Single or Multiple Conversational Agents? An Interactional Coherence Comparison

Ana Paula Chaves[1,2]
[1] UTFPR-CM
Campo Mourão, Brazil
anachaves@utfpr.edu.br,
acs549@nau.edu

Marco Aurelio Gerosa[2]
[2] Northern Arizona University
Flagstaff, AZ, USA
Marco.Gerosa@nau.edu

ABSTRACT
Chatbots focusing on a narrow domain of expertise are in great rise. As several tasks require multiple expertise, a designer may integrate multiple chatbots in the background or include them as interlocutors in a conversation. We investigated both scenarios by means of a Wizard of Oz experiment, in which participants talked to chatbots about visiting a destination. We analyzed the conversation content, users' speech, and reported impressions. We found no significant difference between single- and multi-chatbots scenarios. However, even with equivalent conversation structures, users reported more confusion in multi-chatbots interactions and adopted strategies to organize turn-taking. Our findings indicate that implementing a meta-chatbot may not be necessary, since similar conversation structures occur when interacting to multiple chatbots, but different interactional aspects must be considered for each scenario.

Author Keywords
Chatbot; Dialog agent; human-agent communication.

ACM Classification Keywords
H.5.2. Information interfaces and presentation (e.g., HCI): Natural language.

INTRODUCTION
Conversational agents, also called chatbots, have drawn attention from both academia and industry [9,66]. Commercially available systems include conversational agents that talk to users through text or voice [43,51,60]. Pandorabots Platform hosted more than 285,000 conversational agents in April 2017. It is expected that messaging platforms take places of many websites and apps with graphical user interfaces [13].

Most available chatbots are designed to perform tasks in a highly specific domain. For instance, there are chatbots able to book flights and hotels, recommend restaurants, organize calendars [58], and provide tourist information [15]. These chatbots do not interact with each other, and a human still must combine the received information to make a decision.

To mitigate this issue, some practitioners suggest developing meta-chatbots [1,21,39] to combine available chatbots and assemble a single chatbot to satisfy the user's multiple-domain needs. Alternatively, chatbots could be adapted to participate in a multi-party conversation, federating user requests across a community of agents. Designing a meta-chatbot requires coordination of specific-domains chatbots in background; and putting multiple interlocutors in a single chat tool may create conversational problems, such as the increase in occurrences of entangled messages, disruptive sequences, and topic exchange, which may result in incoherent conversations [4,25,45,54]. Indeed, the literature of human-human and computer-mediated conversations have shown that two- and multi-party interactions have different characteristics [23,25,65].

In this paper, we focus on comparing the interactional coherence in both scenarios, following the method proposed by Herring [25]. For this purpose, we conducted a Wizard of Oz experiment with 24 participants divided into two groups. Participants from the first group interacted with a meta-chatbot, which worked as a facade to three different specialized chatbots; participants from the second group interacted directly with the chatbots in a single conversation.

Our main contribution is to identify differences in interaction when users engage either single or multiple persona chatbots for an information gathering task. We found evidence that organizing turn-taking is more strongly related to the number of interlocutors than to the coherence imposed by the conversational structure. As domain-specialized chatbots may lower user expectancy and can have better natural language processing (NLP), since the vocabulary and topics become more restricted, our results indicate that this may be a feasible approach to integrate the myriad of specialized chatbots available. We showed that several aspects of the conversation dynamics are not significantly different, although participants reported feeling more confused in multi-chatbots than in single-chatbot interactions. Users decide for themselves which agent to interact with, thus effectively offloading some of the NLP burdens to the intelligence of the user. We conclude with insights for designing conversational agents.

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

ACM ISBN 978-1-4503-5620-6/18/04$15.00
https://doi.org/10.1145/3173574.3173765

BACKGROUND

Conversational agents are computer programs that interact with humans using natural language [56]. Although the idea of conversational agents is not new [9], advances in areas such as artificial intelligence [7], natural language processing [28], and cognitive systems [44] have boosted their growth.

The social responsiveness trigged in humans when interacting with computers [12,41,42] favors the use of conversational agents. However, recent studies showed that people change behavior in human-machine communication [40,57]. Besides, the literature on human-to-human interaction has discussed differences between two- and multi-party interactions [23,25,29,65]. However, little is known about scenarios in which conversational agents are interlocutors. In the next section, we discuss why comparing two- vs. multi-party conversations are timely and important in human-chatbot interactions.

Single- vs. multi-party interactions

Most currently available chatbots are developed to manage only two-party conversations (a single conversational agent interacting with a single user). Two-party interactions are easier to manage because the chatbot and user intercalate turns-to-speak throughout the conversation [3]. However, current limitations in NLP prevent a single chatbot to hold conversations in a wide domain of expertise [17], and several tasks demand multi-domain knowledge. For example, planning a trip involves accommodations, food, things to do, weather forecasts, safety, transportation, and so forth. Developing a single chatbot capable of holding multi-domain conversations is still an open challenge [20].

Alternatively, some studies suggest to maintaining the conversation in a two-party scenario, but manage the multiple-domain aspect of the interaction in background. Practitioners named it as meta-chatbot [1,21,39], which is a single chatbot that is capable of wrapping the knowledge of domain-specific chatbots that can be provided by multiple vendors. In this line, Griol and Molina [20] propose an architecture to integrate agents in a single interface.

In contrast, domain-specific chatbots may participate in multi-party conversations, where each agent has its own expertise and somehow coordinate the conversation to each other to perform a task. For example, Bayser et al. [3] propose an architecture for multi-party conversational systems, where a "mediator chatbot" is responsible for calling domain-specific chatbots to the conversation.

Nevertheless, there is a lack of studies to compare the alternatives. In this paper, we analyze whether the number of chatbots as interlocutors impacts the conversation structure and users' behaviors. Inspired by previous research in interactional coherence [2,4,14,45], we analyzed each scenario from two perspectives: turn-taking and sequential coherence. The next sections discuss how these concepts have been investigated in the previous literature.

Turn-taking

Turn-taking represents the agreement that each person engaged in a conversation will have the opportunity to talk [52]. In two-party conversations, it is expected that participants alternate turns [52]. In multi-party scenarios, the conversations may include additional protocols, because participants negotiate the next speaker [16]; for example, explicitly assigning a turn [52,63] or self-selecting a turn when realizing that the current turn is complete [2]. Current chatbots are not prepared to self-select turns, and only interact when a turn is explicitly assigned to them [65]. Recent studies [48,50] analyze turn-taking in multi-party scenarios including conversational agents. In these studies, the conversational agents do not interactively communicate, but only answer a request when they are required to do so. In the example presented by Bayser et al. [3], the turns are mostly coordinated by the mediator chatbot, who explicitly call the specialized chatbots to speak. Uthus and Aka [65], on the other hand, propose a chatbot that self-select a turn to introduce a topic from external sources, but this chatbot does not consider the topic being discussed.

In this paper, we analyze whether users behave differently when one or more conversational agents are involved in a turn-taking process (deciding who should talk). The research question that guided this analysis was:

RQ1: Does the multiplicity of chatbots as interlocutors influence how people organize turn-taking?

Conversation analysis theory [52] argues that interlocutors work to avoid gaps and overlaps between messages in face-to-face conversations. By contrast, in the computer-mediated communication field, Anderson et al. [2] state that lengthy strategic pauses work as a cue to another interlocutor to self-select a turn; Porcheron et al. [48] observed mutual production of silence when people interact with a single conversational agent in a multi-party scenario, which shows an agreement to hand over the turn to the agent.

In this paper, we observed users' behaviors when they finish their turns and wait for chatbots' responses. We also discuss whether the gaps between messages caused more overlaps when interacting with a single or multiple chatbots. The research question that guided this analysis was:

RQ2: Does the multiplicity of chatbot as interlocutors influence the occurrence of topic overlap?

Sequential coherence

The chatbot discussed by Uthus and Aka [65] can self-select a turn to introduce new topics, but it does not preserve the sequential coherence, because it does not consider the topic being discussed. Sequential coherence is the unification of two complementary principles in conversation [25]: adjacency pairs and relevance. Adjacency pairs describe how responses or follow-up to a previous turn should occur adjacent to one another in a temporal sequence. An adjacency pair comprises at least two utterances: a first pair

part (e.g., question) and the correspondent second pair part (e.g., response) [25,52]. Relevance requires that the responses or follow-ups should be semantically close or clearly related to the previous turns [52]. Understanding sequential coherence helps to identify reply-to relationships and reconstruct the interaction among interlocutors [12].

The sequential organization implied by adjacency pairs is a challenge in computer-mediated communication [4]. The adjacency pairs are regularly disrupted by intervening, irrelevant messages [25], which violates sequential coherence. Berglund [4] argues that participants can create coherence despite the disrupted turn adjacency. Some techniques highlighted in the literature [4,12] are the use of quotations, speaker selection, and lexical relation.

Most research on interactions with chatbots focuses on the relevance of the chatbots' utterance to the associated adjacency pair part [31,37,69], but does not address the miscommunication caused by disrupted turn adjacency [4]. To avoid disruptive messages, Garrido et al. [15] propose a chatbot that asks the users if they want more information in each proposition. Although it apparently reduces disruptive messages, it also decreases interactivity and turns the user into a passive *'yes-no'* responder. Linden et al. [36] propose that the chatbot suggests a solution to a user request and expects that the user will critique it. Then, the chatbot refines the solution based on the user's critique. Although this approach is more likely to maintain interactivity, it also has a highly complex model with databases containing many variables, for instance, multiple sub-domains for travel planning. Both studies focus on two-party conversations. We could not find studies on disruptive sequential coherence that consider multi-party conversations involving chatbots.

In this paper, we expect variance in the users' reactions to having a single (multi-domain) chatbot or multiple (specialized) chatbots introducing new topics, which might be perceived as disruptive from the user's point of view. Hence, we focus on analyzing the disruptive adjacency pairs, guided by the following research question:

RQ3: Does the multiplicity of chatbot as interlocutors influence sequential coherence when the chatbots proactively insert topics in conversation?

METHOD

We performed a Wizard of Oz (WOZ) experiment [8], as done in several other chatbots studies [6,62]. In WOZ experiments, "subjects are told they are interacting with a computer system, though in fact, they are not" [22]. This technique helps designers to consider unique qualities of human-machine communication in initial design stages and to understand users' expectations when interacting with conversational systems [8]. Besides, this method enables to control for some limitations of NLP.

To make wizards sound like chatbots, we simulated a retrieval-based approach [5], where wizards selected responses based on keywords. If the keywords were not in the database, they did not answer. The responses were built to sound like a chatbot (simple vocabulary/compositions), and the wizards were not allowed to adapt them. The same database was used for both scenarios.

We first selected the context for conversation, built the databases, and instructed the wizards. After that, we selected participants who would talk to the supposed "chatbots." We collected the following data: conversation content; researcher's qualitative impressions about participants' behaviors during the interaction; and participants' impressions. The following subsections detail each step.

Selecting the context

We selected a travel-planning context for our experiment. When planning a trip, it is common to look for information online and talk to family, friends, and colleagues [18]. However, it can be difficult to make a decision when there are many options [46], and online travel information searches can be overwhelming [34,47]. Previous studies [41,42] showed that people apply social rules and expectations to computers. Therefore, chatbots could serve as word-of-mouth travel information providers, helping visitors find useful information just by asking, rather than by combining and filtering information from heterogeneous sources, which are increasing with the rise of smart cities [10].

Building the knowledge database

Since the experiment involves exploring a destination, we chose a tourist city: Natal, Brazil. We selected three domains for the conversational agents—nature (the coast, beaches, national parks, etc.); culture (architecture, museums, theaters, and history); and shopping (handmade crafts, malls, and souvenirs)—and built a database for each domain. We collected recurrent information from multiple travel websites (e.g., TripAdvisor) and organized it in pairs of keyword-responses in a spreadsheet, which the wizards searched to find an appropriate statement and copy it to the chat.

Training the wizards

We selected three senior undergraduate students from a Computer Science major to work as wizards. They accepted a confidentiality agreement and underwent two training sessions. Each wizard was assigned one domain; hence, they manipulated only one database. When acting in a *multi-chatbots interaction*, each one interacted, representing one conversational agent. When acting in a *single-chatbot interaction*, only one wizard actively talked to the participant, but three wizards worked together to provide answers, following the same pattern as the other group.

The wizards were trained to answer a user's proposition (whether it was addressed to them or not) only if it contained keywords from their database. When the proposition was addressed to a particular chatbot, the wizard should always answer, and use a default error message when the database had no related answer. In addition, they could proactively

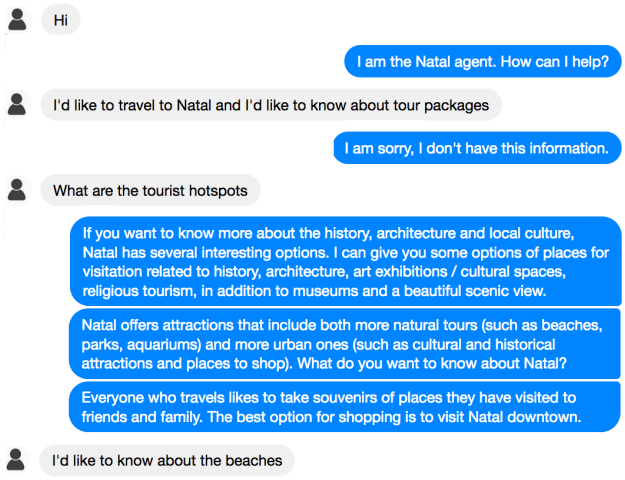

Figure 1. Interaction between P01 and Natal Agent

insert new topics when either they did not know how to answer a question (for example, after a default error message) or keywords related to their knowledge base showed up in the conversation. In the last case, the goal was attracting attention to an aspect that the agent knew about. When no conversational agent had knowledge about the topic requested by the user, the chatbot which answered the last relevant answer sent the default error message. Our wizards were instructed to handle typos and to avoid sending duplicated messages if users corrected a typo in a subsequent message, similarly to most APIs for chatbot development.

Three main aspects were emphasized to the wizards. First, they were to ensure that the only difference between conversational agents' behaviors in a *single-* versus a *multi-chatbots interaction* was the user's awareness about whether one or more agents talked to them. Hence, in both scenarios, users would receive the same number of responses, originated from different chatbots or from a single chatbot. Figures 1 and 2 depict this characteristic for interactions in both scenarios.

Second, when participating in *a multi-chatbots interaction*, they were to answer a user proposition only if the statement was directed to them or related to their domain. This means that participants should not expect to receive three answers for each request. Finally, wizards would not directly talk to each other. All messages were addressed to the participant.

Participants and groups

Undergraduate students from a Brazilian university were invited to participate. Twenty-four students accepted the invitation and voluntarily participated (18 male and six female). Participants were between 18 and 29 years-old and used Internet every day. Twenty-one participants affirmed that they used chat tools in general (for instance, WhatsApp or Facebook Messenger) every day, while the remaining 3 stated that they used them periodically. Regarding the interaction with chatbots, most participants (21) periodically interacted with chatbots, while only one participant reported

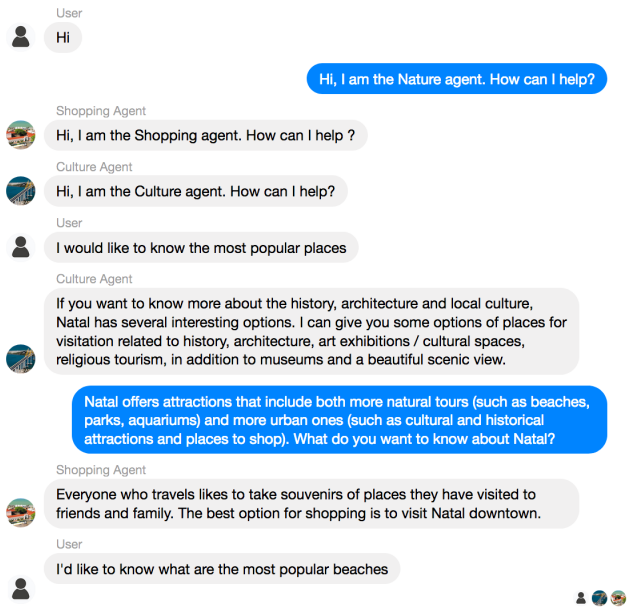

Figure 2. Interaction between P02 and Nature, Culture, and Shopping agents

to use chatbots every day. Two participants had never interacted with a chatbot.

Each participant randomly received an identification number between one and 24; those with odd numbers were assigned to the *single-chatbot interaction* group, and those with even numbers were assigned to the *multi-chatbots interaction* group. The *single-chatbot interaction* group represents the participants who interacted with a single conversational agent, simulating a scenario where a meta-chatbot represents three different specialized chatbots. The *multi-chatbots interaction* group comprises the participants who interacted with three specialized chatbots in a multi-party chat.

The experiment

The experiment was conducted in a lab in which each participant chatted individually with the chatbots. After a brief introduction to the study, the participant was invited to connect to a Facebook account. The participant could connect to his/her own Facebook account or use an account provided by the researchers, as his/her preference. Once they logged in, the researcher invited them to open a Facebook Messenger chat with the chatbot(s) they should interact with. During this step, the researcher introduced the chatbot (or multiple chatbots) and its/their expertise. A printed task description was provided for reference. The task stated that the participant should "talk to the chatbot(s) to decided where they should go" in their first day in the destination.

The researcher invited the participants to "think aloud," so we could collect qualitative impressions about the interaction. While participants chatted with the chatbots, a researcher remained in the lab at a reasonable distance; this was important to foster participants to talk. Some participants did not talk, and they were not pressured to do

so. In this case, the researcher remained distant and silent, so that the participant did not feel troubled by her presence.

Each interaction lasted from 7 to 14 minutes. We did not control the duration, the interaction protocol, or the kind of information the participants search for, because we expected them to complete the task as close to a real interaction as possible. The only restriction was that the source of information to complete the task should be the chat with conversational agents.

The researcher in the lab expected to observe how the participants reacted to the interaction and whether they felt that the task had been successfully completed. In some interactions, the researcher conducted a short debrief session to collect additional data and to clarify doubts about the participant's reactions.

DATA ANALYSIS

We analyzed interactional coherence following the method proposed by Herring [24]. Herring observed that communication characteristics such as the lack of simultaneous feedback and disrupted turn adjacency have an impact on interactional coherence for task-oriented conversation; hence, it helps to compare whether including multiple chatbots as interlocutors in the same chat turns the conversation into a more disruptive scenario. We performed a qualitative analysis of the data collected—textual conversation, observation, and users' impressions reported during debrief sections—combined with quantitative evidence gathered from the conversation structure analysis.

We analyzed interactional coherence in terms of turn-taking and sequential coherence [25], two aspects frequently studied in the discourse analysis literature [25,52,54]. Regarding **turn-taking**, we analyze how the organization of the conversation in terms of turns-to-talk differ from using a single and multiple chatbots in the same chat. We also analyze whether the gap between user messages and conversational agent responses increases overlap [25]. As in a text-based interaction, we do not have "simultaneous talk;" rather, we analyze overlap by looking at *entangled* messages, which are distant from their response by more than one proposition [29,65].

Regarding **sequential coherence**, since we expect the chatbots' utterances to have similar relevance, we focused on investigating whether the frequency of disruptive sequences varies between groups when conversational agents proactively insert new topics.

Finding and coding references

For the first step, we applied Dynamic Topic Analysis (DTA) [26] to code the conversation. We applied this coding style because it adheres to Herring's definition of interactional coherence and provides a visualization tool for the conversation structure analysis [24]. We coded the conversation to identify: role (Chatbot or User); topic; adjacent pairs (which proposition X the proposition Y refers to); message type (the implied action in the proposition: *greetings*, *given option*, *general information*, *request*, *option choice*, *error*, and *user/bot insert topic*); average of the semantic distance (to what degree the proposition is related to its associated adjacent part pair; see [26] for details); and comments.

We focused our analysis on the *User/Bot insert topic* type to determine how these messages impacted sequential coherence. Users/bots were considered inserting a topic when their proposition shift the topic being discussed. When the adjacency pair (proposition-response) was sequential, coherent, and semantically related, *comments* received an *NA* value. The other possible codes were *ignored* (when the proposition has no further responses—there is no "second pair part" to the proposition), *misunderstanding* (when the subsequent proposition is semantically distant), *entangled* (when the proposition is distant from its adjacent-pair by more than one proposition), *repetition* (when the proposition repeats previous proposition), and *modification* (when the proposition repeats the idea of a previous one, but with different words). We focused on *entangled* messages to understand overlap in turn-taking, and on *ignored* messages, which support the analysis of sequential coherence.

Building Discourse Structures

We summarized data according to the metrics discussed in the DSA process. We classified each proposition into one discourse element type [29]: *seed* (a proposition that begins a new thread), *chain* (a message with one reply), *fork* (a message with multiple replies), *tail* (a message inside a thread with no further replies), and *isolated* (a message without reference or replies).

Analyzing Discourse Structures

We performed a quantitative analysis to measure wizards' behavior to verify whether they presented equivalent behavior in both scenarios. We measured the frequency of both the chatbot's *seed* messages and the average of semantic distance per conversation per group.

For turn-taking organization, we considered mostly qualitative evidence, focusing on the textual conversation and users' reports. We used quantitative measures to compare the overlaps in each group, represented by the distance between a proposition and its response. In contrast, for sequential coherence, we measured the frequency of chatbots' *seed* messages *followed* (creating a new chain in the conversation structure) or *ignored* (creating a disruptive sequence) by the user. We also measured the frequency of *ignored* topics *somehow discussed* and *never discussed* per conversation.

The qualitative analysis also revealed other relevant results, and we report them in a separate section, since they are not guided by the presented research questions, but have important implications for design.

To convert the corpora in 24 numerically-comparable observations, we normalized each metric as the frequency of occurrences inside the conversation and compared the mean

of the proportions between groups for each variable. We report numerical significance of comparisons using Student's t-test (for normally distributed data, checking normality with the Shapiro-Wilk test), or an equivalent simulation-based test (for non-normally distributed data, using a permutation test with 10,000 replicates). We consider a p-value < 0.05 as statistically significant.

FINDINGS

We analyzed a total of 781 propositions, divided into 366 messages in *single-chatbot* (30.50 messages per conversation on average, σ=6.11), and 415 messages in *multi-chatbots interactions* (34.58 messages per conversation on average, σ=7.81). We expected *single-chatbot* interactions to be smaller than *multi-chatbots interactions*. However, conversations did not significantly vary in length (t=1.43, df=20.80, p-value=0.17). If we remove greetings messages (all the three conversational agents say *"hello"* at the beginning and *"you are welcome"* at the end of conversations), the average of all propositions is close across the groups: 318 for *single-chatbot* (26.50 on average, σ=6.11) and 319 for *multi-chatbots* interactions (26.58 on average, σ=7.80).

Our findings are split into three subsections: turn-taking processes and sequential coherence, which were guided by the general research question; and other general findings, which came to light during qualitative analysis and have important implications for conversational agent design.

Turn-taking process

We investigated whether the participants of the two groups behaved differently in terms of the turn-taking process, answering **RQ1**. We found out that people behave differently when interacting with a single or with multiple chatbots. As expected, in the *single-chatbot* interactions, the participants naturally alternated turns with the chatbots. However, in *multi-chatbots* interactions, we observed that users failed to organize turn-taking, and some of them even expended some effort trying to define a turn-taking protocol.

In a *multi-chatbots* scenario, some participants tried to use the chatbots' names to assign the next turn and expected to receive only one response. For instance, when M-P04 said "*Hello*" and three chatbots answered the greeting, he said aloud: "*Who should I talk to?*", and wrote "*Nature Agent, which tours do you suggest?*" He complained when other chatbots answered his utterance and ignored all chatbots' attempts to speak when he did not call them nominally, resulting in a rather linear conversation. In contrast, some participants expected that all chatbots would answer all propositions, presuming that they would have one-to-one conversations with each chatbot, but at the same time. These participants expected alternate turns with the chatbot, where the "first pair part" in an adjacency pair would result in three different "second pair parts." For instance, at the beginning of the conversation, M-P20 waited until all the conversational agents had answered. She was surprised when she asked about hiking and the Shopping Agent did not answer. She said "*Nice! It didn't answer because hiking is not a shopping tour!*" During the interaction, M-P20 only sent *"I want to…"* propositions (e.g. *"I want to eat," "I want to listen to music,"* and so forth), expecting to know which suggestion each chatbot would provide. Participant M-P12 said aloud, *"There are only two of them answering me,"* after receiving some messages from Nature and Culture chatbots. Then, when the Shopping chatbot sent a message (related to the latest Culture chatbot message), he said: *"Besides being late in the conversation, he said something unrelated to my question."*

However, in a *single-chatbot* scenario, none of them complained when the conversational agent inserted different topics. For instance, when S-P01 asked about "*touristic spots,*" the conversational agent sent three responses: the first one about culture, the second one about nature, and the third one about shopping (as would be in a *multi-chatbots* set). Nevertheless, S-P01 did not seem uncomfortable with that; he simply chose one topic, ignored the other two, and continued the conversation.

Notably, conversational agents' behaviors did not change between both scenarios, particularly regarding the insertion of *seed* messages. Comparing the proportion of chatbots' *seed* messages per conversations, we observed that, on average, 17% of messages were chatbot's seeds in single-chatbot scenarios, versus 15% in multi-chatbots scenarios, σ=0.06 for both scenarios (t=0.63, df=21.93, p-value=0.53). Therefore, the difference in the conversation only related to the users' awareness about how the conversation took place. Even so, users behaved differently regarding the turn-taking process in each scenario.

> We conclude that the multiplicity of chatbots as interlocutors influences the importance of establishing a perceived protocol for turn-taking. In *multi-party interactions*, a well-defined turn-taking protocol guides users on negotiating **who** should have the next turn (or with whom they should talk first), rather than which topic they should talk about first.

The second aspect we investigated relates to the impact of the gap between users' messages and conversational agents' responses (**RQ2**). Our first observation is that response-time is an issue. Curiously, the problem was mainly reported in *single-chatbot interactions*, in which most participants complained about delayed responses, even when the delay was only a few seconds. This probably happened because with *multi-chatbots* the delay was mostly for the first message. For *single-chatbot*, there was an additional delay for each message, since only one wizard had an active role.

Users explicitly reported that they expected the response to appear immediately when they sent a message. For instance, during his interaction, S-P09 said aloud "*It is taking a while… did it crash?*" In the debrief section, when asked if something annoyed him during the interaction, S-P11 reported that it is annoying when conversational agent "*took a while to respond.*" Participant S-P21 even started to use her

smartphone while waiting for the chatbot's answer. No participant in the *multi-chatbots* scenario mentioned or reacted to the delay.

Although users complained about delay more frequently in *single-chatbot interactions*, we found no significant difference in overlap between groups (t=1.70, df=20.72, p-value=0.10). When measuring *entangled* messages, the mean distance between a statement and its response per conversation was, on average, 1.73 (σ=0.48) and 2.03 (σ=1.95) messages in *single-* and *multi-chatbots* scenarios, respectively.

We conclude that response-time was less annoying in the *multi-chatbots interactions*, but we lack enough evidence to state that the multiplicity of chatbots caused more *entangled* messages, because in both scenarios the common behavior was to wait for answers rather than start another thread.

Sequential coherence

Regarding sequential coherence, we analyzed the effects of the conversational agent's active role. Since the only difference between both scenarios should be the number of interlocutors, we do not expect conversations to be semantically different. In fact, the average of the semantic distance per conversation between groups was not significant (t=0.06, df=19.82, p-value=0.96).

Therefore, we explored sequential coherence from the users' perspective by analyzing whether the chatbots' initiative to insert new topics disrupted or enriched the conversation sequence (**RQ3**). To answer this question, we analyzed what happens in a conversation when a chatbot sends a *seed* message (looking at the *Comment* aspect in the conversation structure analysis). Analysis of each *seed* message suggested that users *followed* or *ignored* the *seed* message, but the proportions of *seed* messages *ignored* by users were considerably large in both groups, as presented in Table 1. There was no statistically significant difference between the proportion of *ignored seeds* per conversation (t=0.92, df=21.82, p-value=0.37) or the proportion of *followed messages* per conversation (t=-1.02, df=21.70, p-value=0.32) between groups.

Interactions	*Ignored* chatbots' *seed* messages		*Followed* chatbots' *seed* messages	
	Mean	σ	Mean	σ
Single-	0.65	0.18	0.33	0.18
Multi-	0.72	0.16	0.27	0.16

Table 1: Mean of the frequency of chatbots' *seed* messages tagged as *ignored* and *followed* per conversation.

In *single-chatbot interactions*, users ignored on average 65% of the chatbots' *seed* messages and, in *multi-chatbots interactions*, they ignored on average 72% of the chatbots' *seed* messages. *Seed* messages received at least one response from a user (users followed a chatbot' *seed* message) in only 33% and 27% of cases, for *single-* and *multi-chatbots interactions,* respectively. In *single-chatbot interactions*, two messages were classified as dismissed, because S-P01 did not *ignore* chatbot's propositions, but did not *follow* the topic either. For instance, he answered a *seed* message with a *"No, thank you"* message. In *multi-chatbots interactions*, users did not send any dismissing messages. The *seed* messages with no answers reduced the sequential coherence of the conversation [25], regardless of the multiplicity of chatbots as interlocutors.

After noticing the large number of ignored messages, we devised a sub-research question to investigate whether the phenomenon was related to the topics being discussed: **RQ3.1:** `Is the seed message's topic related to the reason why users ignored it?`

When we looked at the topic of *ignored seed* messages, we observed that most topics were somehow discussed throughout the interaction, as summarized in Table 2. Two different phenomena explain this observation:

1. ***Useless messages****: ignored seed* message sometimes referred to a topic that had already been discussed; for instance, S-P07 ignored *seed* messages of three different topics (Culture, Shopping, and Nature), but when he did, he had already chatted about all those topics and probably already had enough information;
2. ***Intrusive messages****: ignored message* topics emerged later in conversation; in this case, chatbots inserted a topic that was not at first followed by the user, but was later requested by the user when another topic was perceived as finished. For instance, M-P02 ignored *seed* messages about Culture three times; later, she started a new topic asking about "*historical places.*" Sometimes, users followed a *seed* message the second (or even the third) time the chatbot inserted it, probably because it was the perceived proper time to start it. In this case, users are not simply ignoring messages, but rather sorting their interests.

Interactions	*Ignored* topics somehow discussed		*Ignored* topics never discussed	
	Mean	σ	Mean	Σ
Single-	0.63	0.32	0.38	0.33
Multi -	0.58	0.33	0.37	0.33

Table 2: Mean of the frequency of *ignored* topics *somehow discussed* and *never discussed* per conversation.

Table 2 shows that the frequency of *ignored topics* that were tagged as *never discussed* is around 38%, regardless of the scenario. The frequency of *ignored topics* tagged as *somehow discussed* correspond to 58% and 63% in single- and multi-chatbots interactions, respectively. Another intriguing fact is that some users explicitly reported not knowing what else to ask for both scenarios (S-P09, S-P11, S-P15, M-P02, M-P22) after a few minutes of interaction. We also noticed that some users did not follow the *seed* messages, but they drew ideas from chatbots' messages about what else they could ask. For instance, the following excerpt was extracted from M-P16's interaction:

M-P16: *What is the closest place to shop?*
Shopping agent: *[general information about Midway Mall]*

Shopping agent: *In Capitania das Artes you can buy products from local artists in the Artist's Store.*
Culture agent: *The Artist's Store is located at Capitania das Artes, headquarters of the city's Cultural Foundation, which hosts contemporary exhibitions of native artists. (...) The historic building and the beautiful view of the Potengi River are worthwhile.*
M-P16: *Which is the nearest museum?*

In this excerpt, the Culture agent was not introducing a new topic, but rather it was pointing out more information about the place where Artist's Store is located. Although conversation is still focused on shopping, the Cultural agent intervention likely inspired the user to ask about museums, since it talked about *"historic buildings"* and *"exhibitions open to the public."*

Once more, the frequency of *ignored topics somehow discussed* (p-value= 0.75, simulation-based test) and *ignored topics never discussed* (p-value=1, simulation-based test) per conversation is not significantly different between groups; hence, we do not have evidence to affirm that the multiplicity of chatbots as interlocutors influenced these behaviors.

We conclude that, regardless of the number of chatbots involved in the conversation, *seed* messages disrupt sequential coherence; but, they also might provide insights about conversational agents' knowledge and how to explore it. However, it is important to define an interactional protocol that establishes the relevant transition point to insert them, reducing the substantial number of disruptive and *ignored* messages.

Other findings

This section presents other findings that came to light during the qualitative analysis and have implications for design.

Don't talk too much: several participants expressed their frustration about long messages. Some conversational agents' responses comprised more than 5 lines of text (presenting information about a specific place or describing options). Participants have complained about the time they spent reading those messages. In addition, conversational agents must identify when conversations should end. Some participants in our experiment explicitly declared a decision. For instance, the following propositions were written by S-P19 and M-P04, respectively: *"Ok, I think I will go to the museum"* [S-P19]; and *"Ok, I will visit Morro do Careca"* [M-P04]. These propositions stated a decision. However, the wizards had been instructed to end conversation only when the user sent a greeting for ending. So, after those propositions, users received responses giving information about museums or the site. Sometimes, these propositions were repetitions or additional information not requested by the user. In both cases, at the end of the conversation, users reported that they were confused or undecided.

Human likeness: although participants had similar profiles (see Participants and Groups section), the way they perceived the conversational agent differed. Some participants were extremely polite (using more formal vocabulary and interjections to demonstrate interest and engagement, apologizing, and so forth). For instance, the following sentences are all propositions sent by participant S-P19: *"I may have expressed myself wrong." "Could you help me with that?" "Yes, I would like to know more about Museum of Sacred Arts." "Alright, you've said that before."*

On the other hand, some participants (most of them pursuing Computer Science majors) talked based on keywords. For instance, M-P20 based his conversation on *"I want to..."* questions. Participant S-P13 sent several propositions with only one word (such as *"food," "fairs," "pubs"*). Also, some participants assumed that chatbots would not understand typos. S-P21 rewrote at least two of his own propositions, due to typos). S-P09, M-P12, and S-P13 sent question marks as a subsequent proposition, assuming that chatbots' would not know the previous proposition was a question. Curiously, some participants referred to the conversational agent (all occurrences in the *single-chatbot* set) using a gendered pronoun.

Dealing with language complexity: since people hold highly complex conversations; conversational agents should be able to understand and deal with some level of complexity. During our analysis, we observed three types of complexity problems: multiple topics in one proposition, restrictions, and comparisons.

Sometimes users talked about two different topics in a single proposition. For instance, M-P12 started his interaction with the following proposition: *"I would like to travel to a place with a beach and with good cultural and accommodations options"* [M-P12]. For a human, this proposition is easily broken into three different parts (*beaches*, *culture*, and *accommodation*), and a travel agent probably would address all those topics during conversations, without requiring the user to specify them again. However, when the conversational agents responded propositions with this kind of structure, they either chose one topic or sent a few messages in a row, one per topic they found in the proposition, which increased the number of ignored messages (M-P12 received two responses in a row–from the Nature and Cultural chatbots–and ignored both, subsequently starting a new topic). Regardless of the response strategy, conversational agents missed the users' request trace, and, unless users asked again, they did not address the referred to, but undiscussed, topic.

Additionally, users imposed restrictions. For instance, S-P11, who was a participant with a *Working* task, asked: *"Which is the best touristic spot to visit in the morning?"* The conversational agent responded to this proposition with information about touristic spots but ignored the *'in the morning'* restriction. Another example of restrictions was provided by M-P18, who tried to lead chatbots to recommend a place with two combined features: natural/local landscape and shopping. However, chatbots responses always addressed only one feature at a time.

Restrictions can also be defined when users send a question as a response to a previous question. For instance:

Nature agent: *"Would you be interested in visiting Pipa Beach at night?"*
M-P18: *"Pipa Beach has stores?"*

M-P18 intended to answer the questions, but her answer depended on whether the beach has stores. However, the user did not receive a proper answer, since the agent talked about stores and Pipa, but did not exactly relate the two features.

Finally, many users wanted answers to evaluative questions, which require subjective responses, such as *"which is the best beach in Brazil?"* [S-P05], *"which are the best options in this city?"* [S-P19], *"Why should I visit Natal?"* [S-P21]. The "best" beach or options would depend on the individual's interests. The reason why S-P21 should visit Natal is probably not the same as why someone else should.

DISCUSSION

In this section, we discuss our findings and their implications for design. We situate the findings in terms of the literature, providing insights for chatbots research and design.

Turn-taking: our results show that people defined a perceived turn-taking protocol only in *multi-chatbots* scenarios, although chatbot behavior regarding inserting new topics was similar in both scenarios. Psychology studies of attention argue that exposing people to more than one stimuli requires cognitive effort to focus attention [11,55]. In *single-chatbot interactions*, users maintained their attention to the only agent talking to them, alternating turns, as suggested in the conversation analysis literature [52]. They probably received different topics as options and naturally used their next turn to choose an option and steer the conversation toward their interests. However, in *multi-chatbots interactions*, the fact that there were three agents gave rise to confusion, especially at the beginning of the conversation, when users first received three different stimuli (see Figures 1 and 2). They had to decide not only which topic to discuss, but also which agent to answer. To do so, participants tried to formally select the next speaker, a commonly adopted turn-taking protocol in human-human interactions [16,52,54], or simulate one-on-one conversations with all the three chatbots at the same time, alternating turns with them. Moreover, reported confusion and observed need to establish a turn-taking protocol can also be related to human-likeness. Since all the chatbots aimed to help the user, participants might feel uncomfortable about not giving them an opportunity to speak. In human-human interaction, it could be considered rude "to exclude some person from a conversation or to withhold a response to what someone else has said and in this way to ignore them" [59]. Bayser et al. [3] delegate the turn-taking organization to a moderator chatbot, which invites other chatbots to the conversation. When this moderator is not available, the chatbots must recognize the right moment to speak, especially when the utterance inserts a new topic.

Regarding gaps between messages, we could not relate delays to the occurrence of *entangled* messages (mostly because our participants did not start another thread, but rather waited for answers). This user behavior probably relates to the claim that pauses help to manage turn-taking in computer-mediated communication [2,48], in contrast to the "no gaps, no overlap" in spoken conversation [52]. When participants sent a message, the pause meant that the turn was complete and the chatbots should answer. However, the delay was too large from participants' perspective. In this sense, our findings reiterate previous literature in human-computer interactions [32,33,61], where the outcomes on responsiveness highlight emotional, psychological, and behavioral effects. Design strategies could involve consolidated awareness techniques [30,38], which would inform the user that the chatbot is "thinking," or even "distractor" messages (e.g., "Let's see what I can suggest to you"), which would keep users engaged in the interaction.

Sequential coherence: Sequential coherence analysis showed that conversational agents should know the proper time to insert a new topic, so as not to disrupt conversation with unwanted messages, as discussed by Herring [25], especially when more than one chatbot is involved. In a *multi-chatbot* interaction, when more than one chatbot inserts new topics, users may understand that the agents are self-selecting a turn [2] and starting a new thread. As discussed in the previous section, users found this confusing and disruptive, as also observed in other contexts [68]. Conversational agents must be designed with a well-defined protocol that enables them to hold interactive conversation and not overload users with worthless information.

However, findings also show that the proactive behavior helps users learn to explore the database. Since some users reported not knowing what to ask after a few minutes, it is important to find ways to engage users in new topics and avoid prematurely ending the conversation. Uthus and Aha [65] underscored the necessity of developing chatbots capable of participating actively in a chat. The solution presented by Garrido et al. [15] presents more options from the database for each proposition, guiding the user to explore it, but it also decreases interactivity and turns the user into a passive *'yes-no'* responder. The approach posited by Linden et al. [36] is more likely to maintain interactivity; however, it has a highly complex model for multi-domain chatbots. Thus, a protocol that maintains the sequential coherence and interactivity in a multi-domain scenario is an open challenge.

Don't talk too much: according to Woodburn et al. [67], the frequency of turn exchange in chatting is high when compared to face-to-face communication, producing an intermittent exchange [2]. In addition, Hill et al. [27] state that human-agent interactions last longer than human-human interactions, but messages are shorter when chatbots are involved. Conversations that include chatbots should, therefore, be structured toward shorter messages and more frequent turn alternation [2,27]. However, as discussed in the

sequential coherence topic, it is also important to foster the user's curiosity to avoid prematurely ending conversation. Thus, conversational agents should provide valuable details in shorter messages.

Human likeness: we observed that users hold different perceptions about conversational agents' *human likeness*. The literature has demonstrated concerns about how people's behaviors differ when they interact with other humans versus with artificial agents [27,40,49,57]. Results show different behaviors in agreement, confidence, conscientiousness, and other issues. When helping people to make decisions, it is important to inspire trust. Likely due to poor prior experiences, some participants that said they had interacted with conversational agents before revealed that they did not believe that the chatbots would give them appropriate information. Previous work has also shown that it is important to predict users' perceptions about human-likeness and adapt agents' behaviors to meet users' expectations [35]. Hence, research on psychological and behavioral aspects related to users' satisfaction and technology acceptance when interacting with an artificial agent is relevant and should be considered for interaction design.

Dealing with language complexity: although context is already highlighted by conversational agents literature [19,53,64], and the NLP field has overcome several challenges [28], problems related to discussing many topics in the same sentence and making restricted requests remain open challenges. Although APIs for chatbot development already provide solutions to parse requests with multiple topics, this poses challenges to integrate already available, specialized chatbots. The combination of specific outputs from different chatbots into a single, consistent response would require intelligence on the meta-chatbot or the adaptation of the multi-chatbots, which may not be possible. Regarding comparative questions, we believe that profile information could help. Conversational agent design could take advantage of users' personal information about preferences and interests to match those with places' characteristics and provide useful comparative information. For instance, a useful answer to S-P21's question about the best beach, the agent could respond, *"Because Natal has beautiful beaches with great waves where you can surf,"* provided the agent knew about S-P21's interest in the sport.

Threats to validity

Internal threats mostly relate to the database and subjectivity bias. Although our goal was not to validate the knowledge base, inconsistency or lack of information might cause frustration. To address this threat, we collected information from different and accredited data sources and only inserted information confirmed by more than one source. However, since qualitative analysis involves subjectivity, researchers' judgments might introduce a bias in the results.

External threats relate to participants' representativeness: our participants were all students around the same age. Furthermore, because of the characteristic of the university, all the participants had background in STEM and the number of males and females were unbalanced. Thus, results cannot be directly generalized, and more experiments should be performed considering other profiles. Besides, the small number of participants prevents quantitative analysis from providing conclusive results. Quantitative results should be considered as indicators that help to support our conclusions.

We used a Wizard of Oz approach to control for limitations of natural language processing, inspired in previous chatbots studies [6,62]. However, it is not possible to the wizards reproduce computational power, which introduced a time delay in both scenarios. Nevertheless, sequential coherence issues, such as confusion, highlighted in our results were more evident on multi-chatbots scenarios, where the participants did not complain about response time.

Finally, we recognize that human-chatbot interactions are a multi-faceted problem, and there are several conversational aspects where single- and multi-chatbots interactions may differ. We focused on interactional coherence because it may influence the success of task-oriented conversations [25]; hence, we focused on the measures that directly influence turn-taking and sequential coherence. We plan future replications of this study as an opportunity to investigate other conversational aspects and setup variances, such as the use of real chatbots; different number of interlocutors; the impact of repetition, misunderstanding and modification messages; theme; task; and participant profiles.

CONCLUSION

This paper compared users' behaviors when interacting with single or multiple chatbots to gather information for decision-making. Human participants reported different communication experience, but the conversation structure remained similar between conditions. More specifically, users reported feeling more confused in multi-chatbots interactions, where they tried to adopt strategies to organize turn-taking. Also, we found that the conversational agents' proactive behaviors disrupt sequential coherence, but also help users to explore a knowledge database; hence, proactive behavior should be developed with a well-defined protocol that reduces disruptive messages.

Our results suggest that when designing a conversational agent to help users make decisions, there is no evident reason to split the knowledge amongst more than one agent. Therefore, it is possible to leverage already available conversational agents and combine their knowledge to build a richer database. In this case, designers can choose between: a meta-chatbot user interface that orchestrates multiple agents in the background and overrides the multi-party aspect of interaction; or a multi-chatbot user interface with a well-defined interactional protocol that reduces stimuli overload. In both cases, designers should consider the importance of providing clues on how to explore chatbots' knowledge, but also avoid sequential coherence disruption caused by unimportant messages.

REFERENCES

1. Pelumi Aboluwarin. 2016. Chatbots—Igniting Division of Labour in AI. Retrieved April 7, 2017 from https://businessofmessaging.com/chatbots-igniting-division-of-labour-in-ai-1430fcc85c8d.
2. Jeffrey F Anderson, Fred K Beard, and Joseph B Walther. 2010. Turn-Taking and the Local Management of Conversation in a Highly Simultaneous Computer-Mediated Communication System. *Language@Internet* 7, 7.
3. Maıra Gatti de Bayser, Paulo Rodrigo Cavalin, Renan Souza, Alan Braz, Heloisa Candello, Claudio S. Pinhanez, and Jean-Pierre Briot. 2017. A Hybrid Architecture for Multi-Party Conversational Systems. *CoRR* abs/1705.01214. Retrieved from http://arxiv.org/abs/1705.01214
4. Therese Örnberg Berglund. 2009. Disrupted turn adjacency and coherence maintenance in instant messaging conversations. *Language@ Internet* 6, 2.
5. Parminder Bhatia, Marsal Gavalda, and Arash Einolghozati. 2017. soc2seq: Social Embedding meets Conversation Model. *arXiv preprint arXiv:1702.05512*.
6. Heloisa Candello, Claudio Pinhanez, David Millen, and Bruna Daniele Andrade. 2017. Shaping the Experience of a Cognitive Investment Adviser. In *International Conference of Design, User Experience, and Usability*, 594–613.
7. Justine Cassell. 2000. Embodied conversational interface agents. *Communications of the ACM* 43, 4: 70–78.
8. Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies—why and how. *Knowledge-based systems* 6, 4: 258–266.
9. Robert Dale. 2016. The return of the chatbots. *Natural Language Engineering* 22, 5: 811–817.
10. Auriol Degbelo, Carlos Granell, Sergio Trilles, Devanjan Bhattacharya, Sven Casteleyn, and Christian Kray. 2016. Opening up smart cities: citizen-centric challenges and opportunities from giscience. *ISPRS International Journal of Geo-Information* 5, 2: 16.
11. Jon Driver. 2001. A selective review of selective attention research from the past century. *British Journal of Psychology* 92, 1: 53–78.
12. B.J. Fogg. 2003. Computers as persuasive social actors. In *Persuasive Technology*, B.J. Fogg (ed.). Morgan Kaufmann, San Francisco, 89–120.
13. Asbjørn Følstad and Petter Bae Brandtzæg. 2017. Chatbots and the new world of HCI. *interactions* 24, 4: 38–42.
14. Tianjun Fu, Ahmed Abbasi, and Hsinchun Chen. 2008. A hybrid approach to web forum interactional coherence analysis. *Journal of the Association for Information Science and Technology* 59, 8: 1195–1209.
15. Piedad Garrido, Javier Barrachina, Francisco J Martinez, and Francisco J Seron. 2017. Smart tourist information points by combining agents, semantics and AI techniques. *Computer Science and Information Systems* 14, 1: 1–23.
16. Will Gibson. 2009. Negotiating textual talk: Conversation analysis, pedagogy and the organisation of online asynchronous discourse. *British Educational Research Journal* 35, 5: 705–721.
17. Ben Goertzel. 2014. Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence* 5, 1: 1–48.
18. Google Travel Study. 2014. *The 2014 Traveler's Road to Decision*. Ipsos MediaCT. Retrieved September 1, 2017 from https://storage.googleapis.com/think/docs/2014-travelers-road-to-decision_research_studies.pdf.
19. David Griol, José Manuel Molina, and Zoraida Callejas. 2014. Modeling the user state for context-aware spoken interaction in ambient assisted living. *Applied intelligence* 40, 4: 749–771.
20. David Griol and José Manuel Molina. 2017. Building multi-domain conversational systems from single domain resources. *Neurocomputing*.
21. Nicole Gustas. 2016. Chatbots: Next Big Thing or Cash Grab?. Retrieved April 8, 2017 from https://www.metafilter.com/158975/Chatbots-Next-Big-Thing-or-Cash-Grab.
22. Melita Hajdinjak and France Mihelic. 2003. Wizard of Oz experiments. In *EUROCON 2003. Computer as a Tool. The IEEE Region 8*, 112–116.
23. John Heritage. 2008. Conversation analysis as social theory. *The new Blackwell companion to social theory*: 300–320.
24. Susan C Herring and Andrew J Kurtz. 2006. Visualizing dynamic topic analysis. In *Proceedings of CHI'06*.
25. Susan C. Herring. 1999. Interactional coherence in CMC. *Journal of Computer-Mediated Communication* 4, 4: 0–0.
26. Susan C. Herring. 2003. Dynamic topic analysis of synchronous chat. In *New Research for New Media: Innovative Research Methodologies Symposium Working Papers and Readings*, 47–66.
27. Jennifer Hill, W Randolph Ford, and Ingrid G Farreras. 2015. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior* 49: 245–250.
28. Julia Hirschberg and Christopher D Manning. 2015. Advances in natural language processing. *Science* 349, 6245: 261–266.
29. Torsten Holmer. 2008. Discourse structure analysis of chat communication. *Language@ Internet* 5, 9: 1–19.
30. Raymond B Jennings, Erich M Nahum, David P Olshefski, Debanjan Saha, Zon-Yin Shae, and Chris Waters. 2006. A study of internet instant messaging and chat protocols. *IEEE Network* 20, 4: 16–21.
31. Aditya Joshi, Anoop Kunchukuttan, Pushpak Bhattacharyya, and Mark James Carman. 2015.

SarcasmBot: An open-source sarcasm-generation module for chatbots. In *WISDOM Workshop at KDD*.
32. Christin Kohrs, Nicole Angenstein, and André Brechmann. 2016. Delays in human-computer interaction and their effects on brain activity. *PloS one* 11, 1.
33. Werner Kuhmann, Wolfram Boucsein, Florian Schaefer, and Johanna Alexander. 1987. Experimental investigation of psychophysiological stress-reactions induced by different system response times in human-computer interaction*. *Ergonomics* 30, 6: 933–943.
34. Tania C Lang. 2000. The effect of the Internet on travel consumer purchasing behaviour and implications for travel agencies. *Journal of vacation marketing* 6, 4: 368–385.
35. Min Kyung Lee, Sara Kiesler, and Jodi Forlizzi. 2010. Receptionist or information kiosk: How do people talk with a robot? In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, 31–40.
36. Greg Linden, Steve Hanks, and Neal Lesh. 1997. Interactive assessment of user preference models: The automated travel assistant. In *User Modeling*, 67–78.
37. Daniel Macias-Galindo, Wilson Wong, John Thangarajah, and Lawrence Cavedon. 2012. Coherent topic transition in a conversational agent. In *Proceedings of the 13th Annual Conference of the International Speech Communication Association (InterSpeech), Oregon, USA*.
38. Panos Markopoulos and Wendy Mackay. 2009. *Awareness systems: Advances in theory, methodology and design*. Springer Science & Business Media.
39. Hareesh Maturi. 2016. Meta Chatbot: Enabling collaboration between chatbots. Retrieved April 7, 2017 from https://www.linkedin.com/pulse/meta-chatbot-enabling-collaboration-between-chatbots-hareesh-maturi.
40. Yi Mou and Kun Xu. 2017. The media inequality: Comparing the initial human-human and human-AI social interactions. *Computers in Human Behavior* 72: 432–440.
41. Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues* 56, 1: 81–103.
42. Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 72–78.
43. Daniel Nations. 2017. What is Siri? How Can I Use It and How Can Siri Help Me? A Look at Apple's Personal Assistant for iOS. Retrieved March 21, 2017 from https://www.lifewire.com/what-is-siri-help-1994303.
44. Ahmed K Noor. 2015. Potential of cognitive computing and cognitive systems. *Open Engineering* 5, 1: 75–88.
45. Jacki O'Neill and David Martin. 2003. Text chat in action. In *Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work*, 40–49.
46. Jeong-Yeol Park and SooCheong Shawn Jang. 2013. Confused by too many choices? Choice overload in tourism. *Tourism Management* 35: 1–12.
47. Aneta Pawlowska. 2016. Tourists and social media: Already inseparable marriage or still a long-distance relationship? Analysis of focus group study results conducted among tourists using social media. *World Scientific News* 57: 106–115.
48. Martin Porcheron, Joel E Fischer, and Sarah Sharples. 2017. Do animals have accents?: talking with agents in multi-party conversation. In *ACM CSCW'17*.
49. Andrew B Raij, Kyle Johnsen, Robert F Dickerson, Benjamin C Lok, Marc S Cohen, Margaret Duerson, Rebecca Rainer Pauly, Amy O Stevens, Peggy Wagner, and D Scott Lind. 2007. Comparing interpersonal interactions with a virtual human to those with a real human. *IEEE transactions on visualization and computer graphics* 13, 3: 443–457.
50. Stuart Reeves. 2017. Some conversational challenges of talking with machines. *Talking with Conversational Agents in Collaborative Action Workshop on CSCW 2017*.
51. Microsoft Research. 2014. Anticipating More from Cortana. Retrieved March 21, 2017 from https://www.microsoft.com/en-us/research/blog/anticipating-more-from-cortana/.
52. Harvey Sacks and Emanuel A Schegloff. 1995. *Lectures on conversation: Volumes I & II*. Wiley Online Library.
53. Claude Sammut. 2001. Managing context in a conversational agent. *Linkoping Electronic Articles in Computer & Information Science* 3, 7.
54. Emanuel A Schegloff. 1990. On the organization of sequences as a source of "coherence" in talk-in-interaction. *Conversational organization and its development* 38: 51–77.
55. Brian J Scholl. 2001. Objects and attention: The state of the art. *Cognition* 80, 1: 1–46.
56. Bayan Abu Shawar and Eric Atwell. 2007. Chatbots: are they really useful? In *LDV Forum*, 29–49.
57. Nicole Shechtman and Leonard M Horowitz. 2003. Media inequality in conversation: how people behave differently when interacting with computers and people. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 281–288.
58. Jenna Sheffield. 2016. The Ultimate Travel Bot List. Claire by 30 seconds to fly. Retrieved April 7, 2017 from https://www.30secondstofly.com/ai-software/ultimate-travel-bot-list/#The_Future_of_Travel_Bots.
59. Jack Sidnell. 2011. *Conversation analysis: An introduction*. John Wiley & Sons.
60. Nick Statt. 2016. Why Google's fancy new AI assistant is just called 'Google'. Retrieved March 21, 2017 from

https://www.theverge.com/2016/5/20/11721278/google-ai-assistant-name-vs-alexa-siri.

61. André J Szameitat, Jan Rummel, Diana P Szameitat, and Annette Sterr. 2009. Behavioral and emotional consequences of brief delays in human–computer interaction. *International Journal of Human-Computer Studies* 67, 7: 561–570.
62. Indrani Medhi Thies, Nandita Menon, Sneha Magapu, Manisha Subramony, and Jacki O'Neill. 2017. How Do You Want Your Chatbot? An Exploratory Wizard-of-Oz Study with Young, Urban Indians. In *IFIP Conference on Human-Computer Interaction*, 441–459.
63. David Traum and Jeff Rickel. 2002. Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2*, 766–773.
64. Gabriele Trovato, Massimiliano Zecca, Tatsuhiro Kishi, Nobutsuna Endo, Kenji Hashimoto, and Atsuo Takanishi. 2013. Generation of humanoid robot's facial expressions for context-aware communication. *International Journal of Humanoid Robotics* 10, 01: 1350013.
65. David C Uthus and David W Aha. 2013. Multiparticipant chat analysis: A survey. *Artificial Intelligence* 199: 106–121.
66. Alessandro Vinciarelli, Anna Esposito, Elisabeth André, Francesca Bonin, Mohamed Chetouani, Jeffrey F Cohn, Marco Cristani, Ferdinand Fuhrmann, Elmer Gilmartin, Zakia Hammal, and others. 2015. Open challenges in modelling, analysis and synthesis of human behaviour in human–human and human–machine interactions. *Cognitive Computation* 7, 4: 397–413.
67. R Woodburn, R Procter, J Arnott, and A Newell. 1991. A study of conversational turn-taking in a communication aid for the disabled. *People and Computers*: 359–371.
68. Jun Xiao, Richard Catrambone, and John Stasko. 2003. Be quiet? evaluating proactive and reactive user interface assistants. In *Proceedings of INTERACT*, 383–390.
69. Zhou Yu, Ziyu Xu, Alan W Black, and Alexander Rudnicky. 2016. Chatbot evaluation and database expansion via crowdsourcing. In *Proceedings of the RE-WOCHAT workshop of LREC, Portoroz, Slovenia*.