Exploring the Role of Conversational Cues in Guided Task Support with Virtual Assistants

Alexandra Vtyurina[*]
University of Waterloo
Waterloo, Ontario
avtyurin@uwaterloo.ca

Adam Fourney
Microsoft Research
Redmond, Washington
adamfo@microsoft.com

ABSTRACT

Voice-based conversational assistants are growing in popularity on ubiquitous mobile and stationary devices. Cortana, as well as Google Home, Amazon Echo, and others, can provide support for various tasks from managing reminders to booking a hotel. However, with few exceptions, user input is limited to explicit queries or commands. In this work, we explore the role of implicit conversational cues in guided task completion scenarios. In a Wizard of Oz study, we found that, for the task of cooking a recipe, nearly one-quarter of all user-assistant exchanges were initiated from implicit conversational cues rather than from plain questions. Given that these implicit cues occur in such high frequency, we conclude by presenting a set of design implications for the design of guided task experiences in contemporary conversational assistants.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords

conversational systems; conversational cues; task support

1. U: Alright. Uh... What ingredients do I need?
2. A: One-half teaspoon of chili powder.
3. *U: Okay.*
4. A: One-eighth teaspoon of dried oregano.
5. *U: Oregano. Okay.*
6. A: One pinch onion powder.
7. *U: Okie-doke.*
8. A: Cayenne pepper to taste.
9. *U: Sounds good.*
10. U: Alright, can you repeat that?
11. A: Cayenne pepper to taste.
12. U: Uh, I meant all the ingredients.

Figure 1. An example exchange between a user (U) and the agent (A) during the Wizard of Oz study. Italic user utterances are implicit conversational cues – utterances that advance the conversation and move the user closer to their goal, without the user asking an explicit question nor giving an imperative command.

INTRODUCTION

The popularity of intelligent assistants, such as Cortana, Alexa, and Siri has been steadily growing over the past few years [1]. People are using voice-based assistants on their phones, personal computers (PCs), and internet of things (IoT) devices to complete short tasks, such as setting up an alarm clock or checking weather forecasts [17, 21], as well as for multi-step tasks such as making a restaurant reservation or booking a taxi. Increasingly, devices like the Amazon Echo or the Google Home are placed in the kitchen [23], and are used at mealtime to set timers or to manage short processes related to cooking [12]. As such, companies have begun to develop experiences that are tailored to this scenario, aiming to help users follow recipes; or, more generally, to follow step-by-step instructions [24, 28]. While these guided task completion scenarios can be quite rich and complex [4, 6, 18], the protocol for communicating with contemporary voice assistants is comparatively simple and constrained, and usually follows the <*trigger word, question, answer*> paradigm. For example, *–"Hey Google, what time is it in Montreal?", –"The time in Montreal, QC is 12:48 PM"*.

In this paper, we explore potential interactions that occur in the moments following, or in lieu of, users' explicit <*trigger word, question, answer*> triples. To investigate these utterances, we ran a high-fidelity Wizard of Oz study in which we asked people to interact with a conversational agent as they prepared a simple culinary recipe. As participants engaged with the agent, we observed them using numerous verbal conversational cues. In the context of task guidance, we view such cues as requests that are neither clearly phrased as questions (e.g., "*What do I do next?*", "*What else?"*), nor as imperative commands ("*Read me the next step.*", *"Next step."*), yet nonetheless, serve to advance the conversation and move the user closer to completing their task. Importantly, the cues are highly dependent on the context in which they are spoken, and cannot be easily interpreted in isolation.

We report the prevalence and roles of conversational cues in this scenario. Furthermore, we argue that a system that is able to recognize and correctly act on these cues can achieve

[*] Work done while at Microsoft Research.

CHI 2018, April 21–26, 2018, Montreal, QC, Canada

ISBN 978-1-4503-5620-6/18/04. . . $15.00
DOI: https://doi.org/10.1145/3173574.3173782

high levels of user satisfaction, even when constrained to a simple response model (e.g., limited to sentence selection for question answering [32]).

The remainder of the paper is structured as follows. We present background material and describe the Wizard of Oz experiment. We introduce the taxonomy of verbal conversational cues for a task-oriented dialogue. We investigate the different purposes of short affirmative utterances (e.g., "*Okay.*"), as well as conversational cues that repeat a previous system response. The paper concludes by presenting design implications for future voice-enabled systems.

BACKGROUND

Human-to-human conversation is an immensely complex process, filled with implicit cues – both verbal and nonverbal [19, 14, 9, 30, 5]. These *conversational cues* allow people to better communicate agreement or disagreement, emotional state, and whether they successfully perceived and processed the information [22]. These types of interactions are very natural and effortless, and people may exhibit them even when interacting with automatic systems that are incapable of perceiving or reciprocating these signals [17, 20]. Perceiving, understanding, and reacting to conversational cues is especially important in conversations that aim to provide guided task support [14, 13], which are the focus of this paper.

In the domain of guided task support, researchers have also explored human-agent interactions, and the potential to provide automated task guidance with a conversational system. Closest to our work is research conducted by Martins et al. in [18]. Martins et al. explored how a semi-automatic assistant could guide the user in following cooking recipes[1]. However, the system responded only to a fixed set of 1-word commands (e.g., "*next*", "*previous*", "*repeat*", and "*how*"), leaving little room for natural discourse.

In a similar vein, Bohus et al. [7] describe a dialogue management tool – RavenClaw – that was used to build multiple task-support oriented dialogue systems [4, 6, 26]. RavenClaw is described to have a two-tier architecture: a layer that captures domain-specific information, and a layer that is responsible for turn-taking and grounding behavior, and requests like "help", "resume", "repeat", etc. that are task-independent. Though RavenClaw can process and understand explicit requests for confirmation and disambiguation, it does not handle the types of implicit requests that we describe in this paper.

More recently, new solutions appeared in the market that offer simple step-by-step guidance for cooking recipes. The most prominent examples are Google Home's recipe support [24], as well as Amazon's Alexa support of recipes from *allrecipes.com* [28]. However, interactions are mainly constrained to the protocol of <*trigger word, question, answer*> described in the introduction.

However, human-to-human communication is not limited to strict turn-taking: People interrupt each other and provide unsolicited feedback. Sacks et al. [27] analyze and detail rules according to which the turn-taking behaviour occurs. Likewise, Clark & Schaefer [9] develop a model of speakers contributing to a discussion. In these works, conversational cues play an important role in signalling, and in allowing participants to establish common ground. Building on this line of research, Porcheron et al. [25] characterize the changes that happen when an automatic conversational assistant is introduced as one of the discourse participants. In our Wizard of Oz study, the simulated conversational agent relied on many of the same verbal cues to identify when to advance the conversation.

Finally, conversations can often be broken down into a series of dialogue acts (e.g., statements, questions, acknowledgements, back-channel, etc.). The work by Stolke et al. [29] demonstrates how such dialogue acts can be identified using probabilistic models. One challenge is that, for various dialogue acts, the interpretation of a given cue or utterance is highly context dependent. To this end, Gravano et al. [13] study short affirmative cues, such as "*alright*", "*mm-hm*", "*okay*", etc. and their roles in conversations: agreeing with the interlocutor, displaying interest, cueing the start of a new topic. They explore a number of features that could be used to correctly identify the goal of the speaker, including: lexical, timing, phonetic, acoustic and other properties. We too observe these behaviors, and we further observe a phenomenon in which people repeat – sometimes over an over – a response previously spoken by the agent. We explore the role of these cues in the specific context of following a guided task.

Considering the latest improvements in speech recognition [31], the tooling now exists to capture and transcribe user utterances in high fidelity. We believe, that the next step should be applying this technology to further improve user experience with voice-based interfaces.

EXPERIMENT

In this work, we developed a high-fidelity Wizard of Oz simulation to study the role of conversational cues in guided task scenarios. We describe the protocol and apparatus below.

Procedure

We invited 10 participants (6 male, 4 female, average age 30), to engage with a simulated conversational assistant with the goal of preparing a simple culinary recipe. Out of the 10 participants, 2 reported having used an intelligent assistant earlier that day, 5 – earlier that week, and 1 each – earlier that month, more than a month ago and never. 8 people said they usually enjoyed cooking, and 6 said they cooked often.

The experiment took place in an office at our research facility. Participants were briefed upon arrival, but were not instructed on what natural language commands to use when communicating with the conversational agent, nor were the participants informed that the agent was a simulation. Instead, the participants were simply instructed to naturally converse with the agent in order to prepare a spice rub recipe[2]. This recipe was chosen because it includes numerous preparation steps and ingredients, but makes limited use of cooking surfaces or appliances. I.e., it is ideal for a lab environment.

[1] Martins et al. also experimented with car repair scenarios.

[2] http://allrecipes.com/recipe/17338/tasty-bbq-corn-on-the-cob/

The experiment began when a participant uttered the phrase “start cooking”, and concluded when the participant completed the penultimate step of the recipe (the final step involved grilling the corn on a barbecue). During the experiment, interactions were mediated via a speakerphone, which relayed user utterances to an operator seated in another room. The operator then selected responses from a preset list, which were then played back to the participant in a computer-synthesized voice [3]. All participant actions were audio and video recorded.

Upon concluding the recipe, participants were asked to complete the NASA Task Load Index (TLX) [15] and System Usability Scale [8] (SUS) questionnaires. Given our research focus, and the simulation aspect of this experiment, these questionnaires served primarily as a check to ensure that the simulation was of sufficient quality and completeness to warrant the further investigation of the subtler aspects of the human-agent interaction.

Finally, we conducted semi-structured interviews and debriefed participants about the simulation.

Apparatus

The success of this experiment depended on the fidelity of our Wizard of Oz simulation. Here, our goal was to minimize latency and ensure consistency across repeated responses. As noted above, we developed a preset list of computer-synthesized audio responses from which an experimenter – the wizard – could select. Mimicking existing recipe agents [28, 24], the response list included the recipe’s ingredients, and each sentence from its instructions, as distinct candidate answers. Additionally, we included relevant culinary definitions (e.g., “*a pinch*”, “*to taste*”, etc.), and a “*no answer*” response to handle questions that were out of scope. We also allowed a free-form response to be typed by the Wizard, but it was rarely used, and a post-experiment analysis showed that this option was used mainly to produce “*yes*” or “*no*” responses. We believe these efforts were successful – no participants reported suspecting that they were interacting with a simulation.

RESULTS

In this section, we quickly review results of the TLX and SUS evaluations, then describe the most common explicit requests and implicit conversational cues that we observed in the study.

General impressions

All 10 participants successfully completed the recipe, taking an average of 6.56 minutes (min = 3.22, max = 8.57) and 19 conversational turns (min = 9, max = 27) to reach the final step. The simulation received favorable scores on both the TLX (median = 21.25, IQR = 14), and SUS (median = 83.75, IQR = 13) scales. Notably, participants reported low frustration (median = 25, IQR = 16) and low effort (median = 20, IQR = 21) on the TLX. Likewise, participants reported high levels of confidence (median = 5, IQR = 1), and low levels of inconsistency (median = 1, IQR = 0.75) via the SUS. Taken together these findings suggest that the simulation was of sufficient quality and completeness to effectively ground the analysis that follows.

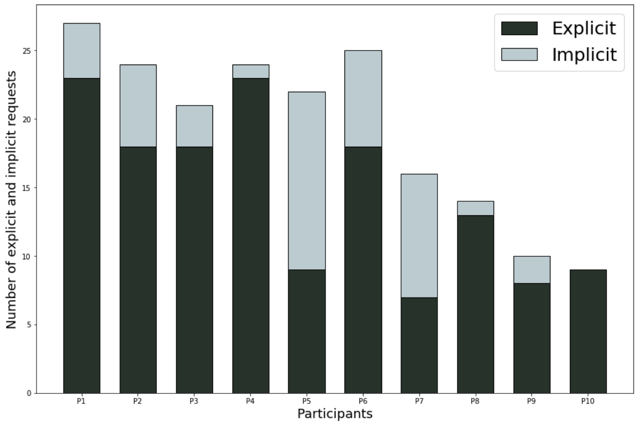

Figure 2. Distribution of categories across participants

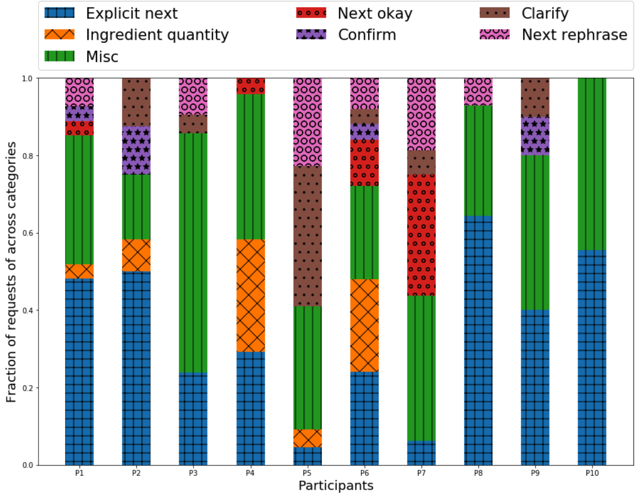

Figure 3. Distribution of categories across participants

Conversational Cues

In an initial briefing, participants were instructed to speak with the agent naturally, as if they were conversing with another person. In fact, only a single explicit command was mentioned to participants: “start cooking”, that activated the system. Given these limited constraints, participants very quickly adopted a highly conversational style of dialogue, rich with conversational cues. As an example, after the agent read the very first ingredient to P6, she simply responded with “*Okay*”, then waited for the bot to continue listing the second ingredient (Figure 1).

Given the frequency and richness of these cues, we set out to study the phenomena in detail. We begin by simply counting the number of cases in which a system response was triggered by a conversational cue, as defined earlier. Namely, system responses were manually labelled as either resulting from an explicit question or statement (e.g., “what is the next step?”), or from an implicit dialogue cue (e.g., “*Okie dokie*”). To do this, two researchers (authors of this paper) independently labelled 50 bot responses, achieving high inter-rater reliability (Cohen’s $\kappa = 0.92$). One researcher then continued to label the remaining 142 responses. In the end, 46 bot responses (24%), were deemed to have been initiated by an implicit

[3] https://responsivevoice.org/

conversational cue, similar to those outlined above. Figure 2 shows the absolute counts of implicit (light grey) and explicit (dark grey) user requests for each participant.

While labelling the aforementioned interactions, we observed that implicit cues serve numerous intents. We examine intents of implicit as well as of explicit requests, below, beginning with the latter.

Explicit Intents

A total of 146 responses were initiated from an explicit query or imperative command. We briefly describe the 5 most popular intents, representing 58.3% of all user requests. We then discuss the intents of the implicit conversational cues in the next section.

Explicit Next

The most common request simply asked for the next ingredient or step of the recipe (Figure 3, blue). These questions constituted 32.8% of all user requests, and varied greatly in their phrasing: E.g., "*What is step number two?*", "*Next*", "*What else?*", etc.

Ingredient Quantity

Comprising 8.9% of all requests, were those asking about ingredient quantities (Figure 3, orange). Examples include: "*How much chili powder do I need?*", "*How much of each do I need?*". The lexical variety was quite low for this category.

Repeat

7.8% of all requests, were for the agent to repeat a prior response (Figure 3, green; i.e., incorporated in "misc"). Examples include: "*Can you please repeat what you just said?*", or "*Can you repeat that?*".

List Ingredients

A smaller fraction of all requests – 6.2% – were to list all, or some subset of the ingredients (Figure 3, green; "misc"). Examples include: "*Okay, thank you. Do I need any other ingredients?*", "*Okay. Is there anything else?*", and "*Alright. So... I've got a few spices in front of me. Can you tell me the spices I need to make the rub for the corn?*". As these examples illustrate, these requests varied considerably in diversity and complexity.

Definitions

Finally, 2.8% of all requests asked for the definitions of ingredients, or cooking procedures (Figure 3, green; "misc"). Examples include "*What is an ear of corn?*", "*What do you mean to taste? How much should I put?*", "*How high is medium-high heat is?*".

In the next sections, we consider implicit requests made to the system.

Implicit Intents

When interacting with the simulated agent, nine of ten participants initiated responses using some form of implicit conversational cues. These cues served a number of distinct intents, as described below.

1. A: Garlic powder to taste.
2. U: Yup. Done that.
3. A: Salt and pepper to taste.
4. U: I don't have salt. Can I omit the salt?
5. A: Yes.
6. *U: Okay. Pepper's in already.*
7. A: Step number 3: blend in the softened butter.
8. U: So I pretend to do that now?
9. *U: Okay, everything's blended.*
10. A: Step number four: apply this mixture to each ear of corn, and place each ear onto a piece of aluminum foil big enough to wrap the corn.
11. *U: Okay. So I applied it to the corn.*
12. A: Step number five: Wrap like a burrito, and twist the ends to close.
13. *U: Ends are twisted.*

Figure 4. Example of a user (U) showing readiness to proceed by repeating previous system response (utterances 6, 9, 11 and 13).

Implicit next

Given the nature of the task, requests for the next step, or next ingredient were the most common. However, 15.6% of the requests were not phrased as explicit questions. Rather, in 7.8% of cases, participants used short positive utterances, such as "*yup*", "*alright*" to signal that the current step was completed and they were ready to proceed. Figure 1 shows an example of such interactions, as the user's utterances "*3: Okay.*" and "*7: Okie-doke.*" signal that she is prepared to continue.

In another 7.8% of cases, participants would paraphrase the step they have just completed, to signal that they were ready to go on to the next step, expecting the system to read the next instruction or ingredient in response. Interaction of this type are outlined in Figure 4, where in utterances 6, 9, 11, and 13 the participant is describing the last completed instruction in his own words, showing that he is done with this step and is ready to move on.

The first row of Table 1 illustrates the counts of next step requests using short positive utterances, paraphrase and explicit questions.

Grounding behavior

During the experiment, we noticed, that although the participants did not know what parts of their speech the system could and could not understand, they would still respond to the system's statements. The purpose of these responses is to let the other speaker – in our case the system – know that the information has been processed and accepted, and that the

Purpose	All	Paraphrase / Repeat	Okays	Explicit
Next	93	15	15	63
Ack	48	16	32	n/a
Memory	32	19	13	n/a

Table 1. Distribution of user utterances requesting next item on the list (Next), showing acceptance of previous system response (Ack), and utterances spoken to keep short term memory updated (Memory).

1. U: How much onion powder?
2. A: One pinch onion powder.
3. *U: One pinch. Okay,* and how much oregano?
4. A: One-eighth teaspoon of dried oregano.
5. *U: Okay.*

Figure 5. Example of acknowledgement by the user (U) with okay's and repetitions (utterances 3, 5).

dialogue may continue. In the literature, this has been referred to as *grounding behavior* [9].

Grounding behavior can also be exhibited using short positive utterances, as well as partial, or verbatim repetitions of the previous content. These behaviors have been called "acknowledgements", "demonstration" and "display" [9]. Grounding cues closely resemble those of the *implicit next* category and are chiefly differentiated by how they are manifested in a conversational turn. For example, utterance 3 in figure 5 shows a participant paraphrasing the agent's prior response, then using a short affirmative phrase ("*Okay*"), and finally, without pause, proceeding to explicitly ask about the next ingredient. This timing pattern precludes these cues from having an *implicit next* intent.

The second row of Table 1 gives counts of different types of grounding behavior that has been observed in the study (we considered an utterance to be a repetition if it either partially or fully, repeats a system response verbatim or paraphrased). During the study, we noticed that people 8 of 10 participants repeated a system response out loud, at least once, while preparing the recipe.

Rehearsing behavior

We also noticed a curious phenomenon. Our participants talked to themselves while they were in the process of completing a step. With this sort of "memory rehearsal" behavior people refresh and maintain items in their short-term memory [10], which is believed to rely on the same pathways as language and speech. Consequently, people often narrate recipes, as Figure 6 demonstrates. In that case, the participant was repeating the name of the ingredient he was looking for, while he was looking for it.

Clarifications and Confirmations

Additionally, response repetitions came as clarifying questions (Figure 7). Whenever people didn't understand the system's response, had doubts about its correctness, or needed more detailed information, they would often repeat a part of the system's response that was not clear, expecting it to provide more thorough explanation.

1. A: Garlic powder to taste.
2. *U: Garlic powder...*
3. *U: Garlic powder to taste...*
4. *U: Garlic powder to taste... Okay, one second.*
5. *U: Garlic powder... Garlic powder...*

Figure 6. Example of a user (U) repeating the response to himself while completing the step (utterances 2, 3, 4, 5)

Closely related to clarifying questions are those that seek confirmations. We observed 10 (5.2%) such user utterances throughout the experiment. Their purpose is to confirm user's belief about a step in the recipe. An example is listed in Figure 8. Here the first part of utterance 2 is reiterating previous content and user's actions, while the second part serves as a cue for confirmation.

1. A: One quarter cup butter, softened.
2. *U: One quarter of the butter?*
3. A: One quarter cup butter, softened.
4. U: One quarter cup butter. Okay.

Figure 7. Example of a user (U) asking for clarification on the previous response (utterance 2).

IMPLICATIONS, LIMITATIONS AND FUTURE WORK

As it has been shown above, conversational cues constitute a large portion of interactions between a user and a conversational agent. A conversational system that is able to recognize and act upon these requests will enable its users to converse using a more natural-style of language, yielding high satisfaction scores.

In our Wizard of Oz study, participants did not need to issue a trigger word to initiate interactions, and we believe that this property is one reason we observe such a high frequency of short conversation cues such as "*Okay*" and "*Yup*". One of the advantages of our simulation was that it was listening to the user at all times, which could be challenging to implement in practice (there are both technical limitations and privacy concerns). However, in our study, most of these conversational cues followed shortly after an agent's prior responses. Leaving the microphone on for a few moments after each response may be an acceptable compromise and could allow for a more seamless dialogue flow.

We have also observed that, despite having different intents, many conversational cues and utterances transcribe into the same lexical representation. However, contemporary virtual assistant frameworks follow a pipeline architecture, transcribing user utterances prior to doing intent classification [2, 3]. In such an architecture, correct classification of these conversational cues will be challenging if not impossible. To extend their functionality, frameworks and SDKs should include information about prosody, and other acoustic features. These features have already proven to be valuable in improving the detection of dialogue acts [29, 16], which can be seen as a similar – but coarser-grained – taxonomy of spoken intents. Likewise, we believe it will be important for situated agents to model a user's attention, either through acoustic features alone [11], or through gaze, so as to facilitate addressee detection. This will allow more conversational cues to be captured in the first place, by allowing the mic to stay on between utterances, and perhaps by eliminating wake words altogether.

Finally, we realize that our singular focus on recipes might raise questions about the generalizability of the findings. Prior

1. *U: So I applied all the ingredients on the corn, and then applied the softened butter and wrapped it with the aluminum. Right?*
2. A: Correct.
3. U: Perfect. What's next?

Figure 8. Example of a user (U) confirming an existing belief about a recipe step (utterance 1).

work has studied the importance of conversational cues in human-human task-oriented dialogue over a range of tasks [12, 13]. Our Wizard of Oz study shows that the importance of these cues extends to at least one class of human-agent task-oriented dialogue: cooking while interacting with a voice-based conversational assistant. Though our protocol employed a simulated agent, we took steps to ensure that our simulation was convincing, and was close in fidelity to existing voice assistants. To this end, we believe it is reasonable to expect to see categories like implicit next, grounding, clarification, and confirmation, in other human-agent task-oriented conversations.

CONCLUSION

Current voice-based conversational assistants mostly abide by the <*trigger word, question, answer*> paradigm, which constrains user interactions, and a number of implicit conversational cues are missed as a result. In this work we have considered a set of common implicit verbal cues exhibited by users of a simulated conversational assistant for the task of cooking a culinary recipe. We have described these cues and their intents in detail and have provided a set of design implications for designing task-oriented conversational systems.

REFERENCES

1. 2017. Alexa, Say What?! Voice-Enabled Speaker Usage to Grow Nearly 130% This Year. `https://www.emarketer.com/Article/Alexa-Say-What-Voice-Enabled-Speaker-Usage-Grow-Nearly-130-This-Year/1015812`. (2017). [Online; retrieved 5-Jan-2018].
2. 2017. Alexa Skills Kit. `https://developer.amazon.com/alexa-skills-kit`. (2017). [Online; retrieved 5-Jan-2018].
3. 2017. The Cortana Skills Kit. `https://docs.microsoft.com/en-us/cortana/getstarted`. (2017). [Online; retrieved 5-Jan-2018].
4. G. Aist, J. Dowding, B. A. Hockey, M. Rayner, J. Hieronymus, D. Bohus, B. Boven, N. Blaylock, E. Campana, S. Early, G. Gorrell, and S. Phan. 2003. Talking Through Procedures: An Intelligent Space Station Procedure Assistant. (2003), 187–190. DOI: `http://dx.doi.org/10.3115/1067737.1067781`
5. Dan Bohus and Eric Horvitz. 2014. Managing human-robot engagement with forecasts and... um... hesitations. In *Proceedings of the 16th international conference on multimodal interaction*. ACM, 2–9.
6. Dan Bohus and Alexander Rudnicky. 2005. LARRI: A language-based maintenance and repair assistant. *Spoken multimodal human-computer dialogue in mobile environments* (2005), 203–218.
7. Dan Bohus and Alexander I. Rudnicky. 2003. RavenClaw: Dialog management using hierarchical task decomposition and an expectation agenda. In *Proceedings of the 8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003*. ISCA, 203–218.
8. John Brooke and others. 1996. SUS – A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.
9. Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive science* 13, 2 (1989), 259–294.
10. Fergus I.M. Craik and Robert S. Lockhart. 1972. Levels of processing: A framework for memory research. *Journal of verbal learning and verbal behavior* 11, 6 (1972), 671–684.
11. Florian Eyben, Felix Weninger, Lucas Paletta, and Björn W. Schuller. 2013. The acoustics of eye contact: detecting visual attention from conversational audio cues. In *Proceedings of the 6th workshop on Eye gaze in intelligent human machine interaction: gaze in multimodal interaction*. ACM, 7–12.
12. David Graus, Paul N. Bennett, Ryen W. White, and Eric Horvitz. 2016. Analyzing and Predicting Task Reminders. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*. ACM, 7–15.
13. Agustín Gravano, Julia Hirschberg, and Štefan Beňuš. 2012. Affirmative cue words in task-oriented dialogue. *Computational Linguistics* 38, 1 (2012), 1–39.
14. Barbara J. Grosz and Candace L. Sidner. 1986. Attention, Intentions, and the Structure of Discourse. *Comput. Linguist.* 12, 3 (July 1986), 175–204.
15. Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology* 52 (1988), 139–183.
16. Julia Kiseleva, Kyle Williams, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Predicting user satisfaction with intelligent assistants. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 45–54.
17. 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 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5286–5297.
18. Filipe Martins, Joana Paulo Pardal, Luís Franqueira, Pedro Arez, and Nuno J. Mamede. 2008. Starting to cook a tutoring dialogue system. In *Spoken Language Technology Workshop, 2008. SLT 2008. IEEE*. IEEE, 145–148.
19. Masahiro Mizukami, Koichiro Yoshino, Graham Neubig, David R. Traum, and Satoshi Nakamura. 2016. Analyzing the Effect of Entrainment on Dialogue Acts.. In *SIGDIAL Conference*. 310–318.

20. Kellie Morrissey and Jurek Kirakowski. 2013. “Realness” in Chatbots: Establishing Quantifiable Criteria. In *International Conference on Human-Computer Interaction*. Springer, 87–96.
21. Kevin Murnane. 2017. IFTTT Survey Provides Insight Into What People Do With Amazon's Echo And Google's Home. https://www.forbes.com/sites/kevinmurnane/2017/07/12/ifttt-survey-provides-insight-into-what-people-do-with-voice-controlled-assistants. (2017). [Online; retrieved 12-July-2017].
22. Duyen T. Nguyen and Susan R. Fussell. 2016. Effects of Conversational Involvement Cues on Understanding and Emotions in Instant Messaging Conversations. *Journal of Language and Social Psychology* 35, 1 (2016), 28–55.
23. Sheryl Ong and Aaron Suplizio. 2016. Unpacking the Breakout Success of the Amazon Echo. https://www.experian.com/innovation/thought-leadership/amazon-echo-consumer-survey.jsp. (2016). [Online; retrieved 20-October-2016].
24. Emma Persky. 2017. Now we're cooking – the Assistant on Google Home is your secret ingredient. https://www.blog.google/products/assistant/cooking-with-the-assistant-google-home-your-secret-ingredient/. (2017). [Online; retrieved 12-July-2017].
25. Martin Porcheron, Joel E. Fischer, and Sarah Sharples. 2017. "Do Animals Have Accents?": Talking with Agents in Multi-Party Conversation. (2017), 207–219. DOI: http://dx.doi.org/10.1145/2998181.2998298
26. Antoine Raux, Brian Langner, Alan W. Black, and Maxine Eskenazi. 2003. Let’s go: Improving spoken dialog systems for the elderly and non-natives. In *Proceedings of the 8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003*. ISCA, 753–756.
27. Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* (1974), 696–735.
28. AllRecipes Staff. 2016. Introducing a Cool New Way to Cook: Allrecipes on Amazon Alexa. http://dish.allrecipes.com/introducing-allrecipes-on-amazon-alexa/. (2016). [Online; retrieved 5-Jan-2018].
29. Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics* 26, 3 (2000), 339–373.
30. David R. Traum and Elizabeth A. Hinkelman. 1992. Conversation Acts in Task-Oriented Spoken Dialogue. *Computational intelligence* 8, 3 (1992), 575–599.
31. Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. 2016. Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256* (2016).
32. Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. 2014. Deep learning for answer sentence selection. *arXiv preprint arXiv:1412.1632* (2014).