# User Perceptions of a Conversational Robot Interface

**Natalie Parde**
University of Illinois at Chicago
Chicago, IL, USA
parde@uic.edu

**Rodney D. Nielsen**
University of North Texas
Denton, TX, USA
rodney.nielsen@unt.edu

## ABSTRACT

Spoken dialogue systems can manifest in a variety of media, including personal devices, smart home hubs, and robots. We analyze the outcomes of a usability evaluation focused on the latter; specifically, a conversational robot that discusses books. We find that factors closely correlated with likeability include users' abilities to understand the robot, and their perceptions of its intelligence. We recommend that frameworks for future conversational robots prioritize dialogue that challenges users cognitively.

## KEYWORDS

spoken dialogue systems; conversational systems; human-robot interaction

## INTRODUCTION

Speech interfaces are increasingly implemented to achieve both task-based and conversational goals. We evaluate the usability of a companion robot designed to engage users in spoken dialogue about creative metaphors in books [3]. We analyze its conversational performance in terms of nine factors (Table 1), and find (Table 2) that likeability (S5) is positively correlated with measures of clarity (S1), intelligence (S6), and inclination for regular use (S4). Intriguingly, there also exists a strong correlation between dialogue naturalness (S7) and the robot's capacity to identify interesting discussion topics (S8). Although many conversational interfaces focus on surface-level small talk, these findings suggest that users desire deeper discussions with their artificial companions.

| Statement | Mode | Median |
|---|---|---|
| S1: I found Grace easy to understand. | 4 | 4.0 |
| S2: I knew what I could say or do at each point of the dialogue. | 3 | 3.5 |
| S3: The system worked the way I expected. | 4 | 4.0 |
| S4: I would like to use this system regularly. | 3 | 3.5 |
| S5: I like interacting with Grace. | 5 | 4.5 |
| S6: Grace seems smart. | 3 | 3.0 |
| S7: Grace's dialogue seems natural. | 2 | 3.0 |
| S8: Grace asked interesting questions about the text we were discussing. | 3 | 3.5 |
| S9: It made sense for Grace to ask the questions we discussed. | 4 | 4.0 |

**Sidebar 1: Mode and median ratings for each survey statement (*n*=26).**

*CHI 2019, May 4–9, 2019, Glasgow, Scotland UK*

| | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 |
|---|---|---|---|---|---|---|---|---|---|
| **S1** | - | .6 | .5 | .6 | .8† | .6 | .6 | .6 | .5 |
| **S2** | | - | .4 | .5 | .6 | .3 | .2 | .2 | .3 |
| **S3** | | | - | .2 | .4 | .1 | .2 | .0 | .4 |
| **S4** | | | | - | .8† | .6 | .6 | .6 | .2 |
| **S5** | | | | | - | .8† | .4 | .5 | .2 |
| **S6** | | | | | | - | .3 | .6 | .2 |
| **S7** | | | | | | | - | .7† | .5 |
| **S8** | | | | | | | | - | .6 |
| **S9** | | | | | | | | | - |

**Sidebar 2: Correlation between survey statements († indicates statistical significance, confidence interval is $r \pm 0.51$).**

## METHODS

We conducted our evaluation using the *Reading with Robots* platform [3]. The platform utilizes a NAO[1] robot named Grace to engage users in conversation about novel metaphors [5] in books, automatically generating questions about those metaphors using the *Questioning the Author* framework [2, 4]. Interactions are conducted via spoken dialogue; we logged participant dialogue using a Wizard-of-Oz approach to avoid confounding factors associated with automatic speech recognition performance. We recruited 26 college students (12 male, 14 female) to interact with the robot in half-hour sessions.[2] Participants discussed *Pride and Prejudice* [1] with the robot, and upon finishing the session they completed a paper-based survey containing nine Likert-scale questions (answers ranged from 1 (Strongly Disagree) to 5 (Strongly Agree)). Participants were compensated with $10 gift cards.

## EVALUATION

Mode and median responses to each survey statement are shown in Table 1, and Pearson's correlation ($r$) between participants' responses to the statements are shown in Table 2. Correlations were considered statistically significant (indicated as such in Table 2 using †) when $p < \alpha'$, where $p$ was the likelihood that uncorrelated data would produce similar $r$ values and $\alpha'$ was the adjusted $\alpha$ ($\alpha = 0.05$) computed using Rom's correction [6]. The highest-scoring survey statements were S1, S3, S5, and S9. We found positive, statistically significant correlations between the following statement pairs: {(S1, S5), (S4, S5), (S5, S6), (S7, S8)}.

## CONCLUSIONS

The results provide insight into the factors that participants associated most with robot likeability—its clarity (S1) and intelligence (S6),[3] indicating that improved speech synthesis and the incorporation of additional domain expertise may be worthwhile subjects of investigation for next-generation conversational robots. They also suggest an intrinsic link between language generation quality and an ability to capture user interest. Overall, these findings lend support to the notion that likeability is intertwined with deeper discussions reminiscent of human-human conversations, rather than surface-level small talk. We recommend that future conversational robot frameworks prioritize realistic, cognitively challenging dialogue.

## ACKNOWLEDGMENTS

This material was based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant 1144248, and the National Science Foundation under Grant 1262860. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

[1] https://www.softbankrobotics.com/emea/en/nao

[2] Although some participants completed additional follow-up sessions, we focus only on the first session here.

[3] The correlation between S4 and S5 is unsurprising but confirms the importance of likeability.

## REFERENCES

[1] Jane Austen. 2004. *Pride and Prejudice*. Barnes  Noble Classics.

[2] Isabel L. Beck and Margaret G. McKeown. 2006. *Improving Comprehension with Questioning the Author: A Fresh and Expanded View of a Powerful Approach*. Scholastic. https://books.google.com/books?id=gcw0AAAACAAJ

[3] Natalie Parde. 2018. Reading With Robots: Towards a Human-Robot Book Discussion System for Elderly Adults. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18) Doctoral Consortium*. Association for the Advancement of Artificial Intelligence, New Orleans, Louisiana. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16173

[4] Natalie Parde and Rodney D. Nielsen. 2018. Automatically Generating Questions about Novel Metaphors in Literature. In *Proceedings of the 11th International Conference on Natural Language Generation*. Association for Computational Linguistics, 264–273. http://aclweb.org/anthology/W18-6533

[5] Natalie Parde and Rodney D. Nielsen. 2018. Exploring the Terrain of Metaphor Novelty: A Regression-based Approach for Automatically Scoring Metaphors. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*. Association for the Advancement of Artificial Intelligence, New Orleans, Louisiana. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/download/16120/16089

[6] Dror M. Rom. 1990. A Sequentially Rejective Test Procedure Based on a Modified Bonferroni Inequality. *Biometrika* 77, 3 (1990), 663–665. http://www.jstor.org/stable/2337008