: Current Approaches and Future Directions. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2018*, AndrÃĽ Berton, Udo Haiber, and Wolfgang Minker (Eds.). TUDpress, Dresden, 175–185.

[36] Michael McTear, Zoraida Callejas, and David Griol. 2016. *The conversational interface: Talking to smart devices*. Springer. https://doi.org/10.1007/978-3-319-32967-3

[37] Raveesh Meena, José Lopes, Gabriel Skantze, and Joakim Gustafson. 2015. Automatic Detection of Miscommunication in Spoken Dialogue Systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 354–363. https://www.aclweb.org/anthology/W15-4647

[38] Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, et al. 2014. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23, 3 (2014), 530–539. https://doi.org/10.1109/TASLP.2014.2383614

[39] Roger K Moore. 2017. Is Spoken Language All-or-Nothing? Implications for Future Speech-Based Human-Machine Interaction. In *Dialogues with Social Robots*, Jokinen K. and Wilcock G. (Eds.). Lecture Notes in Electrical Engineering, Vol. 427. Springer, Singapore, 281–291. https://doi.org/10.1007/978-981-10-2585-3_22

[40] M Granger Morgan, Baruch Fischhoff, Ann Bostrom, and Cynthia J Atman. 2002. *Risk communication: A mental models approach*. Cambridge University Press.

[41] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for How Users Overcome Obstacles in Voice User Interfaces. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, Article 6. https://doi.org/10.1145/3173574.3173580

[42] Katashi Nagao. 2019. Symbiosis between Humans and Artificial Intelligence. In *Artificial Intelligence Accelerates Human Learning*. Springer, 135–151. https://doi.org/10.1007/978-981-13-6175-3_6

[43] Jakob Nielsen. 1993. *Usability Engineering*. Academic Press, Inc.

[44] Sarah Perez. 2018. *Google says it sold a Google Home device every second since October 19*. https://techcrunch.com/2018/12/28/smart-speakers-hit-critical-mass-in-2018/

[45] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, Article 640. https://doi.org/10.1145/3173574.3174214

[46] Martin Porcheron, Joel E Fischer, and Sarah Sharples. 2017. Do animals have accents? Talking with Agents in Multi-Party Conversation. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing*. 207–219. https://doi.org/10.1145/2998181.2998298

[47] Antonio Roque and David Traum. 2008. Degrees of Grounding Based on Evidence of Understanding. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*. 54–63. https://www.aclweb.org/anthology/W08-0107/

[48] Johnny Saldaña. 2015. *The Coding Manual for Qualitative Researchers*. Sage.

[49] Emanuel A Schegloff. 2007. Sequence Organization in Interaction: A Primer. In *Conversation Analysis*. Vol. 1. Cambridge University Press.

[50] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I Hong. 2018. Hey Alexa, What's Up? A Mixed-Methods Studies of In-Home Conversational Agent Usage. In *Proceedings of the Designing Interactive Systems Conference (DIS '18)*. Association for Computing Machinery, New York, NY, USA, 857–868. https://doi.org/10.1145/3196709.3196772

[51] Phoebe Sengers and Bill Gaver. 2006. Staying Open to Interpretation: Engaging Multiple Meanings in Design and Evaluation. In *Proceedings of the 6th conference on Designing Interactive systems (DIS '06)*. Association for Computing Machinery, New York, NY, USA, 99–108. https://doi.org/10.1145/1142405.1142422

[52] Gabriel Skantze. 2005. Exploring Human Error Recovery Strategies: Implications for Spoken Dialogue Systems. *Speech Communication* 45, 3 (2005), 325–341. https://doi.org/10.1016/j.specom.2004.11.005

[53] Chairs Constantine Stephanidis, Gavriel Salvendy, Members of the Group Margherita Antona, Jessie Y. C. Chen, Jianming Dong, Vincent G. Duffy, Xiaowen Fang, Cali Fidopiastis, Gino Fragomeni, Limin Paul Fu, Yinni Guo, Don Harris, Andri Ioannou, Kyeong ah (Kate) Jeong, Shin'ichi Konomi, Heidi KrĚĘmker, Masaaki Kurosu, James R. Lewis, Aaron Marcus, Gabriele Meiselwitz, Abbas Moallem, Hirohiko Mori, Fiona Fui-Hoon Nah, Stavroula Ntoa, Pei-Luen Patrick Rau, Dylan Schmorrow, Keng Siau, Norbert Streitz, Wentao Wang, Sakae Yamamoto, Panayiotis Zaphiris, and Jia Zhou. 2019. Seven HCI Grand Challenges. *International Journal of Human-Computer Interaction* 35, 14 (2019), 1229–1269.

https://doi.org/10.1080/10447318.2019.1619259

[54] Lucy Suchman. 2007. *Human-machine reconfigurations: Plans and situated actions.* Cambridge university press. https://doi.org/10.1017/CBO9780511808418

[55] Paul Ten Have. 2007. *Doing Conversation Analysis* (2nd ed.). Sage. https://doi.org/10.4135/9781849208895

[56] Jason Wu, Karan Ahuja, Richard Li, Victor Chen, and Jeffrey Bigham. 2019. ScratchThat: Supporting Command-Agnostic Speech Repair in Voice-Driven Assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 2, Article 63 (2019). https://doi.org/10.1145/3328934

Received October 2019; revised January 2020; accepted March 2020