Towards Personalization of Spoken Dialogue System Communication Strategies

Carla Gordon, Kallirroi Georgila, Volodymyr Yanov and David Traum

Abstract This study examines the effects of 3 conversational traits – Register, Explicitness, and Misunderstandings – on user satisfaction and the perception of specific subjective features for Virtual Home Assistant spoken dialogue systems. Eight different system profiles were created, each representing a different combination of these 3 traits. We then utilized a novel Wizard of Oz data collection tool and recruited participants who interacted with the 8 different system profiles, and then rated the systems on 7 subjective features. Surprisingly, we found that systems which made errors were preferred overall, with the statistical analysis revealing error-prone systems were rated higher than systems which made no errors for all 7 of the subjective features rated. There were also some interesting interaction effects between the 3 conversational traits, such as implicit confirmations being preferred for systems employing a "conversational" Register, while explicit confirmations were preferred for systems employing a "formal" Register, even though there was no overall main effect for Explicitness. This experimental framework offers a fine-grained approach to the evaluation of user satisfaction which looks towards the personalization of communication strategies for spoken dialogue systems.

Carla Gordon
USC Institute for Creative Technologies, 12015 Waterfront Drive, Los Angeles, CA 90094 e-mail: cgordon@ict.usc.edu

Kallirroi Georgila
USC Institute for Creative Technologies, 12015 Waterfront Drive, Los Angeles, CA 90094 e-mail: kgeorgila@ict.usc.edu

Volodymyr Yanov
USC Institute for Creative Technologies, 12015 Waterfront Drive, Los Angeles, CA 90094 e-mail: yanov@ict.usc.edu

David Traum
USC Institute for Creative Technologies, 12015 Waterfront Drive, Los Angeles, CA 90094 e-mail: traum@ict.usc.edu

1 Introduction

Virtual assistant dialogue systems are becoming ubiquitous, as a growing number of people interact with systems such as Apple Siri, Microsoft Cortana, and Google assistant on their smart phones using natural language. Additionally, Virtual Home Assistant systems (VHAs) such as Amazon Alexa and Google Home are being welcomed into increasingly more homes across the world. These systems have become more conversational and human-like than traditional task-oriented dialogue systems, featuring human voices, and special features outside the realm of virtual assistant duties, such as joke telling.

A 2018 Google study suggests that 72% of consumers who own a VHA use it as part of their daily routine, and 41% talk to their VHA as if it were a friend or another human. The study points to the users' use of pleasantries such as "please" and "thank you", and even "sorry" to illustrate the extent to which these systems are perceived as more of a companion than a machine [9]. Communication strategies can vary widely in human-human interaction depending on a number of variables such as age, and cultural or socioeconomic background. For example, Linguist Deborah Tannen has an exhaustive body of research focused on the study of gender differences alone [12, 13, 14, 15]. The fact that human-human communication is so diverse in style and strategy, and humans are increasingly interacting with VHAs in a more naturalistic way, underscores the need to move away from a "one size fits all" model of communication strategy for dialogue systems in the VHA domain, where a significant percentage of users conceptualize interactions to be more akin to human-human communication. Indeed, previous research has shown a dichotomy between users who prefer the system to be more conversational, and those who prefer the formal approach [5].

This suggests that the future of VHA design would benefit from the ability to allow users to personalize the communication strategy of their VHA to better suit their own. To accomplish this, a finer-grained approach to the evaluation of user satisfaction in VHAs is warranted, to tease apart exactly which communicative traits and behaviors are responsible for creating the appearance of Intelligence or Naturalness, and how the interaction of these traits affects user satisfaction. Are people more willing to forgive system misunderstandings if it speaks in a more conversational register? Would a more conversational system seem more intelligent if it provided implicit confirmations instead of explicit ones? These are the kinds of questions which this study seeks to answer.

2 Related Work

Historically, much of the research on evaluating user satisfaction with dialogue systems has focused more on objectively quantifiable measures such as task completion and the length of the interaction. However, for the past few decades, the focus has shifted to the evaluation of more subjective measures. The PARADISE framework

[16] is a well known dialogue evaluation framework which has been used by many researchers to optimize desired qualities such as user satisfaction [10]. Likert scale questionnaires have also been used to evaluate user satisfaction [11] as well as more complex questionnaires, such as the SASSI questionnaire [7].

Even more recently, attention has been paid to evaluating the specific subjective features which contribute to overall user satisfaction. A review of several studies which have focused on evaluating subjective user feedback on user satisfaction have revealed a set of subjective features which have frequently been mentioned by users: Intelligence, Personality, Friendliness, and Naturalness [1, 4, 8]. However, in these previous studies no attempt was made to determine what specific system behaviors give it the appearance of friendliness or intelligence.

By their very nature, it is more difficult to evaluate specific subjective features, such as those described above, than objective measures like word error rate. Researchers have used Likert scales to quantify the degree to which a dialogue system is perceived as intelligent or pleasant, but the subjective nature of these terms makes it difficult to extrapolate exactly which communicative traits and system behaviors are responsible for the user's perception of these features. It is, therefore, necessary not only to analyze which subjective features users find most agreeable, or to what extent they feel these features are present, but to discover what combination of behaviors the system exhibits that leads the user to perceive them as more satisfactory.

A recent study suggested explicit confirmations of user requests have a strong inverse correlation with the perception of Pleasantness, Naturalness, and Personality [5]. It also suggested that a more conversational register has a positive correlation with Personality, but a negative correlation with Intelligence. Additionally, another study found a predictably strong negative correlation between system misunderstandings and overall user satisfaction [2]. A more recent study focused on predicting a number of subjective features from real user dialogues and dialogue participants' ratings or dialogue observers' ratings versus simulated dialogues and dialogue observers' ratings [3]. Interestingly, this study suggested that for 3 subjective features (Intelligence, Naturalness, Overall Quality) learning prediction functions of ratings from simulated data only could be sufficient.

These previous studies informed the use of Register, Explicitness, and Misunderstandings as the set of communicative features combined to create the system profiles for the current study.

3 Experimental Design

In order to determine how each of these communicative traits interacts with the others to affect the overall user experience, as well as the perception of more specific subjective features, we developed a set of 8 system "profiles" each with a different combination of traits. Participants were recruited to interact with and rate these system profiles via a novel Wizard of Oz data collection tool. This produced a corpus of human-system dialogues with ratings for 7 subjective features such as Naturalness

User	(1) Turn up the volume in the bathroom speaker.
System	(2) Roger that.
User	(3) A little bit more, please.
System	(4) Done.
User	(5) And turn off the washer in the garage.
System	(6) I am on it, Sir.

Table 1 A dialogue with a conversational system.

User	(1) Hey, connect the speaker to bluetooth.
System	(2) The speaker is already connected.
User	(3) Please set the washer to rinsing mode.
System	(4) The washing mode is now set to rinsing.

Table 2 A dialogue with a formal system.

and Intelligence (see section 3.2 for a more detailed discussion of these subjective features).

3.1 Communicative Traits

In order to make analysis of these traits more manageable, each trait was treated as binary, rather than a continuum. Therefore, each system was either "conversational" or "formal" in Register, with no distinctions made as to different levels of formality.

Register refers to to the presence or absence of informal, colloquial speech. Systems which were assigned a "conversational" Register would frequently use informal language such as "sure thing" or "I'll get right on that" when responding to user requests. An example of a dialogue with a conversational system can be seen in Table 1. In contrast, systems which were assigned a "formal" Register would never use this kind of informal language, as exemplified by the dialogue in Table 2.

Explicitness refers to whether or not the system gave explicit responses to user requests. Systems which were assigned to the "explicit" group always gave explicit confirmations to user requests, as can be seen in line (4) of the dialogue in Table 2. Likewise, systems which were assigned to the "implicit" group only ever gave implicit confirmations to user requests, as in line (2) of the dialogue in Table 1. Instead of the system explicitly confirming that it has turned up the volume for the bathroom speaker, it gives a general acknowledgement that the user's request has been heard, with the implication that their request will be carried out.

Misunderstandings refer to whether or not the system would make errors indicative of misunderstandings due to a failure in speech recognition or natural language understanding. Systems which were assigned to the "error" group would make at least one error per dialogue, and would recover from that error during the next dia-

User	(1) Connect the AC to the WiFi in the living room.
System	(2) Of course! I connected the WiFi for the living room TV.
User	(3) Turn WiFi off TV in the living room.
System	(4) Sure, I disconnected the WiFi for the living room TV.
User	(5) Turn on the AC WiFi in the living room.
System	(6) I'm on it, turning on the WiFi for the AC in the living room.

Table 3 A dialogue containing a system misunderstanding.

System	**Register**	**Explicitness**	**Errors**
Monkey	Formal	Explicit	Yes
Elephant	Formal	Explicit	No
Giraffe	Formal	Implicit	Yes
Rabbit	Formal	Implicit	No
Kangaroo	Conversational	Explicit	Yes
Raven	Conversational	Explicit	No
Squirrel	Conversational	Implicit	Yes
Lion	Conversational	Implicit	No

Table 4 System profiles and associated communicative traits.

logue turn. An example of a system error can be seen in line (2) of Table 3. Systems which were in the "no error" group would not make any such errors during their interactions with participants.

Profiles These 3 communicative traits were combined to create our 8 system profiles for this study. Table 4 shows the distribution of the traits across the system profiles. Each system profile was given an animal name for reference, and was comprised of a unique combination of the 3 communicative traits. Please see the Appendix section for dialogue examples which illustrate the different communication strategies of the 8 profiles.

3.2 Subjective Features

To gauge the effect of these communicative traits on the user experience, 7 subjective features were chosen to be rated by participants:

- Intelligence
- Friendliness
- Naturalness
- Personality
- Enjoyableness
- Likelihood to Recommend
- Overall Quality

Intelligence, Friendliness, Naturalness, Personality The first 4 features were chosen based on a review of studies designed to evaluate the subjective user experience of interacting with spoken dialogue systems. These are features that are frequently mentioned by users in subjective user feedback as having an effect on overall user satisfaction (see section 2). Additionally, some studies have suggested some of these features, such as Naturalness, may be hard to tie to a specific set of system behaviors, while others like Personality and Intelligence may be at odds with each other, and maximizing one may mean sacrificing the other [5].

Enjoyableness, Likelihood to Recommend, Overall Quality The last 3 general subjective features were chosen as a means of measuring different facets of the user experience, in order to make a more nuanced analysis of overall user satisfaction possible.

3.3 Data Collection Tool

Collecting dialogues for research in this domain is not a trivial task. The best source of data would come from live interaction between participants and a VHA, in an actual home environment where the participant can see and hear if their requests are being carried out properly. Data collected without this environmental context might be less representative of real user interaction, because it would not be clear when the system makes a mistake that does not appear in an explicit confirmation, or if other side-effects impact the dialogue. However, such an experimental setup would be logistically challenging and costly to carry out.

To solve these challenges we made use of a novel Wizard of Oz (WOz) data collection tool which seeks to emulate a real-world home setting [6]. This framework consists of a set of interconnected, web-based GUIs, one for participants and two for the Wizard.

The **User View Interface** (see Figure 1) displays information designed to emulate a virtual home environment. The rooms and their accompanying devices are displayed in the middle of the screen. Each device displays state information such as whether it is on or off, and the settings of its various features. Changes to these settings are shown to the user in real time. At the top of the screen a task is displayed that the user must complete by communicating with a VHA using the text chat function below the device display. In the upper right hand corner the system profile is represented by a picture of an animal. In order to control for a possible confounding influence of animal preference, the animals displayed to the user were randomized by the system between participants, rather than each user seeing the same correspondence between animal and communicative traits presented in Table 4. In this way, one participant's *Monkey* system may use the communication profile of the *Elephant* system, while another participant's *Monkey* system may employ the communication profile of the *Rabbit* system.

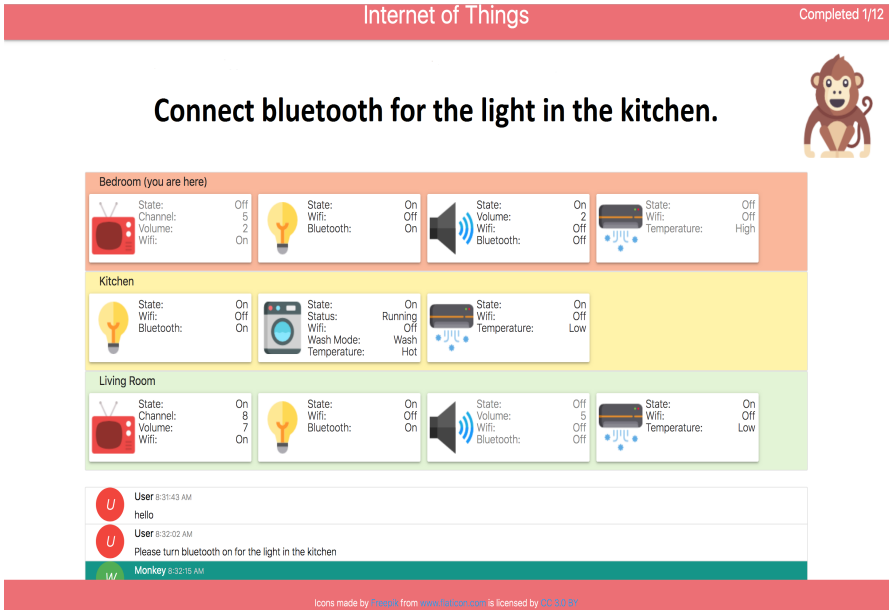

Fig. 1 The User View Interface of the WOz tool.

The **Wizard View Interface** displays the same virtual home environment as the User View Interface, including updates to device settings, but with some additional information needed by the Wizard. This information includes the system profile information, as well as other information not relevant to the current study.

The **Wizard Control Interface** (see Figure 2) allows the Wizard to control device states in the virtual home environment, and also to communicate with the user via template-based GUI buttons. A toggle menu allows the Wizard to switch between 4 different Wizard Control screens, each with a different set of utterance templates which conform to the different combinations of Register and Explicitness which comprise the 8 system profiles. The system profile information serves as a guide for the Wizard. Thus when the profile calls for "conversational" and "implicit" system behavior the Wizard will use the buttons that generate appropriate language for these system behaviors. Misunderstandings were created artificially by the Wizard, by intentionally performing actions that were not congruent with user requests.

4 Data Collection

Participants Eighteen participants were recruited using craigslist.com. As our primary concern was the collection of system ratings data, we did not collect any per-

Fig. 2 The Wizard Control Interface of the WOz tool.

sonal identifying or demographic information from the participants, other than verifying that each was over the age of 18.

Interaction Participants were seated in a quiet and distraction-free environment, engaged in IoT dialogues using the User View Interface of the WOz tool on a Macbook Pro laptop. They were instructed to imagine they were in a home environment and were talking to a VHA in order to accomplish small tasks around the home. Each participant interacted with only 4 of the 8 system profiles, to ensure there was enough interaction with each profile for participants to be able to evaluate them properly.

The WOz tool generated a randomized set of 12 out of a list of 36 pre-determined tasks; 3 tasks per system profile. These tasks were displayed one at a time to the user via the User View Interface. Participants were required to communicate with the VHA via the text chat function, and command it to complete these tasks. Tasks varied based on the type and number of devices involved, the number of rooms involved, and whether the task was to be completed now, or scheduled for a future time. They also ranged from simple requests such as "Turn on the TV in the living room." to slightly more complex requests such as "In 3 minutes, turn off the AC in the bedroom, and turn on the TV in the living room. Set the TV channel to 7.".

Participants were able to see updates to device settings in real time as the Wizard carried out their requests. The Wizard also monitored task completion, and could prevent participants from proceeding to the next task until they had completed the current task. To achieve this, when participants clicked the "next" button, indicating

their desire to move on to the next task, the Wizard would receive a notification that allowed them to reject the participant's request to continue. In this case, a pop-up message would be shown to the user stating "Please make sure the task is complete before moving on to the next task.". Finally, participants were not informed of the WOz nature of the study until after they had completed their interactions.

Questionnaires There were 3 types of questionnaires administered to participants:

- A **Pre-interaction Questionnaire** was administered once to each participant at the beginning of their interaction. This questionnaire collected information about their overall computer literacy, familiarity with VHAs, and whether or not they own one themselves.
- A **Post-task Questionnaire** was administered after each task the participant completed. This questionnaire asked them to rate the system's *Intelligence*, *Friendliness*, *Naturalness*, and *Personality*, as well as how much they *Enjoyed* the interaction, how likely they are to *Recommend* the system to a friend, and the *Overall Quality* of the interaction, on a 7-point Likert scale (1: low, 7:high).
- Finally, a **Post-interaction Questionnaire** was administered at the end of each participant's interaction with the WOz tool, asking them to rank the 4 systems they interacted with from best to worst.

5 Results

To examine the effects of the 3 communicative traits on user satisfaction, a multivariate two-way analysis of variance was performed, and the results are summarized below.

5.1 Main Effects

There was a significant main effect found for Errors for all measures except Naturalness. Oddly, systems which produced errors were rated more highly than systems which did not produce errors, regardless of Register and Explicitness (see p-values and means in Table 5). This contradicts much previous research, and a deeper discussion of this anomaly can be found below in Section 6.

There were no statistically significant main effects found for either Register or Explicitness. However, there were a few interesting interaction effects found.

Feature	Mean w/ Errors	Mean w/o Errors	p-value
Intelligence	5.67	5.11	.022
Friendliness	5.93	5.38	.005
Naturalness	5.75	5.42	.124
Personality	5.76	4.94	.000*
Likelihood to Recommend	5.30	4.48	.004
Enjoyableness	5.49	4.61	.001
Overall Quality	5.69	4.74	.000*

Table 5 Means and p-values for the main effect of Errors, (*) denotes highly significant p-values less than .001.

5.2 Interaction Effects

Register*Explicitness: There is a significant interaction effect on the measure of Overall Quality of the system (p = .044). An independent samples t test showed systems that were Formal and Explicit (M = 5.69) rated far better than Formal and Implicit (M = 4.9), with a p-value of .016. Conversely, Conversational systems scored higher if they were Implicit (M = 5.23) rather than Explicit (M = 5.04), although this difference was not statistically significant. Nevertheless, this suggests that if a system uses a Formal register it should give explicit confirmations, whereas if it is more Conversational implicit confirmations might be preferred.

Register*Errors: There is a significant interaction effect between Register and Errors on the measure of Personality (p = .032). For both Conversational and Formal systems, those that produced errors were rated higher (Conversational Mean = 5.9, Formal Mean = 5.63) than those that did not (Conversational Mean = 4.63, Formal Mean = 5.27), although further analysis revealed that this difference was only significant for Conversational systems ($p < .001$). Overall, systems that used a Conversational register and produced errors were rated most highly on the measure of Personality.

Explicitness*Errors: There is a significant interaction effect between Explicitness and Errors on the measures of Enjoyableness (p = .024) and Overall Quality (p = .007), with systems that made errors and were explicit receiving the highest ratings for both Enjoyableness (Mean with errors = 5.83, Mean no errors = 4.4, $p < .001$) and Overall Quality of the system (Mean with errors = 6.17, Mean no errors = 4.56, $p < .001$). The systems that did not make errors scored higher if they were implicit, both for Enjoyableness (Mean implicit = 4.83, Mean explicit = 4.39) and Overall Quality (Mean implicit = 4.92, Mean explicit = 4.57), although these differences were not statistically significant. This suggests that if a system is prone to errors, users would prefer it give explicit confirmation of user requests, whereas if a system makes fewer errors, implicit confirmations might be preferred.

Register*Explicitness*Errors: The interaction of all three conversational traits had a statistically significant effect on the system's perceived Friendliness (p = .05).

Rank	Mean Rating	System	Register	Explicitness	Errors
1	5.83	Kangaroo	Conversational	Explicit	Yes
2	5.56	Monkey	Formal	Explicit	Yes
3	5.21	Squirrel	Conversational	Implicit	Yes
4	4.97	Giraffe	Formal	Implicit	Yes
5	4.84	Elephant	Formal	Explicit	No
6	4.83	Lion	Conversational	Implicit	No
7	4.71	Rabbit	Formal	Implicit	No
8	4.06	Raven	Conversational	Explicit	No

Table 6 Ranking of system profiles, based on ratings averages for Enjoyableness, Likelihood to Recommend, and Overall Quality.

6 Discussion

There are some unexpected trends revealed in the ratings data. As illustrated in Table 6 and discussed in Section 5.1, an analysis of the ratings for Enjoyableness, Likelihood to Recommend, and Overall Quality, shows that the top 4 ranked systems are those which produced errors, and the top two are systems that were Explicit in their responses. One possible explanation for this phenomenon is that the systems were rated more favorably because they recovered quickly from their errors, and gave explicit confirmations so the user knew the errors had been addressed. As mentioned in Section 3.1, the policy was for the Wizard to recover from any errors during the next dialogue turn. This suggests that perhaps it is not a complete lack of errors, but rather the ability to recover quickly from errors that makes a system better overall.

This raises some interesting questions about what "user satisfaction" really means, and how best to evaluate it for dialogue systems in the VHA domain. The results of this study show that this sample of users prefers to interact with a VHA that makes occasional errors, as long as it recovers from them quickly, because it gives the system "personality". Indeed, the statistical analysis revealed a very strong statistical significance for errors in the ratings for Personality, with those systems that made errors (M = 5.76) averaging almost a full point higher in ratings than those that did not (M = 4.94) as can be seen in Table 5. This runs contrary to what common sense suggests, but makes more sense within the context of the 2018 Google study, cited earlier, which found almost half of people surveyed communicate with their VHA systems as if they were another human [9]. Misunderstandings are a natural part of human-human communication, so it stands to reason that a system which makes occasional errors could be seen as more human-like. This would seem to indicate that, for a significant percentage of the population, personality matters more than accuracy.

Additionally, the interaction effects suggest that there are strategies which can be employed to maximize user satisfaction based on the limitations of a certain system, or specific use case scenario. For example, if a family owns a VHA which they keep in a noisy family room and frequently misunderstands their requests, the interaction effect found between Explicitness and Errors suggests that the system would benefit

from giving explicit confirmations of user requests to maximize user satisfaction. Further, as mentioned in Section 2, previous research has suggested a divide among tested populations between those preferring a conversational system and those preferring a formal one. If a system were to employ different dialogue modules to allow users to choose between a formal or conversational style, the interaction effect between Register and Explicitness shows that user satisfaction can be improved by utilizing implicit confirmations for the conversational module, and explicit for the formal module.

7 Conclusion

We discussed the development of a fine-grained evaluation framework for VHA dialogue systems. This approach sought to examine the main and interaction effects of Register, Explicitness, and Misunderstandings on overall user satisfaction, as well as more specific subjective features, such as Personality. A surprising trend was found in which systems that made errors were rated more favorably overall, and on all specific subjective features. This suggests that the context of interaction for VHA systems may make maximizing user satisfaction more a matter of minimizing errors, and the recovery time from them, rather than eradicating them entirely. Finally, the interaction effects showed how certain communicative traits can be combined to form conversational strategies that can maximize user satisfaction in different use cases.

Future work in this area could examine not just the presence or absence of errors, but also the amount of errors and how quickly and correctly they are resolved. It would also be interesting to look more carefully at whether different users have different preferences for combinations of system traits or perhaps whether circumstances like the complexity of the dialogues have an impact.

Acknowledgements This work was funded in part by Samsung Electronics Co., Ltd., and partly supported by the U.S. Army. Statements and opinions expressed do not necessarily reflect the policy of the United States Government, and no official endorsement should be inferred.

Appendix

The following are examples of dialogues for a single task, generated by participant interactions with each of the 8 system profiles. These examples are provided as a means of illustrating the differences in interaction between the 8 system profiles.

The Task Users were presented with the following task: “Stop the washing machine in the kitchen and then turn it off, then turn the speaker volume to 9 in the living room.”.

NoError Systems Below you will find dialogue examples for the systems which did not make errors. These were the 4 worst performing systems overall.

Raven	**Conversational*Explicit*NoErrors**
User	(1) Turn off the washing machine and turn up the living room speaker to 9.
System	(2) I can't turn the washer off while it is running.
User	(3) Stop washing machine.
System	(4) I stopped the washer.
Lion	**Conversational*Implicit*NoErrors**
User	(1) Stop the washing machine in the kitchen, then turn it off.
System	(2) Sure thing! I took care of that for you.
User	(3) Increase volume on living room speaker to 9.
System	(4) You got it. I handled that.
Elephant	**Formal*Explicit*NoErrors**
User	(1) Stop the washing machine in the kitchen and turn it off, then turn the speaker volume to 9 in the living room.
System	(2) I stopped the washer.
System	(3) I turned the washer off.
System	(4) I turned the volume to 9 for the speaker in the living room.
Rabbit	**Formal*Implicit*NoErrors**
User	(1) Stop and then turn off the washing machine in the kitchen.
System	(2) Ok. Done.
User	(3) Turn the speaker volume to 9 in the living room.
System	(4) I will do that.

Table 7 Dialogue examples for NoError systems.

Error Systems Below you will find dialogue examples for the systems which did make errors. These were the 4 best performing systems overall.

It may not be immediately clear what the errors are for the Squirrel and Giraffe systems, since they only gave implicit confirmations of requests. The error in the Squirrel system is that the washing machine is only stopped, and not turned off, requiring the user to restate the request to turn it off in line 3 of the Squirrel dialogue in Table 8. The error for the Giraffe system is that the speaker volume was set to 8 instead of 9, as evidenced by the user restating their request in line 11 of the Giraffe dialogue in Table 9.

References

1. Artstein, R., Traum, D., Boberg, J., Gainer, A., Gratch, J., Johnson, E., Leuski, A.: Listen to my body: Does making friends help influence people? In: Proc. of FLAIRS. Florida, USA (2017)
2. Georgila, K., Gordon, C., Choi, H., Boberg, J., Jeon, H., Traum, D.: Toward low-cost automated evaluation metrics for Internet of Things dialogues. In: Proc. of IWSDS. Singapore (2018)

Kangaroo	**Conversational*Explicit*Errors**
User	(1) Turn off washing machine in the kitchen.
System	(2) I can't turn the washer off while it is running.
User	(3) Then stop it and turn it off.
System	(4) Sure thing! I am stopping the washer. I am turning the washer off.
User	(5) Living room speaker volume 9.
System	(6) You got it. I am turning the volume to 8 for the speaker in the living room.
User	(7) I said 9.
System	(8) Sorry! I am turning the volume to 9 for the speaker in the living room.
Squirrel	**Conversational*Implicit*Errors**
User	(1) Stop the kitchen washing machine, then power off.
System	(2) I'm on it!
User	(3) Power off washing machine.
System	(4) No problem.
User	(5) Set speaker level to 9 in living room.
System	(6) I'm taking care of that right now.

Table 8 Dialogue examples for Error systems: Conversational.

Monkey	**Formal*Explicit*Errors**
User	(1) Stop washer and turn off.
System	(2) Yes. I will stop the washer. I will turn the washer off.
User	(3) Living room speaker volume 9 please.
System	(4) I will turn the volume to 8 for the speaker in the living room.
User	(5) Volume 9.
System	(6) I will turn the volume to 9 for the speaker in the living room.
User	(7) Thank you.
System	(8) You're welcome.
Giraffe	**Formal*Implicit*Errors**
User	(1) Hello giraffe.
System	(2) Hello.
User	(3) Stop washing machine in kitchen.
System	(4) Ok, I will do that.
User	(5) Turn the washing machine in the kitchen off.
System	(6) I did that.
User	(7) Hello giraffe.
System	(8) Hello. Is there anything else I can do for you?
User	(9) Turn up speaker volume in living to 9.
System	(10) Ok. I will.
User	(11) Turn up speaker volume in living room to 9.
System	(12) Yes. I will do that.

Table 9 Dialogue examples for Error systems: Formal.

3. Georgila, K., Gordon, C., Yanov, V., Traum, D.: Predicting ratings of real dialogue participants from artificial data and ratings of human dialogue observers. In: Proc. of LREC. Marseille, France (2020)
4. Geutner, P., Steffens, F., Manstetten, D.: Design of the VICO spoken dialogue system: Evaluation of user expectations by Wizard-of-Oz experiments. In: Proc. of LREC. Las Palmas, Spain (2002)

5. Gordon, C., Georgila, K., Choi, H., Boberg, J., Traum, D.: Evaluating subjective feedback for Internet of Things dialogues. In: Proc. of SemDial:AixDial. Aix-en-Provence, France (2018)
6. Gordon, C., Yanov, V., Traum, D., Georgila, K.: A Wizard of Oz data collection framework for Internet of Things dialogues. In: Proc. of SemDial:LondonLogue. London, UK (2019)
7. Hone, K.S., Graham, R.: Towards a tool for the Subjective Assessment of Speech System Interfaces (SASSI). Journal of Natural Language Engineering **6**(3-4), 287–303 (2000)
8. Hurtig, T.: A mobile multimodal dialogue system for public transportation navigation evaluated. In: Proc. of MobileHCI. Helsinki, Finland (2006)
9. Kleinberg, S.: 5 ways voice assistance is reshaping consumer behavior - think with google (2018). URL https://www.thinkwithgoogle.com/consumer-insights/voice-assistance-consumer-experience/
10. Möller, S., Ward, N.: A framework for model-based evaluation of spoken dialog systems. In: Proc. of SIGDIAL. Columbus, Ohio, USA (2008)
11. Paksima, T., Georgila, K., Moore, J.D.: Evaluating the effectiveness of information presentation in a full end-to-end dialogue system. In: Proc. of SIGDIAL. London, UK (2009)
12. Tannen, D.: You Just Don't Understand: Women and Men in conversation. Morrow (1990)
13. Tannen, D.: Gender and Conversational interaction. Oxford University Press (1993)
14. Tannen, D.: Gender and Discourse. Oxford University Press (1994)
15. Tannen, D., Kendall, S., Adger, C.T.: Conversational Patterns across Gender, Class, and Ethnicity: Implications for Classroom Discourse, pp. 75–85. Springer Netherlands (1997)
16. Walker, M., Kamm, C., Litman, D.: Towards developing general models of usability with PARADISE. Journal of Natural Language Engineering **6**(3-4), 363–377 (2000)