



Savcheva, D. and Foster, M. E. (2018) Comparing User Responses to Limited and Flexible Interaction in a Conversational Interface. In: Proceedings of the 6th International Conference on Human-Agent Interaction (HAI '18), Southampton, UK, 15-18 Dec 2018, pp. 368-370. ISBN 9781450359535.

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

© The Authors 2018. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in the Proceedings of the 6th International Conference on Human-Agent Interaction (HAI '18), Southampton, UK, 15-18 Dec 2018, pp. 368-370. ISBN 9781450359535. <https://doi.org/10.1145/3284432.3287189>.

<http://eprints.gla.ac.uk/176188/>

Deposited on: 3 January 2019

Comparing User Responses to Limited and Flexible Interaction in a Conversational Interface

Dilyana Savcheva

School of Computing Science, University of Glasgow
2091499S@student.gla.ac.uk

Mary Ellen Foster

School of Computing Science, University of Glasgow
MaryEllen.Foster@glasgow.ac.uk

CCS CONCEPTS

• **Human-centered computing** → **User studies**; **Natural language interfaces**; • **Computing methodologies** → **Discourse, dialogue and pragmatics**;

ACM Reference Format:

Dilyana Savcheva and Mary Ellen Foster. 2018. Comparing User Responses to Limited and Flexible Interaction in a Conversational Interface. In *6th International Conference on Human-Agent Interaction (HAI '18), December 15–18, 2018, Southampton, United Kingdom*. ACM, New York, NY, USA, 3 pages. <https://doi.org/10.1145/3284432.3287189>

1 INTRODUCTION

The principles governing written communication have been well studied, and well incorporated in interactive computer systems. However, the role of spoken language and in human-computer interaction, while an increasingly popular modality, still needs to be explored further [3]. Evidence suggests that this technology must further evolve in order to support more “natural” conversations [2], and that the use of speech interfaces is correlated with a high cognitive demand and attention [4]. In the context of spoken dialogue systems, a continuum has long been identified between “system-initiative” interactions, where the system is in complete control of the overall interaction and the user answers a series of prescribed questions, and “user-initiative” interactions, where the user is free to say anything and the system must respond [5]. However, much of the work in this area predates the recent explosive growth of conversational interfaces.

In this study, we develop two versions of a conversational skill using a modern conversational interface, the Amazon Alexa Toolkit: a basic version, where the interaction is fully system-driven (“Cooking Buddy”), and an enhanced version that supports more sophisticated and flexible user interaction (“Cooking Buddy Pro”). We compare user responses to the two systems using a range of objective and subjective measures. In a user study, the subjective responses to both skills were largely positive, with no difference among the participants; however, the error rate was significantly lower for the participants who used Cooking Buddy Pro. Also, the free-form comments about Cooking Buddy Pro were more positive than those for the basic Cooking Buddy skill.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

HAI '18, December 15–18, 2018, Southampton, United Kingdom

© 2018 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-5953-5/18/12.

<https://doi.org/10.1145/3284432.3287189>

- User: Alexa, ask Cooking Buddy to tell me a recipe with cherries and duck.
- CB: The best recipe I found is: . . . , published by: . . . , with a social rank of: . . . Do you like this recipe?
- User: No.
- CB: Okay. Let’s try this one: . . . Do you like this recipe?
- User: Yes.
- CB: Great! Here are the ingredients for this recipe: . . .

Figure 1: Sample interaction with Cooking Buddy

- User: Alexa, ask Cooking Buddy Pro to tell me a recipe with bananas.
- CBP: Okay. Let’s try this one: . . . Do you like this recipe?
- User: Yes
- CBP: Great. If you ask me, I can tell you the ingredients, the publisher or the link to the recipe. What do you want me to do?
- User: Tell me the ingredients, please.
- CBP: Here are the ingredients: . . .
- User: Thank you.
- CBP: You are welcome. Is there anything else you would like me to do for you?
- User: Tell me the link for the recipe.
- CBP: Okay. You can now find the link to the recipe on a card in your mobile app. Goodbye!

Figure 2: Sample interaction with Cooking Buddy Pro

2 SKILLS DEVELOPED

To study the influence of interaction style, two versions of a conversational skill were developed using the Amazon Alexa API were developed. The selected application domain was cooking recipes, so the skills were developed based on the API provided by the website **Food 2 Fork** (<http://food2fork.com/>), which allows recipes to be browsed by ingredient. Two versions of the skill were developed: a basic version (“Cooking Buddy”), which permitted only a simple, task-focussed interaction, and an enhanced version (“Cooking Buddy Pro”), which added mixed-initiative and social chat features to the basic interaction. Sample conversations with both systems are given in Figures 1 and 2.

3 USER EVALUATION

To compare the two versions of the system, a user evaluation was carried out involving 28 participants (14 female), with ages ranging from 18 to 49; all participants were fluent English speakers, and 7 of them were native. For the purposes of the study, two different task types were defined: one where the users were given full step-by-step instructions (Figure 3), and one where only the overall target was specified (Figure 4). The study involved a 2x2 between-participants factorial design, with the two factors being the skill under evaluation (Cooking Buddy vs Cooking Buddy Pro) and the task specification (step-by-step vs target only).

- (1) Say (Cooking Buddy // Cooking Buddy Pro) to begin the conversation.
- (2) Request a recipe with **chicken** and **lemon**.
- (3) (Answer "No" // Request the next recipe).
- (4) Write down the name of the recipe.
- (5) Obtain the ingredients for the recipe.
- (6) Say "Exit" to leave.

Figure 3: Sample "step-by-step" task instructions

- (1) Use (Cooking Buddy // Cooking Buddy Pro) to find the *second* recipe with **cherries** and **chocolate**.
- (2) Write down the name of the recipe and its ingredients.

Figure 4: Sample "target only" task instructions**Table 1: Errors during interactions**

	Step by Step	Target Only	Mean
CB	4.33 (2.42)	2.38 (1.69)	3.21 (2.19)
CBP	1.57 (1.90)	1.00 (1.41)	1.29 (1.69)
Mean	2.86 (2.41)	1.64 (1.69)	2.25 (2.13)

Each participant first carried out a recipe-finding task using the Food 2 Fork website on a laptop, and then used either Cooking Buddy or Cooking Buddy Pro to carry out one task; the interaction with the skills took place on an Amazon Echo Dot smart speaker. The skills and task specification were alternated so that 7 participants carried out each task type with each of the two skills.

After completing the tasks, each participant filled out a questionnaire consisting of three main sections: an initial section to gather demographic information, a second section based on the Subjective Assessment of Speech System Interfaces (SASSI) questionnaire [1], and a final section with three open-ended questions to gather any further information about the experience. The SASSI section covered five of the six factors included in the full SASSI questionnaire (System Response Accuracy, Likeability, Cognitive Demand, Annoyance, Habitability) and included a total of 17 of the full 34 questions, each rated on a 7-point Likert scale ranging from "Strongly disagree" to "Strongly agree".

4 RESULTS

4.1 Performance analysis

User performance was assessed by examining the logs of the interactions as recorded by Amazon CloudWatch. For each participant, an **error** was counted every time the system entered a situation where an invocation failed as a result of handled/unhandled exceptions, timeouts, permission errors, or any other situation that caused Alexa to terminate. Table 1 shows the mean number of error situations encountered per interaction across all of the conditions.

A 2x2 between samples ANOVA was used to analyze the results. A significant main effect of the Skill was found ($F(1,1) = 7.44, p < 0.05$), suggesting that Cooking Buddy Pro ($M=1.29, SD=1.69$) yielded fewer errors on both tasks compared to Cooking Buddy ($M=3.21; SD=2.19$). No significant main effect was found for the Task ($F(1,1) = 2.94, p=0.99$) as well as a non-significant interaction ($F(1,1)=.827, p=0.37$).

4.2 SASSI questions

The data from the SASSI questionnaire was collected and analyzed as follows. First, the ratings for each negative item were reversed, so that a high score is always positive. Secondly, the individual scores for each of the five attributes were calculated as a sum of the ratings of the items composing it. An ANOVA test yielded no significant main effects of any of the 5 attributes: System Response Accuracy, Likeability, Cognitive Demand, Annoyance, Habitability or the total score of each group. Furthermore, no significant interactions between Skill and Task were found.

4.3 Free-form comments

Users identified a number of issues on the free-form portion of the questionnaire. Some participants identified basic functional issues with the spoken interaction – e.g., difficulty in understanding user accents, and issues with the speed of information presentation. In comparison with the Food 2 Fork website, some participants felt that it was more fun and particularly useful for tasks such as cooking where the hands were occupied; however, others preferred the website because it was more responsive and was able to provide visual representations. Overall, the comments for CBP were generally more positive than those for CB, and participants mentioned features such as its friendliness.

5 DISCUSSION

The goal of this experiment was to compare a basic, system-initiative conversational skill with one that included more aspects of mixed-initiative interaction and social chat. Although no difference was found between the user ratings of the two systems on the SASSI questionnaire, the study did show that interactions with the mixed-initiative CBP system included significantly fewer error situations, suggesting that from a performane point of view, a more "human-like" system provides more efficient interaction. Also, the responses to the free-form questionnaire suggested that users had more positive reactions overall to the CBP system compared to the CB system.

One possible explanation for the SASSI results is that interacting with this sort of conversational system was a novel experience for nearly all of the participants – in fact, approximately 2/3 of the participants had never used a spoken dialogue system before. So the ratings may have been influenced by the novelty effect – it is possible that with a longer-term interaction or a within-subjects design, the differences between the systems would have become more apparent on the questionnaire.

In conclusion, the results of this study suggest that the CBP performed slightly better than CB—however, this result was not reflected in the subjective user evaluation. To address this issue further, future studies should probably be designed to minimise the novelty effect of conversational interaction, perhaps through repeated interactions with each system; in that case, the impact of the different interaction styles would become more evident in the user ratings. Also, a more task-based follow-up study (in which participants are actually using the skill to perform real meal-selection tasks) might also produce stronger results. Note that it is not necessarily the case that such a follow-up study would favour CBP: depending on the specific user task, a system-initiative interaction might actually be more functional [5].

REFERENCES

- [1] Kate S Hone and Robert Graham. 2000. Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering* 6, 3-4 (2000), 287–303.
- [2] Gustavo López, Luis Quesada, and Luis A Guerrero. 2017. Alexa vs. Siri vs. Cortana vs. Google Assistant: a comparison of speech-based natural user interfaces. In *International Conference on Applied Human Factors and Ergonomics*. Springer, 241–250.
- [3] Anna-Brita Stenstrom. 2014. *Introduction to Spoken Interaction*, An. Routledge.
- [4] David L Strayer, Joel M Cooper, Jonna Turrill, James R Coleman, and Rachel J Hopman. 2017. The smartphone and the driver's cognitive workload: A comparison of Apple, Google, and Microsoft's intelligent personal assistants. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 71, 2 (2017), 93.
- [5] V. W. Zue and J. R. Glass. 2000. Conversational interfaces: advances and challenges. *Proc. IEEE* 88, 8 (Aug. 2000), 1166–1180. <https://doi.org/10.1109/5.880078>