A Conversation Analysis of Non-Progress and Coping Strategies with a Banking Task-Oriented Chatbot

Chi-Hsun Li, Su-Fang Yeh, Tang-Jie Chang, Meng-Hsuan Tsai, Ken Chen, Yung-Ju Chang
National Chiao Tung University, Hsinchu, Taiwan
{armuro}@cs.nctu.edu.tw, {funing314.iem05g}@g2.nctu.edu.tw,
{sfy.iem07g, tjchang.cs08g, kent31.iem05g}@nctu.edu.tw, {clairetsai818}@gmail.com

ABSTRACT
Task-oriented chatbots are becoming popular alternatives for fulfilling users' needs, but few studies have investigated how users cope with conversational 'non-progress' (NP) in their daily lives. Accordingly, we analyzed a three-month conversation log between 1,685 users and a task-oriented banking chatbot. In this data, we observed 12 types of conversational NP; five types of content that was unexpected and challenging for the chatbot to recognize; and 10 types of coping strategies. Moreover, we identified specific relationships between NP types and strategies, as well as signs that users were about to abandon the chatbot, including 1) three consecutive incidences of NP, 2) consecutive use of message reformulation or switching subjects, and 3) using message reformulation as the final strategy. Based on these findings, we provide design recommendations for task-oriented chatbots, aimed at reducing NP, guiding users through such NP, and improving user experiences to reduce the cessation of chatbot use.

Author Keywords
chatbot; conversation analysis; breakdowns; non-progress; coping strategies

CSS CONCEPTS
•Human-centered computing~Human computer interaction (HCI)~Interaction paradigms~Natural language interfaces

INTRODUCTION
The market popularity of text-based conversational agents (text-based CAs), known as chatbots is growing. Chatbots have been deployed in online platforms in various fields [9], and in 2018, more than 300,000 chatbots were said to be active on Facebook Messenger alone [5]. Task-oriented chatbots in particular are attracting considerable attention because, by focusing on helping users perform specific tasks, they can serve as important alternatives to live customer support, mobile apps and websites. However, the quality of task-oriented chatbots' interaction designs has not kept pace with their rapid growth in popularity; and we argue that this problem can be ascribed chiefly to lack of understanding of how users use chatbots in their daily lives. Various researchers have sought to develop better natural-language processing techniques, or to reduce recognition errors [22,26], since conversation breakdowns can be caused by difficulties with the complexities of natural-language [25].

Researchers have also started to develop guidelines for the chatbot interaction design. For instance, Jain et al. [12] explored how first-time users communicated with several kinds of chatbots and generated a set of guidelines based on the findings, and Ashktorab et al. [4] studied which strategies users prefer chatbots to adopt to repair conversation breakdowns. However, the resulting guidelines have thus far been based on studies in which the participants were given specific interaction instructions or scenarios. Therefore, their uses of chatbots were not driven by their own day-to-day needs, and the realism of the obstacles to human-chatbot interaction reported in these studies remains uncertain. Likewise, unknown are the frequency of these obstacles, how users deal with them, and which of them are most likely to cause users to break off communication with a chatbot. We argue that obstacles to conversation, or the non-progress (NP) of a conversation, between a human and a task-oriented chatbot are just as important to address as improving the usability of a website or mobile app. Moreover, it might be possible to anticipate NP and prioritize it for repair if we have a better idea about when and how often it occurs, and which subtypes of NP are more likely to lead to repair by users vs. abandonment of the conversation.

To answer these questions, we conducted conversation analysis on a three-month conversation log involving 1,685 users' 17,136 conversational exchanges with a chatbot maintained by one of the top digital-banking institutions in Taiwan. We focused on a task-oriented banking chatbot because financial services has thus far led other industries in its use of bots and artificial intelligence more generally [1]. From observing extensive real-world use of this chatbot, we hope to answer the following research questions: 1) What are the categories of frequent NP? 2) What strategies for coping with NP do users adopt most often? And 3) Are there specific conversational clues that a user is likely to abandon his/her dialogue with a task-oriented chatbot?

As well as practical design recommendations based on its findings, this paper provides three major contributions. First,

CHI 2020, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04...$15.00.
DOI: https://doi.org/10.1145/3313831.3376209

it presents the first data-driven typology of NP, comprising 12 observed subtypes, and identifies five types of content that are unexpected and challenging for typical chatbots to accurately recognize and thus to respond to. Second, it identifies 10 distinct strategies users employ to cope with NPs, and close links between some of those strategies and specific types of NP, including those that often cause users to abandon communication with a chatbot. And third, it delineates signs that users are about to cease their conversations with a chatbot, including three consecutive NPs; repetitive use of same kinds of coping strategy; and message reformulations as the last coping strategy.

RELATED WORK

Researchers have studied how users interact with CAs [7,16,17,20,21,30,32], including chatbots. Some have discussed how user experience might be improved through enriching the personalities of voice-user interfaces (VUIs) and chatbots [7,16,17,21,24], while others have studied factors that affected users' preference of voice and text input [11,32]. Meanwhile, people's willingness to use CAs were found to be influenced by several variables, such as systems' low reliability and users' poor mental models of what a VUI is capable of [20], and conversation breakdowns during the interaction with chatbots [4,12]. However, communication failure due to chatbots' difficulties with handling natural language are still commonplace [3,25].

Various researchers have attempted to address CAs' failures, some of them by focusing on how they repair breakdowns [8,15]. For example, Lee et al. [15] found that users' mental schema regarding service have an impact on their recovery-strategy preferences. Users with more relational outlooks tend to be more satisfied with apologies for mistakes, whereas those with more utilitarian orientations prefer compensation. Ashktorab et al. [4] studied the strengths and weaknesses of repair strategies implemented by a particular chatbot, and found that most study participants preferred that it provides a few guesses and let them decide which one is correct. And Weisz et al. [31] addressed communication breakdowns by teaching users about chatbots and developing their empathy toward them. However, hardly any research has examined the relationship between task-oriented chatbots' conversational breakdowns and the strategies their users adopt to deal with such problems. The two studies most relevant to the present one both focused on VUI. Jiang et al. [13] studied how users reformulated their information requests to VUIs. Because VUI breakdowns were usually due to missing or substituted words, reordering words or changing phonetics could help correct input errors. More recently, Myers et al. [23] identified five forms of conversational breakdown and 10 coping strategies. Of these, hyper-articulation was the most frequently used, and quitting the task was the least.

Probably due to the essential differences between task-oriented chatbots and VUIs, and/or between the datasets that were used, our conversation analysis yielded both a different set of NPs and different coping strategies for dealing with them than either of the aforementioned studies.

DATA CHARACTERISTICS

The banking institution used the Facebook Messenger and LINE Messenger platforms to build its banking chatbot. The dataset we analyzed contained 2,597 users' conversations with the Facebook Messenger chatbot, recorded from May 1, 2017 to July 31, 2017. The data were stored in a spreadsheet, with each row representing one of the dataset's 24,074 exchanges: i.e., one user input followed by one output from the chatbot. In each case, the intent of the user input was analyzed using the IBM Watson conversation understanding service, which also calculated and recorded a level of confidence in its recognition of that intent (Figure 1). To interact with the chatbot, in addition to typing, which was always allowed, the chatbot sometimes provided alternative input modalities based on its assessments of user intentions, including buttons on a card and quick-response buttons (see Figure 2). Because the chatbot did not ask users to provide sensitive personal information such as their usernames, ID/account numbers, and so forth, they could only ask it for general information unrelated to their identities. Accordingly, the services it offered included a currency-exchange converter, introduction to credit cards, a housing-loan evaluator, and investment information. First-time users of this chatbot were informed about our data-collection activities and their purposes, e.g., "Investigation, statistics and research analysis". After seeing this message, the user had to click on an "Agree" button to continue their use of the chatbot. Each unique user had had a different number of interaction sessions with the chatbot; the average number of sessions was 1.7 (SD=2.5, Min=1, Max=52).

DATA CLEANING AND CONVERSATION ANALYSIS

The data was cleaned three times, once prior to coding and twice afterwards. The objective of the first round of cleaning

Channel ID	Input	Intent	Confidence	Output	Time
User1	Input 1	Intent 1	Confidence 1	Output 1	Jun 20, 2017 19:15:47
User1	Input 2	Intent 2	Confidence 2	Output 2	Jun 20, 2017 19:15:55
User1	Input 3	Intent 3	Confidence 3	Output 3	Jun 20, 2017 19:16:07

Figure 1. Sample raw data as entered in the spreadsheet

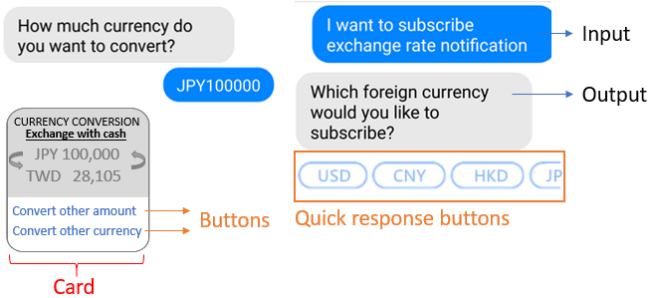

Figure 2. In addition to typing, which was always allowed, the chatbot sometimes provided alternative input modalities based on its assessments of user intentions.

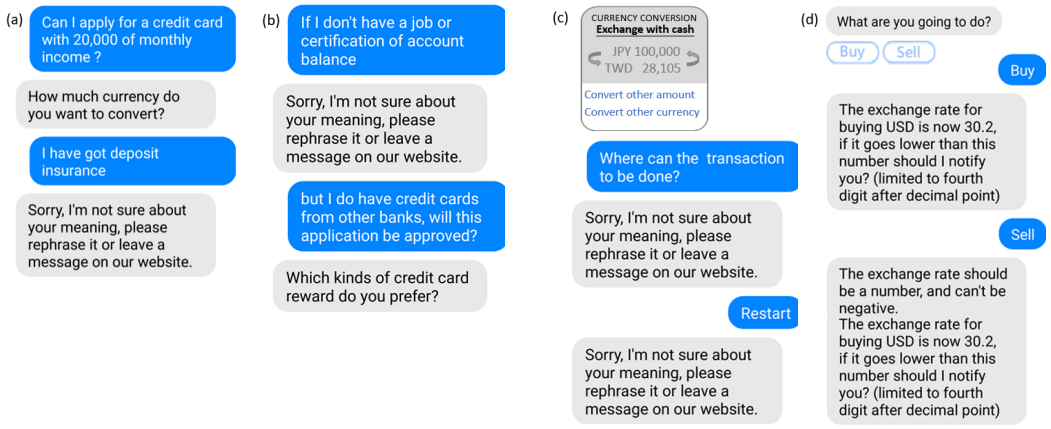

Figure 3. Four examples of unexpected content types, translated from Chinese: (a) Extra explanation; (b) Unfinished message and finishing the unfinished message; (c) Restarting the subject, i.e., the user simply texted 'restart' or repeated the first message; (d) Staying in the previous topic, i.e., the user trying to correct the information input in answer to a previous question topic from the chatbot (the user was subscribing exchange rate notification).

was to remove not replicable message exchanges, while the second and third aimed to filter exchanges for NP analysis. These latter two cleaning processes will be explained in detail in the sections 'Non-progress in Intended Usage' and 'Identifying Abandonment of Chatbot', below.

Not Replicable Exchanges

We felt it was reasonable to assume that choosing an option from a chatbot-provided list would be less likely to cause NP than typing one's own input. Accordingly, we decided to remove exchanges for which, despite visual examination of what the users had seen (see Figure 2), we could not ascertain whether the user had typed his/her own responses or had been prompted to choose a list option. Specifically, we followed the same flow as the recorded conversational exchanges to interacted with the chatbot, and added details about each such replication to the log, marking whether an input was from a card, a button, or a quick-response button. However, we could not replicate all exchanges because some of the chatbot's outputs had changed. To clarify these not replicable exchanges, we consulted the employee in the banking institution responsible for the chatbot project. We learned that these exchanges were not replicable because they were seasonal and had become no longer unavailable. Thus, we removed these not replicable exchanges. In total, we removed all 4,623 exchanges we could not replicate, and this indirectly led to all the conversations of 760 users being dropped, meaning that the dataset used for conversation analysis comprised 19,451 exchanges involving 1,837 users.

Conversation Analysis

Conversation analysis is an inductive process for analyzing how users' conversations are organized into sequences of actions and systematic practices [2]. We examined on a line-by-line basis how such sequences in our data were grouped and situated within particular instances of conversation, and assigned codes to each exchange [18,30]. One of the co-authors generated the initial set of 106 codes. Then, a second coder joined the coding process to help ensure the reliability of the codebook. The two coders then applied their first set of codes to 2% of the full dataset, and iteratively discussed and revised it until consensus was reached regarding all of codes' meanings. The two coders then tested their revised codes with a larger sample of the data, i.e., 10%. In this 10%, the two coders engaged in the same iterative process of discussion and of generating, revising, removing, and combining codes. As each 2% block of the wider dataset was coded – i.e., five times in total – the coders checked reliability again; and at the end of coding this 10% of the data, the final codebook contained 88 codes with a Cohen Kappa of 0.802, indicating high inter-coder reliability [19]. The coders then divided the rest of the data evenly between them and coded it independently. Each code fell into one of six categories: User Input (n=34), Chatbot Output (n=27), Event Subject (n=5), TimeBreak (n=5), Session (n=4), and Next Step Behaviors (n=13). All exchanges in the full dataset were coded with all these 88 codes. The first two of these categories were used for identifying NP, and the other four for identifying coping strategies. It should also be noted here that, because each conversational exchange was associated with an intent supplied by the conversation-understanding service, the coders could, in the case of the Chatbot-output category, clarify whether NP was due to mis-recognition of such intent vs. failure to recognize any intent.

Non-progress in Intended Usage

To identify NP, all exchanges were coded in two dimensions. The first was 'progress', i.e., whether the user's input enabled the chatbot to move the conversation on. For example, progress was deemed to have occurred if the user requested certain information, and the chatbot provided that information; and if the user provided information requested by the chatbot, the chatbot then needed to move on to the next request or provide information based on the user's input. The second dimension was 'usage', i.e., whether the user's conversational content was within the range of the banking service. Any usage beyond the scope of the bank's intent for its chatbot service (hereafter, "unintended usage"), such as attempts to exchange idle pleasantries with the chatbot, not

Recognition Error	Mis-recognition	Non-recognition
Expected content	43.0%	45.2%
Unexpected content/Intention gaps		
Extra explanation	1.6%	2.5%
Restart the subject	0.4%	0.4%
Stay in the previous topic	0.4%	1.3%
Unfinished message	0.8%	2.5%
Finishing an unfinished message	0.7%	1.1%

Table 1. Non-progress types, by frequency

replicable naturally led to NP because the chatbot could not deal with them. Thus, usage coding allowed us to distinguish whether NP occurred because a user entered irrelevant information vs. the chatbot being unable to recognize or mis-recognizing information that actually within its service. Unintended usage comprised 5.54% (n=1,078) of our sample of 19,451 exchanges. Our findings regarding NP do not include these, because our focus was on NP that occurred when the system was used as intended. Though the proportion of unintended usage was small, it seems that some users attempted to chitchat with this task-oriented chatbot, though it was designed to solve banking-related tasks. Among the 18,373 exchanges in which usage was intended, 59.5% consisted of users providing information and 40.5% of users requesting it.

Identifying Abandonment of Chatbot

Given that we could only access the three-month log described above, we did not know the ground truth of whether a user who had broken off communication with the chatbot used it again after July 31, 2017. Nevertheless, we felt it was important to arrive at a fair and reasonable definition of user abandonment of the chatbot service, including a time threshold for it. For example, if that threshold was 30 days, NPs occurring after July 1 could not be counted, because the logs would not be long enough to observe the user's action on the 30th day following the NP. However, NPs occurring in late June would be counted. Thus, there was a tradeoff for setting such a threshold: while a short one would allow less data removal, one that was too short would raise doubts about any claims we made regarding users' non-return to chatbot use. On the other hand, a long threshold would make our claims regarding abandonment of the service more credible, but sacrifice a large amount of data, leaving us less confident in the observed proportions of other data characteristics. Therefore, we examined how long it typically took users to return to the chatbot after they had quit a conversation, and found that four-fifths (79.5%) of them returned within 10 days. Thus, we defined non-returns of more than 10 days as *abandonment* of the chatbot, and did not analyze NPs for which 10 full days of log data was unavailable, and only present results relating to abandonment from the period from May 1 to July 21. However, it is important to note that data from July 22 through July 31 were still utilized for observation of users returning to chatbot use. It is noteworthy that, despite the relative brevity of our 10-day threshold, most (94.5%) of users who had ceased conversing with the chatbot for more than 10 days before June 30 never recommenced using it within the 30-day-plus range we could observe; suggesting that most 10-day non-returns probably did imply abandonment of the chatbot. Nevertheless, we must emphasize that we do not have the ground truth of abandonment, being unable either to confirm or deny that any return occurred after July 31, 2017.

Expected Content	
NP1	Mis-recognition
NP2	Non-recognition
Unexpected Content	
NP3	Extra explanation + Mis-recognition
NP4	Extra explanation + Non-recognition
NP5	Restarting the subject + Mis-recognition
NP6	Restarting the subject + Non-recognition
NP7	Stay in the previous topic + Mis-recognition
NP8	Stay in the previous topic + Non-recognition
NP9	Unfinished message + Mis-recognition
NP10	Unfinished message + Non-recognition
NP11	Finishing an unfinished message + Mis-recognition
NP12	Finishing an unfinished message + Non-recognition

Table 2. Non-progress categories

The final dataset from which we present the percentage of NPs consists of 17,136 exchanges from 1,685 users' conversations, of which 59.1% (N=10,131) consist of information provision, and 40.9% (N=7,005) user requests. This composition is also similar to it of all expected usages.

RESULTS

Frequency of Non-progress Occurrence

In the 17,136-exchange dataset described above, 63.5% (N=10,885) of exchanges were made by clicking on buttons or quick-response buttons. Probably because of the high percentage of these instances that rarely caused NPs, 90.6% of the exchanges (N=15,524) resulted in progress. Among the 9.4% that contained NP (N=1,607), 97.4% of the NP-causing inputs comprised users typing their own words. NP was 63 times more likely when users typed their own words than when they clicked buttons or chose quick responses from a menu (i.e., 25%, 1,565/6,251 vs. 0.4%, 42/10,885). This suggests the strong benefit of providing users with options. Interestingly, users occasionally typed the same response that appeared on a button or quick response presented to them, instead of clicking them; and in such cases, NP sometimes occurred because the user had added other text or symbols such as a question mark, making the chatbot not able to process them perhaps it expected to receive the exact same words appearing on a button or quick response.

We further found that NP occurred much more often when users requested information (88%) than when they provided it (12%). This could have been because users were more likely to use their own phrasing when asking questions than

Message reformulation		
|C1	add words	6.68%
C2	remove words	4.76%
C3	rephrase	8.82%
C4	repeat	5.75%
C5	ask new topic	5.48%
C6	others	3.56%
Quitting		
C7	quit subject temporarily	27.16%
C8	quit conversation temporarily	6.74%
C9	switch subject	13.47%
C10	abandon chatbot service	17.58%

Table 3. Users' strategies for dealing with non-progress

when answering them, due to the chatbot's requests for information containing fairly explicit instructions.

Categories of NPs

We observed two main kinds of NP: expected content and unexpected content (see Tables 1 and 2). More than 88% of the NP-containing exchanges consisted of the former: i.e., the content users entered was discernible to the researchers as what the chatbot expected, but the specific ways they said it were not correctly recognized by the chatbot. It thus either generated the message, *"Sorry, I'm not sure about your meaning, please rephrase it, or leave a message on our website"* (non-recognition), or mis-recognized their intent.

The sizeable minority of NP that was caused by unexpected content, meanwhile, could be divided into five types, as shown in Figure 3. In descending order of frequency, these were: 1) providing extra explanation of a previous input (4% of all NP-causing exchanges); 2) entering an unfinished message (3.3%); 3) attempting to finish an unfinished message (1.8%); 4) staying in a conversational topic after the chatbot had moved on to a new topic (1.7%); and 5) attempting to restart the conversation (0.9%). It was noteworthy that some of these conversational traits, while commonplace in person-to-person text messaging, are challenging for chatbots to handle, due to a general design assumption that users' intent can be completely conveyed via single, discrete messages. Although unexpected content only contributed to 12% of NPs, it had played a disproportionately large role in users' abandonment of chatbot use.

Strategies for Dealing with Non-progress

Users adopted two major strategies when they encountered NP. The first was trying to quit (65%), in descending order of frequency: temporarily changing subject (27.2%); abandon the chatbot service (for more than 10 days) (17.6%); switching the subject (for more than 10 days) (13.5%); and temporarily quitting the conversation (6.7%). In line with prior research [6,14] and our definition of abandonment of the chatbot, we defined temporarily quitting the conversation as exiting it for more than 30 minutes but less than 10 days. We found it interesting that such temporary quitting was the least frequently observed subtype of quitting strategy. The other main strategy type was message reformulation, i.e. re-trying to communicate about the same subject using different formulations (35%). It included rephrasing (8.8%), adding words (6.7%), repeating the same words (5.8%), asking a new topic on the same subject (5.5%), removing words (4.8%), and others including changing symbols, replacing words, abbreviation, correcting wrong words, switching word order, and switching language (3.6%) (see Table 3).

Relationships between Specific Types of Non-progress and Users' Strategies for Dealing with Them

NPs That Led to Chatbot Abandonment (C10)

As Figure 4 shows, unexpected content (NP3-NP12) was generally more likely to lead users to abandon chatbot use (C10) than expected content (NP1 and NP2) was. This can also be observed from Figure 5, showing that 18.7% of C10 resulted from unexpected content, whereas only about 10%

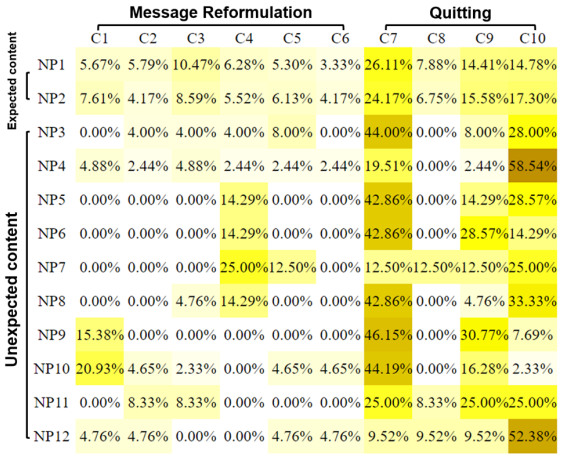

Figure 4. Heat map and percentages for each coping strategy, by NP. Odd-numbered NPs are mis-recognitions, and even-numbered ones are non-recognitions. Darker color represents higher percentage. Here, NPs that usually led to chatbot abandonment and other relatively frequently used coping strategies can be seen.

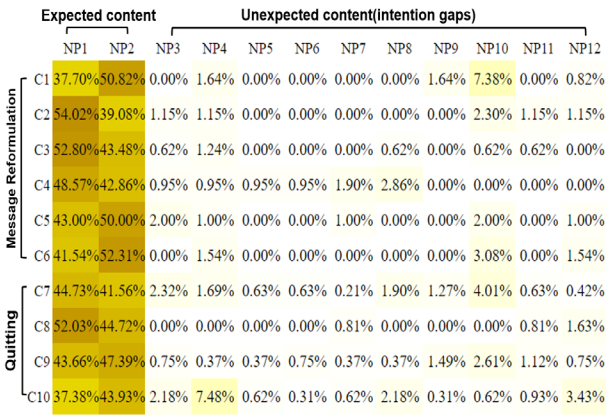

Figure 5. Heat map and percentages of each NP, by coping strategy. Darker color represents higher percentage. Note: the percentages in this figure were dominated by the proportion of the occurrence of the NPs in the dataset; thus, expected content (NP1 and NP2) had much higher percentages than unexpected content (NP3-NP12), but variation within the latter 10 NP types is the focus of the figure.

of other strategies were used in connection with such content. More specifically, NP4 and NP12 were associated with the highest likelihood of C10. Both were related to users adding additional words to their previous inputs, either to complete a previous message or to add more explanation to make it clearer. When users saw the chatbot had failed to understand such additions/corrections, they discontinued use of the chatbot for more than 10 days in more than 50% of cases. NP8 – sticking with a topic that the chatbot had moved on from – also led to a relatively high likelihood of C10, albeit less than that associated with NP4 and NP12. Conceptually, NP8 is similar to NP4 and NP12 in it involves users' messages that refer back to their own previous messages. C10 was also one of the top strategies users adopted to deal with two relatively less common NP types, NP7 and NP11: situations in which the chatbot misrecognized users' intentions to stay on the previous topic, and to finish a previously unfinished message, respectively. NP9 and NP10 were exceptional among the unexpected-content NP types, in that both were associated with very low likelihood of C10. They refer to situations in which users entered unfinished messages and the chatbot was either unable to recognize them (NP9) or mis-recognized them (NP10). The low rate of chatbot abandonment in such cases could have been because users themselves recognized how challenging it was for the chatbot to recognize unfinished messages, and thus were more likely to cope by adding words (C1).

Temporarily quitting the conversation (C8) was only ever used when the chatbot misrecognized users' intentions to stay on the previous topic, or failed to correctly process their messages aimed at completing previous, unfinished ones. Users left temporarily in these situations probably because they suspected that the chatbot would resume later.

NP Types Related to Message Reformulation Strategy

There were some NPs that users mainly used one of the message reformulation to cope with, probably because they could (or thought they could) make sense of how/why such problems had arisen [23]. For example, users were disproportionately likely to add words (C1) to unfinished messages (NP9, NP10), and to repeat themselves (C4) to cope with NP5, NP6, NP7, and NP8. What connected all four of the latter set of NP types was users wanting to correct the flow of the conversation with the chatbot: either by requesting to stay on a topic that the chatbot had moved on from (NP7 and NP8), or to restart the subject (NP5 and NP6). When they found the chatbot was unable to recognize or misrecognized such intentions, they were more likely to repeat what they just said to emphasize their true intentions. This implied that they did not recognize how the chatbot had been designed to deal with conversational flow. Interestingly, when such problems occurred, repetition and chatbot abandonment were both common reactions. Similarly, the message reformulations suggested a clear correspondence between particular NP types and particular strategies: when users in our sample adopted reformulation, it seemed to be because they had a clear idea about how to deal with the NP, instead of simply using trial-and-error, as Myers et al. [23] found in users' interaction with VUI.

Non-recognition vs. Mis-recognition

We found three interesting patterns of coping-strategy distributions across non-recognition, i.e., chatbot incapable of recognizing any intent from the input, and mis-recognition types of NP. The first was a similarity in user responses to NP1 and NP2, probably because the types of errors in these NPs were diverse, so users adopted diverse ways of coping with them. The second involved similarity in users' choices of coping strategies for NPs that were relatively unconnected to adding extra information to previous messages, including restarting (NP5, NP6) and unfinished messages (NP9, NP10), all of which were addressed mainly via a certain similar set of coping strategies, while the other were hardly used at all in such cases. This was probably because users could often recognize the causes of these NPs, and thus focused on specific strategies they were fairly sure would work: e.g., repeating (C4) to deal with NP5 and NP6, and adding words (C1) to deal with NP9 and NP10. The third pattern that involved NPs related to users' adding extra information to their own previous messages (NP3, NP4, NP7, NP8, NP11 and NP12), showed a different pattern. In these six cases, we observed especially strong divergence in the coping strategies used across the non-recognition and mis-recognition NPs. That is, NP7 and NP11 were associated with a broad range of coping strategies, while their direct counterparts, NP8 and NP12, were dealt with using just one or two clear primary strategies. NP3 and NP4, on the other hand, provoked starkly different primary coping strategies: C7 (temporary quitting subject) for NP3, and C10 (abandon the chatbot) for NP4. These results suggest that the adoption of coping strategy is determined by both users' intention, such as restart a subject, and the types of error they saw. i.e., both the key elements of an NP, rather than either by itself – that determined how users coped with the NP.

Signs that Users Would Abandon the Chatbot Service

Although prior research reported that certain types of NPs lead more easily to cessation of chatbot use [4, 12], we are particularly interested in other signs that such cessation is about to occur, potentially including numbers of consecutive NPs and changes in users' strategies for dealing with NPs. This reflects our assumption that chatbots could be designed to detect these events and act to forestall user abandonment.

Three Consecutive Occurrences of Non-progress

The most consecutive NPs that occurred was nine, but this happened in only two conversations. As shown in Figure 6, among all cessations of chatbot use in our data, 90% were preceded by no more than three consecutive errors, and 75% by no more than two consecutive errors. Surprisingly, 45% of abandonment were preceded by just one NP. Conversely, just under 30% of all individuals who abandoned the service long-term did so at the second NP, while just over 15% did so upon experiencing a third. Those who never abandoned the chatbot, on the other hand, were very likely to have experienced *more* consecutive NPs, and we observed a sharp

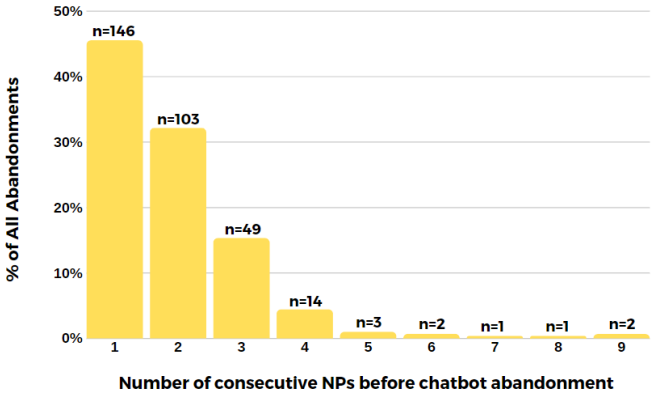

Figure 6. Proportions of users who abandoned chatbot use, by number of consecutive NP errors. Nearly half (45%) of long-term abandonments were preceded by just one NP, and 90% were preceded by no more than three consecutive ones.

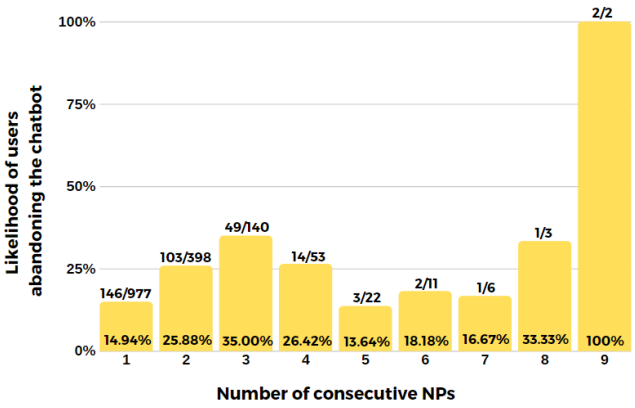

Figure 7. Likelihood of users abandoning the chatbot, by number of consecutive NPs. Just 14.9% of users abandoned the chatbot upon encountering their first NP, but 25.9% abandoned it on encountering the second consecutive NP.

drop in the proportion of users who abandoned the service after four consecutive NPs, as compared to those who had experienced three or fewer (see Figure 6 and Figure 7). This decrease in the likelihood that an individual would abandon the chatbot at or above the four-NP mark could have been because those who could endure four consecutive NPs or more were also more willing to use trial and error to deal with them. Taken as a whole, the above findings imply that error messages at the second and third NPs should be carefully designed to discourage users from giving up.

Repeated Strategies Ending with Reformulation

We also examined the final strategies users adopted within their final session before abandonment of chatbot use. We only examined coping strategies in the last series of consecutive NPs before they abandoned, as shown in Figure 8. For purposes of the figure, the construct of "message reformulation" covers all reformulation strategies, and that of "switching subject" includes both temporary and long-term switching. (Quitting the conversation temporarily was not included, because it would have implied the existence of a subsequent session). This examination revealed two patterns that could potentially be used as warning signs of imminent chatbot abandonment. First, message reformulation was deployed more often than switching subject as users' final coping strategy before they left. This is an especially interesting pattern since the general case show an opposite trend (see Figure 4), i.e. overall switching subject was the most often used strategy to deal with NPs. Second, repeated use of the same strategy set (i.e., either message reformulation or subject switching) was more often used than a mixture of the two immediately before users ceased chatbot use. Among all two consecutive coping strategies, repeated use of message reformulation and switching subject together took place 65.22% of the time; Likewise, among all three consecutive coping strategies, repeated use of message reformulation and switching subject together took place 69.2% of the time. Given that, as discussed earlier, most users left before the third consecutive NP, a user's use of two consecutive message reformulations is probably a strong indicator that he/she will abandon the chatbot as soon as the next NP occurs.

DISCUSSION

NPs and Tactics in Chatbot Interactions

Jain et al.[12] indicated that some users of chatbots abandon them when they feel these agents' functionality or behavior does not meet their expectations. Our study results not only support this point, but further illustrate the high frequency of quitting as coping strategy, even as compared to message reformulation. On the other hand, our results suggest a distinction between task-oriented chatbots and VUIs. Specifically, Myers et al [23] showed that mishearing users and/or mapping their utterances to the wrong intentions were the most frequent errors users encountered when using a VUI calendar manager. In contrast to those findings, we observed that the number of mis-recognition and non-recognition NPs was well balanced. Secondly, while both Stent et al.[29] and Myers et al. [23] found users most often used hyper-articulation, and the latter added that simplification was also often used to cope with recognition errors, our dataset revealed very infrequent use of repeating and removing words, the two coping strategies we studied that were most

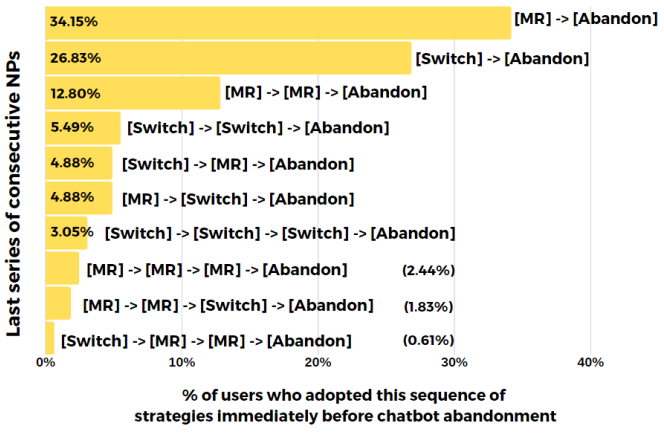

Figure 8. Relative use of message reformulation ("MR") vs. switching subjects ("Switch") as the user's final strategy before chatbot abandonment.

similar to hyper-articulation and simplification, respectively. Moreover, when they were used, it was only to cope with certain NPs. Indeed, the entire category of message reformulation was used fairly infrequently, because the users in our sample more often decided to switch subjects or abandon the chatbot when they encountered NP. A much higher frequency of abandoning the chatbot (as compared to quitting the VUIs in the previous studies aforementioned) could have been because, unlike VUI users, the banking chatbot users perceived that they had alternative means of achieving the same goals: websites, apps, customer support, local bank branches, etc. On the other hand, some of the discrepancies in these study results could be attributable to innate differences between VUI and text input. That is, most users in most situations are capable of ensuring that the text they type is correct, but users of VUI might not know if its errors arise from incorrect speech recognition or from incorrect conversational understanding [20,28]. For the same reason, chatbot users may be less likely to repeat themselves than their VUI-using counterparts, because the former group can almost immediately confirm, via direct visual inspection, that their own input is not garbled (at least in their own estimation). By the same token, when the sampled chatbot users repeated themselves, it was generally part of an attempt to correct the overall conversation flow, rather than to confirm the content of their own inputs. Additionally, our results revealed five categories of user-input content that – unexpectedly, from our point of view – the focal banking chatbot was unable to handle; though arguably, these unexpected inputs would be challenging for mainstream chatbots to handle as well. That is, unlike VUI systems that are not designed to react to only one utterance, but to multiple ones at a time (making them more likely to process the collective meaning of all user utterances [27]), mainstream chatbots built on widely used chatbot platforms such as Facebook Messenger respond when and as the user sends any discrete message. Moreover, the intent of those unexpected inputs that did not present a complete sentence were challenging for a current state-of-the-art language understanding service to accurately recognize. While unfinished/broken-up sentences are common or indeed inevitable in human-to-human texting, they resulted in extensive confusion in this case, unlike in VUI.

Relationships between NPs and Strategies

Both NPs and Types of Recognition Error Matter

Our relatively large conversations-log dataset enabled us to observe several trends in the relationships between particular NPs and the strategies chatbot users cope with them. First, users' messages that had caused the NPs played a vital role in determining such strategies: e.g., adding words to finish a previous, unfinished message, or repeating the same message by way of emphasizing a preferred conversational topic. However, we noted that with certain NPs, the chatbot's fundamental error type – i.e., non-recognition vs. mis-recognition – also influenced users' coping strategies. In particular, we found that the *combination* of the user's intention behind an NP with the type of error the user saw determined what users would do to cope with that NP. For example, the two fundamental types of errors did not lead to considerable differences in coping strategies when the content of the users' messages was expected, but when the NPs were caused by unexpected content, highly divergent coping strategies could result. For instance, NP7 and NP8 were, respectively, the chatbot's mis-recognition and non-recognition of a user's attempt to stick with the existing conversational topic. When the chatbot mis-recognized this, users were very likely to repeat the same message and unlikely to switch subjects, whereas when the chatbot expressed itself unable to recognize it, the users very often ceased communicating with it. Similarly, when the chatbot expressed itself unable to recognize users' supplementary explanations (NP4) or their attempts to finish unfinished messages (NP12), more than 50% of them ceased using the chatbot; but the likelihood of their doing so was roughly halved when the chatbot mis-recognized these intentions as other intentions (NP3 and NP11). And conversely, when users attempted to restart a conversation whose intention had been mis-recognized by the chatbot, the likelihood of their quitting the chatbot was twice as high as it was when the chatbot failed to understand their input completely.

The above-mentioned differences in coping-strategy distribution across the two fundamental types of chatbot-communication error raise two possibilities. The first is that users have divergent interpretations of the meanings of non-recognition and mis-recognition across different NP types, and use different sets of strategies based on such interpretations. And second, depending on their own intentions, users might not have any clear idea about how to cope with certain NPs, even when the error type is clearly identified. Thus, for some NPs, their coping strategies were quite evenly distributed, possibly reflecting a trial-and-error approach, while for others they tended to focus narrowly on just one or two strategies. So, while Jain et al. [12] indicated that some users preferred chatbots to conceal their incapability, and others that they admit to it, our results suggest that both major types of a banking chatbot's recognition errors, either admitting its incapability of recognizing or mis-recognizing users' intent could lead users to abandon it (albeit at different moments). This implies that chatbot designers should not simply be asking themselves which type of error to reveal, but rather, which types to show in the contexts of various user intentions.

Design Implications

Preventing Non-progress: Processing Multiple Messages

Users were most likely to cease using the studied banking chatbot when they encountered an NP type caused by unexpected content (NP3-NP12). Three-fifths of all such occurrences were caused by a user's attempt to refer back to one or more of their previous messages via his/her current one. This led to NPs so readily because, as noted above, current mainstream chatbot platforms have been designed to

respond at once whenever they receive any message, regardless of whether the user has really finished what he/she wants to say. Given how hard it is (even for humans) to correctly detect the meaning of a 'broken' message by reading either half of it in isolation, we recommend that chatbot developers consider having their agents process multiple messages at a time, rather than responding so quickly.

This process could be assisted through detection of whether users' typing action has ceased or is ongoing, and the inclusion of a brief, naturalistic delay between receiving users' input and replying to it. Another approach would be to have the chatbot consider all of a given user's previous messages whenever it receives a new one. Thus, the chatbot would be able to take its own failure to understand a message as a sign that a message might be incomplete and that it should therefore be merged with the preceding or following one. The chatbot could even be designed to withdraw its previous response as soon as it recognized that, due to message incompleteness, its initial assessment of the user's intent was wrong. Giving due consideration to previous messages might also help prevent the chatbot from mistakenly moving on to a new stage of conversation when the user is in fact still referring to a previous message; and this function of dialogue management might emerge as particularly crucial when users' state in the dialog are being tracked by a state tracking [10]. Another option would be allowing users to reply to or comment on specific previous messages (as is already possible in Facebook Messenger), to make specific corrective information available to the chatbot while the conversation is still in progress.

Preventing Non-Progress: Granting Users Flow Control

Currently, users of the target chatbot are only allowed to follow its flow, despite sometimes wanting to reset or correct its conversational flow. Between them, the chatbot's 1) non-recognition of users' intentions to stick with their existing conversational topics and 2) mis-recognition of their intention to restart their subject led to nearly one-third of abandoning chatbot service. Therefore, we recommend that chatbot developers allow users more control of conversational flow. At a minimum, users should be granted an 'official' way to quit or restart the conversation. We also suggest allowing users to rewrite or delete their own messages. This is because, in our data, they sometimes persisted in attempts to make a message understood not only because the previous one had not been appropriately processed, but also because they wanted to make changes to what they had written earlier. If users were allowed to 'undo' their previous message(s), it would give the chatbot scope to re-calculate their intent. But currently, mainstream chatbot platforms do not engage in such re-consideration of revised previous messages. Also, because failed attempts to control conversational flow were often manifested in the data by users repeating the same message, a chatbot capable of detecting repetition could proactively provide its users with options for jumping to particular topics and/or phrases.

Avoidance of Chatbot-use Discontinuation Due to Negative Experiences

We identified several warning signs that users were about to abandon the chatbot. These included, firstly, the occurrence of certain NPs (i.e., NP4, NP8, NP12). Thus, it is important to prevent those NPs from occurring, and this formed the basis of our above recommendation that multiple messages be processed together. Second, the presentation of errors as either mis- or non-recognition prompted widely divergent coping strategies. Thus, developers should distinguish between the kinds of NPs that most often lead to abandoning the chatbot service and those that users can most easily resolve themselves, according to the data presented in Figure 6. Adjusting the confidence threshold for recognizing a user intention might be one way to achieve this. For example, the chatbot revealing its incapability of recognizing extra explanations would tend to abandon chatbot service more likely than if it appeared to recognize them, but wrongly. Assuming that users add extra explanations of some intentions more than of others, developers could lower the confidence threshold for recognizing the former group of intentions, such that the chatbot is more likely to identify them. Likewise, when a given chatbot's mis-recognition of a particular intention has been identified by researchers as highly likely to provoke its users into abandoning a chatbot service, developers can raise its threshold such that it expresses its incapability more frequently. Such actions, especially if applied jointly, appear likely to reduce the incidence of abandonment.

Third, developers could usefully focus on preemptive detection of signs that users are about to abandon the chatbot. One such sign is that repeated and consecutive use of message reformulation or switching subjects to deal with consecutive NPs. A more important sign, however, is the occurrence of consecutive NPs, given that those users in our data who ceased chatbot use at all rarely stayed after seeing the third NP in a row. Thus, we recommend that if a second NP occurs, either someone from customer service takes over the conversation, or that the chatbot provides additional guidance or options. All that being said, however, we do not think users leaving the conversation is always bad, provided that a better alternative avenue of communication exists and that their main reason for leaving is not a feeling of helplessness aroused by the interaction. It is better if they leave because they have been overtly guided to a better alternative instead than because their experience was a negative one, since at worst, the latter phenomenon could have a permanent negative impact on their views not only of the chatbot but of the company or entity that it represents.

Guiding to Other Options and Alternative Channels

Users may be able to express their intentions to chatbots more effectively if they are given alternative ways of doing so from a menu based on their initial input. In creating any such menu, developers should ensure that it considers previous messages and not just the current one, as numerous errors arise from users referring back to their previous

messages as a (generally futile) means of explaining what they really wanted to do. Such menus should provide both *options* and *alternative channels*. Options, such as buttons or quick responses, have been found in the current study to be relatively effective ways of acquiring users' intentions correctly and reducing the incidence of NP. And, in addition to its main purpose (as suggested by Ashktorab et al. [4]), providing some "guesses" of users' intentions and provide possible options could give users a sense that the chatbot was *trying* to recognize their intentions. Options including ways to return or reset should be also provided even when the chatbot thinks it has successfully recognized the user's intention, given the ever-present possibility that it is wrong.

Guiding users to alternative channels of communication is equally important, and perhaps especially useful for users who tend to add hard-to-recognize extra words to their messages or to use colloquial language. Since messages with these features are difficult for conversation-understanding services to process, chatbot designers can guide users to websites, apps, or customer support, including appropriate links and other contact details, rather than simply letting them experience frustration with the chatbot, which might cause them to abandon use of it in the future. In addition, the same users who try to converse with a chatbot as if they are talking to a real person may also be those likely to have unrealistic expectations of its real ability to understand their messages. Directing such users to more appropriate communication channels, while also informing them about what kinds of tasks the chatbot can actually do, may gradually help them develop more realistic expectations.

Limitations and Future Work

Given the characteristics of the conversation logs, our conversation analysis is subject to a number of limitations. First, the analysis of input text did not distinguish between text that users typed themselves and preset response buttons that they could choose to press. Therefore, we could not be certain about what visual elements users saw at the moment of interacting with the chatbot, though we tried our best to replicate each situation by typing the same inputs. Additionally, it is important to note that whether a conversation made progress was influenced by the performance of the conversation-understanding service, and thus, so were our NP categorization results. Second, our analysis has focused on a particular task-oriented chatbot in the banking domain. As a preliminary examination of whether the unexpected inputs we observed were also difficult for other chatbots, we tested them with those chatbots included in a prior study [12] that were still available for use, as well as with several public chatbots operated by large corporations such as Facebook, HubSpot, Hangseng, and HSBC. Based on these tests, we established that unexpected inputs were difficult for most current mainstream chatbot platforms and conversation-understanding services to handle. Given that the banking chatbot we studied adopted mainstream services for both (i.e., Facebook Messenger + IBM Watson), we believe that some of the NPs and user strategies we observed are generalizable to chatbots that use these same mainstream services. Nevertheless, to fully validate the generalizability of our main study's results, similarly detailed conversation logs from additional chatbot services would be needed. In addition, it is questionable that our results would be generalizable across different cultures and languages. Third, because the log only contained three months' worth of data, we could not always be certain whether discontinuations of chatbot use were temporary or permanent. Finally, the dataset did not include any user reflections on their conversations, only the conversations themselves. Thus, we used a neutral term NP instead of "conversation breakdown", which we felt described users' subjective feelings to which we did not actually have access. As such, it is possible that users received output that we saw as normal/progress, but that they saw as abnormal/NP – or vice versa. For this reason, we recommend that the behaviors and patterns identified in this study be followed up with empirical studies involving users' reflections to clarify such issues and explain the observed behavior.

CONCLUSION

Using logs of actual banking customers' interactions with an existing chatbot, we observed numerous occurrences of specific obstacles users encountered and how they dealt with them when interacting with the chatbot. This, the first conversation analysis of the natural use of a task-oriented banking chatbot as a case study of conversational NP, identified 12 types of NP; five types of unexpected content that were especially challenging for the chatbot to recognize; and 10 coping strategies, along their interrelationships with particular NP types. This enabled us to identify specific NPs that were mostly likely to lead to users discontinuing their use of the chatbot, with the most frequent one being the chatbot's inability to understand users' additional words of messages they had previously sent. We were also able to identify signs that users were about to terminate their conversations, including three consecutive NPs, and repeated and consecutive use of message formulation or subject-switching. Finally, we have provided practical design recommendations for task-oriented chatbots that should help to prevent NP, guide users through such events, and help companies detect when chatbot users are on the verge of giving up. Overall, like previous research on the need to gracefully mitigate breakdowns in human-robot interaction [8, 15], the present study has revealed both potential ways to deal with, and obstacles to dealing with, inadequately supported chatbot conversations.

ACKNOWLEDGEMENT

We sincerely thank our shepherd Prof. Andreas Riener for his excellent guidance that helps us significantly improve the paper. We also thank the banking institution for providing the chatbot conversation log. This research was supported in part by the Ministry of Science and Technology, R.O.C (MOST 108-2218-E-009 -050).

REFERENCES

[1] Adobe. 2019 Digital Trends: Financial Services in Focus. Retrieved September 20, 2019 from https://www.adobe.com/content/dam/acom/uk/modal-offers/2019/DT-Report-2019/Econsultancy-2019-Digital-Trends-Financial-Services.pdf

[2] Paul M. Aoki, Margaret H. Szymanski, Luke Plurkowski, James D. Thornton, Allison Woodruff, and Weilie Yi. 2006. Where's the "Party" in "Multi-party"?: Analyzing the Structure of Small-group Sociable Talk. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work* (CSCW '06), 393–402. DOI:https://doi.org/10.1145/1180875.1180934

[3] appliedAI. 8 Epic Chatbot / Conversational Bot Failures [2019 update]. *appliedAI*. Retrieved September 20, 2019 from https://blog.aimultiple.com/chatbot-fail/

[4] Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19), 254:1–254:12. DOI:https://doi.org/10.1145/3290605.3300484

[5] Marion Boiteux. 2018. Messenger at F8 2018. *Medium*. Retrieved September 20, 2019 from https://blog.messengerdevelopers.com/messenger-at-f8-2018-44010dc9d2ea

[6] Lara D. Catledge and James E. Pitkow. 1995. Characterizing browsing strategies in the World-Wide web. *Computer Networks and ISDN Systems* 27, 6 (April 1995), 1065–1073. DOI:https://doi.org/10.1016/0169-7552(95)00043-7

[7] Ana Paula Chaves and Marco Aurelio Gerosa. 2019. How should my chatbot interact? A survey on human-chatbot interaction design. *arXiv:1904.02743 [cs]* (April 2019). Retrieved September 20, 2019 from http://arxiv.org/abs/1904.02743

[8] Sara Engelhardt, Emmeli Hansson, and Iolanda Leite. 2017. Better faulty than sorry : Investigating social recovery strategies to minimize the impact of failure in human-robot interaction. 19–27. Retrieved September 20, 2019 from http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-217595

[9] Jonathan Grudin and Richard Jacques. 2019. Chatbots, Humbots, and the Quest for Artificial General Intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19), 209:1–209:11. DOI:https://doi.org/10.1145/3290605.3300439

[10] Matthew Henderson. 2015. Machine Learning for Dialog State Tracking: A Review. In *Proceedings of The First International Workshop on Machine Learning in Spoken Language Processing*.

[11] Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q. Vera Liao, Khai Truong, and Shwetak Patel. 2018. FarmChat: A Conversational Agent to Answer Farmer Queries. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4 (December 2018), 170:1–170:22. DOI:https://doi.org/10.1145/3287048

[12] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N. Patel. 2018. Evaluating and Informing the Design of Chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference* (DIS '18), 895–906. DOI:https://doi.org/10.1145/3196709.3196735

[13] Jiepu Jiang, Wei Jeng, and Daqing He. 2013. How Do Users Respond to Voice Input Errors?: Lexical and Phonetic Query Reformulation in Voice Search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '13), 143–152. DOI:https://doi.org/10.1145/2484028.2484092

[14] Rosie Jones and Kristina Klinkner. 2008. Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs. 699–708. DOI:https://doi.org/10.1145/1458082.1458176

[15] M. K. Lee, S. Kiesler, J. Forlizzi, S. Srinivasa, and P. Rybski. 2010. Gracefully mitigating breakdowns in robotic services. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 203–210. DOI:https://doi.org/10.1109/HRI.2010.5453195

[16] Q. Vera Liao, Matthew Davis, Werner Geyer, Michael Muller, and N. Sadat Shami. 2016. What Can You Do?: Studying Social-Agent Orientation and Agent Proactive Interactions with an Agent for Employees. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (DIS '16), 264–275. DOI:https://doi.org/10.1145/2901790.2901842

[17] Q. Vera Liao, Muhammed Mas-ud Hussain, Praveen Chandar, Matthew Davis, Yasaman Khazaeni, Marco Patricio Crasso, Dakuo Wang, Michael Muller, N. Sadat Shami, and Werner Geyer. 2018. All Work and No Play? In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), 3:1–3:13. DOI:https://doi.org/10.1145/3173574.3173577

[18] Jane Lockwood. 2017. An analysis of web-chat in an outsourced customer service account in the Philippines. *English for Specific Purposes* 47, (July 2017), 26–39. DOI:https://doi.org/10.1016/j.esp.2017.04.001

[19] Matthew Lombard, Jennifer Snyder-Duch, and Cheryl Bracken. 2005. Practical Resources for Assessing and Reporting Intercoder Reliability in Content Analysis Research Projects. *Retrieved April* 19, (January 2005).

[20] 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* (CHI '16), 5286–5297. DOI:https://doi.org/10.1145/2858036.2858288

[21] 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 *Human-Computer Interaction - INTERACT 2017* (Lecture Notes in Computer Science), 441–459.

[22] Joanne Meredith. 2017. Analysing technological affordances of online interactions using conversation analysis. *Journal of Pragmatics* 115, (July 2017), 42–55. DOI:https://doi.org/10.1016/j.pragma.2017.03.001

[23] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for How Users Overcome Obstacles in Voice User Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), 6:1–6:7. DOI:https://doi.org/10.1145/3173574.3173580

[24] Pernilla Qvarfordt, Arne Jönsson, and Nils Dahlbäck. 2003. The Role of Spoken Feedback in Experiencing Multimodal Interfaces As Human-like. In *Proceedings of the 5th International Conference on Multimodal Interfaces* (ICMI '03), 250–257. DOI:https://doi.org/10.1145/958432.958478

[25] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (CHIIR '17), 117–126. DOI:https://doi.org/10.1145/3020165.3020183

[26] Xin Rong, Adam Fourney, Robin N. Brewer, Meredith Ringel Morris, and Paul N. Bennett. 2017. Managing Uncertainty in Time Expressions for Virtual Assistants. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17), 568–579. DOI:https://doi.org/10.1145/3025453.3025674

[27] Dirk Schnelle and Fernando Lyardet. 2006. Voice User Interface Design Patterns. In *EuroPLoP*.

[28] Ben Shneiderman, Catherine Plaisant, Maxine Cohen, Steven Jacobs, Niklas Elmqvist, and Nicholas Diakopoulos. 2016. *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (6th ed.). Pearson.

[29] Amanda J. Stent, Marie K. Huffman, and Susan E. Brennan. 2008. Adapting speaking after evidence of misrecognition: Local and global hyperarticulation. *Speech Communication* 50, 3 (March 2008), 163–178. DOI:https://doi.org/10.1016/j.specom.2007.07.005

[30] Wyke Stommel, Trena M. Paulus, and David P. Atkins. 2017. "Here's the link": Hyperlinking in service-focused chat interaction. *Journal of Pragmatics* 115, (July 2017), 56–67. DOI:https://doi.org/10.1016/j.pragma.2017.02.009

[31] Justin D. Weisz, Mohit Jain, Narendra Nath Joshi, James Johnson, and Ingrid Lange. 2019. BigBlueBot: Teaching Strategies for Successful Human-agent Interactions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (IUI '19), 448–459. DOI:https://doi.org/10.1145/3301275.3302290

[32] Jennifer Zamora. 2017. I'm Sorry, Dave, I'm Afraid I Can'T Do That: Chatbot Perception and Expectations. In *Proceedings of the 5th International Conference on Human Agent Interaction* (HAI '17), 253–260. DOI:https://doi.org/10.1145/3125739.3125766