Evaluating Subjective Feedback for Internet of Things Dialogues

Carla Gordon[1], Kallirroi Georgila[1], Hyungtak Choi[2], Jill Boberg[1], David Traum[1]
[1]Institute for Creative Technologies, University of Southern California, USA
[2]Samsung Electronics Co., Ltd., Korea
{cgordon,kgeorgila,boberg,traum}@ict.usc.edu
ht777.choi@samsung.com

Abstract

This paper discusses the process of determining which subjective features are seen as ideal in a dialogue system, and linking these features to objectively quantifiable behaviors. A corpus of simulated system-user dialogues in the Internet of Things domain was manually annotated with a set of *system communicative and action responses*, and crowd-sourced ratings and qualitative feedback of these dialogues were collected. This corpus of subjective feedback was analyzed, revealing that raters described top ranked dialogues as *Intelligent*, *Natural*, *Pleasant*, and as having *Personality*. Additionally, certain communicative and action responses were statistically more likely to be present in dialogues described as having these features. There was also found to be a lack of agreement among raters as to whether a direct communication style, or a conversational one was preferred, suggesting that future research and development should consider creating models for different communication styles.

1 Introduction

Objective measures such as task completion and word error rate, while of course essential to the evaluation of task-based dialogue systems, are not the only measures of system performance that should be used. Subjective judgments such as user satisfaction can also be critical, especially if users are expected to interact with the system on a regular basis. This paper focuses on evaluating subjective feedback in the Internet of Things (IoT) domain. The IoT refers to a network of home devices which are connected to the Internet, and can be controlled by a virtual home assistant (VHA) via human-system dialogue interaction. In contrast to dialogue systems designed to facilitate booking travel or restaurant reservations, these new systems occupy a more intimate space in a user's life. They are likely to be used more frequently, and to be perceived as less of a tool and more of a friend (Kleinberg, 2018). For this reason, it is important that research related to this type of dialogue systems places greater emphasis on the user's subjective interaction experience.

There are some natural dichotomies which exist in accordance with the personal communication styles of humans. Some people will prefer that the system have a "personality" and a conversational communication style, as in the example dialogue of Table 1, while others will appreciate a more formal, direct style, as in the example dialogue of Table 2.

Additionally, some people may prefer the system to be explicit in informing the user what actions it is taking (Table 2, line 2), while others may prefer the brevity achieved by more implicit confirmations (Table 1, line 4). Our analysis of a subjective feedback corpus, in conjunction with a manually annotated dialogue corpus, reveals that many of the subjective features mentioned correlate with objectively verifiable system behaviors, such as confirmation of understanding, explicit confirmations of user requests, and grammaticality of utterances. Also, the dichotomies discussed suggest that future research and development of IoT dialogue systems should take into account the user's preference of communication style.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. License details: https://creativecommons.org/licenses/by-nc-nd/4.0/

User	(1) Turn up the volume of 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.

Table 1: Example dialogue 1 (conversational communication style).

User	(1) Connect the speaker to bluetooth.
System	(2) It is already connected.
User	(3) Please set the washer to rinsing mode.
System	(4) The washing mode is now set to rinsing.
User	(5) Thanks.

Table 2: Example dialogue 2 (formal/direct communication style).

While previous research on the links between objective and subjective measures mainly focused on user satisfaction, we establish links between objective measures and more nuanced subjective judgments, namely, *Intelligence*, *Personality*, *Pleasantness*, and *Naturalness*.

2 Related Work

There is an ever-growing body of research concerned with the evaluation of dialogue systems. Most authors distinguish between "objective measures", such as word error rate (in spoken dialogue) and task completion, and "subjective measures", such as user satisfaction and perceived task completion.

PARADISE (Walker et al., 2000) is the most well-known framework for evaluating dialogue systems. PARADISE seeks to optimize a desired quality such as user satisfaction by formulating it as a linear combination of a variety of metrics, such as task success and dialogue cost (e.g., dialogue length). The advantage of this method is that once a desired quality has been formulated as a realistic evaluation function, it can be optimized by controlling the factors that affect it. In the example above, user satisfaction can be optimized by increasing task success and minimizing dialogue length. User satisfaction can be measured via survey questions on a Likert scale (Paksima et al., 2009) or more complex questionnaires, such as the SASSI questionnaire (Hone and Graham, 2000). Most researchers have used PARADISE as a method for establishing links between subjective measures (user judgments) and objective measures. For example, Möller et al. (2007) use PARADISE to establish links between user satisfaction and usability (and other user judgments resulting from the SASSI questionnaire), and objective system features. Callejas and López-Cózar (2008) use statistics to find relationships between interaction parameters (objective measures) and quality judgments (subjective measures). In some cases, subjective human ratings are used only to shed light on why automatic evaluation metrics have failed (Liu et al., 2017).

A review of several studies which have collected subjective user feedback reveals a set of frequently mentioned subjective features such as *Intelligence*, *Personality*, *Pleasantness*, and *Naturalness* (Artstein et al., 2017; Geutner et al., 2002; Hurtig, 2006), four features which were also mentioned frequently by participants in the current study. However, in these previous studies, no attempt was made to provide a more nuanced picture of what "satisfaction" means in terms of these subjective features of the interaction.

3 The IoT Dialogue Corpora

We investigated three related corpora in the IoT dialogue domain (see Table 3). Our initial corpus (Full Dialogue Corpus) consisted of roughly 6200 simulated dialogues (Georgila et al., 2018). The dialogues were written by a team of linguists to be representative of the types of interactions people will typically have with a VHA, and included information about device states before and after each dialogue turn (e.g., whether a device was on or off, or connected to WiFi). The dialogues included potential speech recognition errors leading to system misunderstandings, and instances of system clarification requests. The dialogues also represented a wide variety of devices and tasks. The devices included TV, air conditioner, washer, bulb (light), and speaker, and dialogues assumed there could be multiples of the same device in different locations (e.g., kitchen, bedroom, bathroom, etc.). Tasks could be immediate, such as "turn on the light in the bathroom", or scheduled for completion in the future, such as "turn on the air conditioner

	Full Dialogue Corpus	AMT Task Corpus	Subjective Feedback Corpus
Contents	6200+ dialogues	232 dialogues	6000+ feedback comments
Annotations	system state information	system and user behavior	subjective feedback

Table 3: Information about the corpora used for this research.

in 10 minutes". The dialogues also represented diversity in system communication style, with some dialogues presenting a system that was much more formal, and others one that was more conversational.

It should be noted that despite the best efforts of the linguists to produce data that was as realistic as possible, our corpus lacked certain natural dialogue phenomena that would have been present in real human-machine dialogues, such as pauses, mid-sentence restarts, and self-repairs (Shalyminov et al., 2017). However, the focus of this research was to discover which system behaviors users would find most favorable, and not on classification techniques for producing correct responses to user input. We, therefore, believe that this lack of the natural phenomena mentioned above does not have such a great impact on this line of inquiry as to negate the conclusions drawn from its results.

In order to carry out the crowd-sourcing evaluation, a second smaller corpus of 232 dialogues was extracted from this larger corpus (AMT Task Corpus). Care was taken to ensure that this smaller corpus was also representative of the range of interactions, tasks, and devices found in the larger corpus. The smaller corpus was divided into sets of 5 dialogues for which raters on Amazon Mechanical Turk (AMT) were asked to provide rankings and subjective feedback, in their own words. There were 4 tasks, each providing the rater with 8 sets of 5 dialogues, representing a mix of dialogue tasks and devices, as well as varying degrees of context as to the current state of the devices controlled by the VHA (e.g., "the kitchen light is on, but the bedroom light is off"). Raters ranked the sets of 5 dialogues from best (1) to worst (5), and then explained why they chose this ranking. For the purposes of this paper, we are primarily concerned with the corpus of subjective feedback produced by these tasks.

The Subjective Feedback Corpus included over 6000 individual comments from 199 raters, each associated with a ranked group of 5 dialogues from the smaller dialogue corpus (AMT Task Corpus). The feedback was written in the raters' own words in a text field provided within the ranking questionnaire, in response to the question "Why did you choose to rank the dialogues in this order? What did you like/dislike about these dialogues?". There was no limit imposed on the rater as to how much or how little they could write, and raters varied in the level of detail they provided. Some raters gave short, concise feedback, such as "the highest ranked dialogues just seemed more natural" whereas others provided much more detailed feedback, such as:

> "The first one gave me more information on what the system understood leaving me to know rather than assume it understood. The first was much more friendly. The second one was okay, but just okay. It was straight to the point and not too bad. I would think number two is acceptable. Three, four, and five I didn't like at all."

4 Qualitative Analysis of Rater Feedback

The qualitative analysis of the feedback corpus was carried out using a novel approach. This approach consisted of: (1) Analyzing the overall word frequency for the entire corpus. (2) Manually analyzing a small subset of the corpus to extract the most commonly mentioned features, both negative and positive. (3) Creating semantic clusters which correlate with the features from the previous step, based on the highest ranking words from the word frequency list (e.g., the semantic cluster for *Brevity* contains the words "short", "brief", "concise", and "quick" among others). (4) Analyzing the frequency of the words in each semantic cluster in the feedback corpus to determine how many raters mentioned it, and how often it was mentioned.

The following is a description of the most commonly mentioned features of the dialogues, indicating how many raters (out of 199) mentioned each feature at least once, as well as summarizing the raters' explanations of these features:

Misunderstandings (151) and Effectiveness (106): The most frequently mentioned feature was *Misunderstandings*, and nobody liked them. Raters were very unforgiving of misunderstandings and expected the system to recover from them quickly. An analysis conducted in Georgila et al. (2018) revealed that conversations with multiple misunderstandings were consistently ranked the lowest. In addition, over half of raters mentioned *Effectiveness* in at least one of their comments, which presumably refers to a lack of misunderstandings and correctly executing a task the first time around.

Simplicity (130) vs. Complexity (24): The next most mentioned feature was *Simplicity*. Overall, people largely preferred simple dialogues to complex ones. The words "short and sweet" appear repeatedly in the feedback corpus. In some of the comments which mention *Complexity*, raters did say they preferred the simple dialogues, but also liked that they could have a more detailed and complex conversation if they wanted to.

Confirming (111) and Responsiveness (108): The third and fourth most frequently mentioned features were *Confirming* and *Responsiveness*. Raters showed strong aversion to silence from the system, citing that a lack of responsiveness made for a poorer dialogue. Raters also consistently mentioned system confirmations and requests for clarification as positive dialogue aspects. This includes any time the system repeated rater commands to confirm them, or gave confirmation that a certain command was complete.

Naturalness (101): *Naturalness* is harder to qualify, however some raters clearly stated that this would mean that the system's responses were more "human-like", while others failed to specify what was meant by "natural". A few did mention grammatical mistakes as taking away from the naturalness of the dialogue, and a few talked about disliking "robotic" responses.

Brevity (97): Almost half of raters mention *Brevity* in their comments. Whereas most of the comments on *Simplicity* seem to suggest that a shorter overall dialogue was preferred, many of the comments about *Brevity* imply that shorter individual utterances were preferred as well.

Pleasantness (74) and Rudeness (55): Raters often mentioned words like "kind", "nice", "pleasant", and "polite" when referring to the systems they preferred, and many explicitly mentioned specific behaviors they found rude. Silence in response to user utterances was the most often mentioned rude behavior, but people also disliked when the system used words like "obviously" and "naturally". These were seen as "back talk" and "sass" on the part of the system (although, a small minority of participants expressed an affinity for the system being "sarcastic").

Personality (27) vs. Formality (28): Some raters said that they enjoyed system utterances like "roger that" and "mission complete" that gave the system more *Personality*. They said these kinds of system utterances made them laugh and would enhance their experience. Roughly the same amount of people explicitly mentioned disliking these utterances than liking them, preferring the system to use more formality when responding.

Directness (53): Based on the comments, *Directness* may tie into a number of other features such as *Brevity*, *Simplicity*, and *Formality*. Many of the comments about *Directness* also mention liking the system to be "straightforward" or "precise", and mention disliking phrases linked to the *Personality* comments such as "roger that".

Intelligence (25): Although a proportionally small number of raters explicitly mentioned *Intelligence*, there was enough precedence in previous work for it to be included in this analysis. However, due to the highly subjective nature of this feature, steps were taken to try to determine which behaviors the system displayed that led raters to describe the system as being "intelligent".

5 System Response Annotation Scheme

We created a novel annotation scheme to describe features of the system's communicative and action responses in a dialogue, in order to investigate how the qualitative features from the feedback corpus might be achieved. Some of our annotation labels were motivated by existing schemes (Core and Allen, 1997; Bunt et al., 2012), but we found no scheme that encompassed the breadth of information for which we wanted the system's utterances to be annotated. Annotations fell into 3 broad categories: **Action Assessment, Response Assessment**, and **Linguistic Feature Assessment**.

Assess Action	Assess Response					Assess Linguistic Features		
Action type	**Describe current understanding**	**Acknowledge action**	**Specify state**	**Request**	**Other**	**Specificity**	**Register**	**Grammaticality**
A-something	CU-confirm	AA-past	SS-done	Req-loc	O-null	explicit	Reg-direct	gram
A-nothing	CU-lack	AA-present	SS-NA	Req-dev	O-pleasant	implicit	Reg-conv	ungram
A-valid		AA-future	SS-unclear	Req-time				
A-invalid		AA-ANS		Req-temp				
		AA-AI		Req-other				
		AA-null		Req-action				
				Req-repeat				

Table 4: Taxonomy of annotation categories and subcategories.

Action Assessments are concerned only with determining if any action was taken by the system, and whether that action was valid or invalid, based on the user's request. The Action Assessment annotations represent the *System Action Responses*. Response Assessments are concerned with indicating what the system said to the user, and in what way it communicated that information. For example, do the system responses focus explicitly on the system's actions, or implicitly by describing the system's current state? The Linguistic Features represent the overall communication style of the system. For example, this communication can be explicit or implicit, conversational or more formal and direct. The Response Assessment and Linguistic Feature Assessment annotations comprise the *System Communicative Responses*. Some categories were further broken down into subcategories, as illustrated in Table 4.

For any given utterance, there were at least 5 annotations: an Action Type, a Response Type (Response Assessment), and assessments of Specificity, Register, and Grammaticality. The vast majority of system utterances had only 5 annotations, one from each of the above categories, but occasionally an utterance was annotated with more than one Response Type. The Response Type represents the illocutionary force (Alston, 2000) of a particular system utterance – that is, what the system is trying to communicate to the user – so occasionally more than one annotation was appropriate for a given utterance if it encompassed more than one illucutionary act, such as in the following example: "I couldn't understand, which TV would you like to link to the network?". This utterance was annotated with "CU-lack" indicating that the system lacked an understanding of the user's request (first illocutionary act) and also "Req-loc" since the system requested more information from the user about the location of the device for which it should take action (second illocutionary act). The Response Types "Acknowledge action" and "Specify state" represent a variety of locutionary acts all with the same intended illocutionary force: acknowledgment that the user's request has been fulfilled (or that it cannot be fulfilled). Below is a full accounting of all annotation categories for system communicative and action responses included in our annotation scheme. The full annotation scheme (including annotations of user input) is described in Georgila et al. (2018).

System Action Responses:

Assess action: Assesses what action, if any, was taken by the system for the specific utterance.

- A-something (system does something: "I'm connecting the speaker."),
- A-nothing (system does nothing: "Which speaker?"),
- A-valid (system does requested thing: "U: Turn on the kitchen light. S: I'm turning on the kitchen light."),
- A-invalid (system does not do requested thing: "U: Turn on the kitchen light. S: I'm turning on the porch light.").

System Communicative Responses:

Describe current understanding: Confirms the user's request, or informs the user that it does not understand their request.

- CU-confirm (confirm request before doing: "Do you want me to turn on the kitchen light?"),
- CU-lack (describe lack of understanding: "Sorry I don't understand.").

Acknowledge action: Explicitly acknowledges an action the system has taken, is taking, or will take in the future.

- AA-past (action specified in the past: “The light has been turned on.”),
- AA-present (action specified in the present: “I’m turning on the light.”),
- AA-future (action specified in the future: “I will turn on the light in 5 minutes.”),
- AA-ANS (action not specified: “U: Turn on the light. S: Done.”),
- AA-AI (action impossible: “I can’t open the door while the cycle is running.”),
- AA-null (action is done but not acknowledged: “U: Turn on the light. S: Anything else?”).

Specify state: Implicitly informs the user that an action has been taken, by describing the current state of the system.

- SS-done (implicit action, done: “The light is now on.”),
- SS-NA (implicit action, not applicable: “The light is already on.”),
- SS-unclear (implicit action, unclear: “The light is on.” – it is not clear whether the light was already on or the system performed the action).

Requests: Requests more information from the user.

- Req-loc (missing parameter, location: “Which light?”),
- Req-dev (missing parameter, device: “What should I connect to WiFi?”),
- Req-time (missing parameter, time: “When should I do that?”),
- Req-temp (missing parameter, temperature: “What temperature do you want?”),
- Req-other (missing parameter, other: “What should I connect it to?”),
- Req-action (request more actions: “Is there anything else I can do for you?”),
- Req-repeat (request repeat: “Could you repeat?”).

Other response: Represents system behaviors which do not fit into other categories.

- O-null (equivalent to silence),
- O-pleasant (system pleasantry: “You are welcome.”).

Specificity: Indicates the level of specificity of a given utterance.

- explicit (parameters explicit: “U: Turn on the light. S: The light has been turned on.”),
- implicit (parameters implicit: “U: Turn on the light. S: It has been turned on.”).

Register: Refers to the conversational style of the utterance.

- Reg-direct (direct: “U: Turn on the light. S: I’m turning on the light.”),
- Reg-conv (conversational: “U: Turn on the light. S: Sure thing, the light is now on.”).

Grammaticality: Indicates whether or not a specific utterance was grammatical.

- gram (grammatical: “Which light would you like me to turn on?”),
- ungram (ungrammatical: “What temperature you want me to fix?”).

The above annotations were used as a means to determine which system communicative and action responses may be responsible for the perception of the subjective features mentioned in the raters’ feedback, as discussed in section 4.

Annotation	**Mean (high)**	**Mean (low)**	**p-value**	η^2
Intelligence (# of dialogues: High = 69, Low = 74)				
A-something, A-valid	1.4638	1.4324	.009	.05
O-null	.2319	.6081	.000	.14
O-pleasant	.0000	.0541	.051	.03
ungram	.0000	.1081	.016	.04
A-nothing	1.0725	1.4324	.008	.05
Naturalness (# of dialogues: High = 178, Low = 187)				
O-null	.3503	.4866	.011	.02
CU-Confirm	.2486	.3476	.058	.01
Personality (# of dialogues: High = 51, Low = 56)				
Reg-conv	1.0784	.6786	.023	.05
CU-confirm	.1569	.3393	.029	.05
CU-lack	.1765	.0179	.005	.07
O-null	.2745	.5714	.003	.08
Pleasantness (# of dialogues: High = 144, Low = 161)				
A-nothing	1.0903	1.3540	.005	.03
CU-confirm	.2500	.3665	.035	.02
O-null	.3194	.5217	.001	.04

Table 5: Statistical Analyses: results of Mann-Whitney U tests.

6 Relationship of System Behaviors to Subjective Features

Some subjective features are easier than others to relate to system behaviors. In the case of *Effectiveness*, it is reasonably safe to assume that a system which is capable of completing a requested task would be seen as effective. Likewise, in the case of *Brevity* it is clear that shorter dialogues (or those with shorter utterances) will rank higher on this measure. However, certain subjective features such as *Intelligence*, *Pleasantness*, *Naturalness*, and *Personality* pose a much larger problem in determining which behaviors should be displayed by the system in order to give the appearance of possessing these qualities.

To address this issue, an analysis was conducted which compared the group of highest ranked dialogues to the group of lowest ranked dialogues. That is, for each set in which a rater mentioned a specific feature (e.g., *Intelligence*), the “high” group contains only the highest ranked dialogue, and the “low” group contains only the lowest ranked dialogue. For each group, the total number of occurrences of each annotation was calculated for each dialogue, and the groups were then compared to determine if there was a statistically significant difference in the presence of each behavior. The results of the statistical analyses are summarized in Table 5.

Dialogues ranked highest on *Intelligence* had statistically more valid actions (A-something, A-valid), fewer silences (O-null), fewer pleasantries (O-pleasant), fewer ungrammatical utterances (ungram), and fewer instances of doing nothing (A-nothing). Overall, the only system behavior with a large effect size was system silences; other variables show small to medium effect sizes, indicating that system responsiveness is heavily tied to rater perceptions of *Intelligence*.

Regarding *Naturalness*, the only communicative response for which statistical significance was found was system silences (O-null), and even then the effect size is small. However, this implies that the system should respond to all user utterances, not just questions, in order to appear more “natural”. If the system asks a question and the user says “no” the system should follow up with another general question such as “what would you like me to do then?” instead of simply waiting for the next command. Additionally, confirmations (CU-confirm) were nearly significant, suggesting that dialogues which too frequently confirm user commands (e.g., “should I turn on the light in the living room?”) may be perceived as less natural. Overall, these analyses suggest that *Naturalness* might be particularly difficult to evaluate, perhaps because of competing interpretations of what makes a system seem natural.

Correlated Features	Pearson's r	Correlated Features	Pearson's r
Personality-Pleasantness	.12	Pleasantness-Naturalness	.53***
Personality-Naturalness	.26***	Pleasantness-Intelligence	.30***
Personality-Intelligence	.07	Naturalness-Intelligence	.37***

Table 6: Pairwise correlations (Pearson's r) for *Personality*, *Pleasantness*, *Naturalness*, and *Intelligence* (***: p<.001, **: p<.01, *: p<.05).

Dialogues in which the system was described as having *Personality* were statistically more likely to use a conversational register than a direct one. This result seems intuitive, but it is somewhat surprising that conversational register has the lowest effect size (together with CU-confirm) out of all of the behaviors listed. A factor affecting the perception of system personality to a greater degree was informing the user of the system's lack of understanding (CU-lack), such as "sorry I don't know what you want". Confirmations of understanding (CU-confirm) were associated with lower ranked dialogues, while indicating a lack of understanding was associated with more highly ranked dialogues. In addition, the highest ranked dialogues also had fewer silences (O-null) than the lowest ranked dialogues.

Dialogues ranked as the most *Pleasant* had fewer confirmations of user requests (CU-confirm), and fewer silences (O-null), much like those described as more natural or as having a personality. However, *Pleasantness* was also associated with fewer system utterances in which no action was taken (A-nothing). It is worth noting that none of these behaviors shows a particularly high effect size, indicating that, like *Naturalness*, it may be hard to find a fixed set of features which represent *Pleasantness*, due to competing interpretations of what a pleasant system is.

Table 6 shows pairwise Pearson's correlations for *Personality*, *Pleasantness*, *Naturalness*, and *Intelligence*. These correlations have been calculated based on feedback for the highest ranked dialogues in each set of 5 dialogues presented to the raters.

The only feature which did not correlate significantly with *Intelligence* is *Personality*. This may be indicative of the dichotomy mentioned between those who prefer a more conversational system and those who prefer a more direct system. Dialogues ranked highly on the measure of *Intelligence* contained statistically fewer pleasantries, which is indicative of a more direct communication style, whereas the qualitative analysis revealed that dialogues described as having *Personality* frequently used more conversational utterances such as "roger that".

7 Conclusion and Future Work

The preceding analysis sought to gather subjective rater feedback, in the raters' own words, and evaluate that feedback to determine what subjective features were found as most favorable. It sought also to determine which system communicative and action responses were most closely correlated with the set of subjective features (*Intelligence*, *Naturalness*, *Pleasantness*, *Personality*) mentioned frequently by raters in the current study, and in previous literature. From the above analysis, it is clear that subjective features such as *Intelligence* can be analyzed to determine which system communicative and action responses are likely to give raters the impression that the system possesses these qualities, even though for certain features such as *Naturalness* and *Pleasantness*, this task may be more difficult.

Further study is needed on quantifying the degree to which these subjective measures were perceived in the dialogues. Additionally, as the qualitative analysis suggests, there is a need for future research to determine what behaviors correlate most with different communication styles, so that dialogue systems can be tailored to users' preferences. Finally, we are currently in the process of validating the above findings with real dialogues and user feedback rather than simulated dialogues and rater feedback.

Acknowledgments

This work was funded by Samsung Electronics Co., Ltd., and partly supported by the U.S. Army Research Laboratory. Any statements or opinions expressed in this material are those of the authors and do not necessarily reflect the policy of the U.S. Government, and no official endorsement should be inferred.

References

William P. Alston. 2000. *Illocutionary acts and sentence meaning*. Cornell University Press, Ithaca, NY.

Ron Artstein, David Traum, Jill Boberg, Alesia Gainer, Jonathan Gratch, Emmanuel Johnson, and Anton Leuski. 2017. Listen to my body: Does making friends help influence people? *Proceedings of the International Florida Artificial Intelligence Research Society Conference (FLAIRS).*

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David Traum. 2012. ISO 24617-2: A semantically-based standard for dialogue annotation. *Proceedings of the International Conference on Language Resources and Evaluation (LREC).*

Zoraida Callejas and Ramón López-Cózar. 2008. Relations between de-facto criteria in the evaluation of a spoken dialogue system. *Speech Communication* 50(8–9):646–665.

Mark G. Core and James F. Allen. 1997. Coding dialogs with the DAMSL annotation scheme. *Proceedings of the AAAI Fall Symposium on Communicative Action in Humans and Machines.*

Kallirroi Georgila, Carla Gordon, Hyungtak Choi, Jill Boberg, Heesik Jeon, and David Traum. 2018. Toward low-cost automated evaluation metrics for Internet of Things dialogues. *Proceedings of the International Workshop on Spoken Dialogue Systems Technology (IWSDS).*

Petra Geutner, Frank Steffens, and Dietrich Manstetten. 2002. Design of the VICO spoken dialogue system: Evaluation of user expectations by Wizard-of-Oz experiments. *Proceedings of the International Conference on Language Resources and Evaluation (LREC).*

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):287–303.

Topi Hurtig. 2006. A mobile multimodal dialogue system for public transportation navigation evaluated. *Proceedings of the Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI).*

Sara Kleinberg. 2018. 5 ways voice assistance is shaping consumer behavior. Retrieved from *https://www.thinkwithgoogle.com/consumer-insights/voice-assistance-consumer-experience/.*

Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2017. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Sebastian Möller, Paula Smeele, Heleen Boland, and Jan Krebber. 2007. Evaluating spoken dialogue systems according to de-facto standards: A case study. *Computer Speech and Language* 21(1):26–53.

Taghi Paksima, Kallirroi Georgila, and Johanna D. Moore. 2009. Evaluating the effectiveness of information presentation in a full end-to-end dialogue system. *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL).*

Igor Shalyminov, Arash Eshghi, and Oliver Lemon. 2017. Challenging neural dialogue models with natural data: Memory networks fail on incremental phenomena. *Proceedings of the Workshop on the Semantics and Pragmatics of Dialogue (SemDial).*

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 2000. PARADISE: A framework for evaluating spoken dialogue agents. *Proceedings of the Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics (ACL/COLING).*