Comparing Objective and Subjective Measures of Usability in a Human-Robot Dialogue System

Mary Ellen Foster and Manuel Giuliani and Alois Knoll
Informatik VI: Robotics and Embedded Systems
Technische Universität München
Boltzmannstraße 3, 85748 Garching bei München, Germany
{foster,giuliani,knoll}@in.tum.de

Abstract
We present a human-robot dialogue system that enables a robot to work together with a human user to build wooden construction toys. We then describe a study in which naïve subjects interacted with this system under a range of conditions and then completed a user-satisfaction questionnaire. The results of this study provide a wide range of subjective and objective measures of the quality of the interactions. To assess which aspects of the interaction had the greatest impact on the users’ opinions of the system, we used a method based on the PARADISE evaluation framework (Walker et al., 1997) to derive a performance function from our data. The major contributors to user satisfaction were the number of repetition requests (which had a negative effect on satisfaction), the dialogue length, and the users’ recall of the system instructions (both of which contributed positively).

1 Introduction
Evaluating the usability of a spoken language dialogue system generally requires a large-scale user study, which can be a time-consuming process both for the experimenters and for the experimental subjects. In fact, it can be difficult even to define what the criteria are for evaluating such a system (cf. Novick, 1997). In recent years, techniques have been introduced that are designed to predict user satisfaction based on more easily measured properties of an interaction such as dialogue length and speech-recognition error rate. The design of such performance methods for evaluating dialogue systems is still an area of open research.

The PARADISE framework (PARAdigm for DIalogue System Evaluation; Walker et al. (1997)) describes a method for using data to derive a performance function that predicts user-satisfaction scores from the results on other, more easily computed measures. PARADISE uses stepwise multiple linear regression to model user satisfaction based on measures representing the performance dimensions of task success, dialogue quality, and dialogue efficiency, and has been applied to a wide range of systems (e.g., Walker et al., 2000; Litman and Pan, 2002; Möller et al., 2008). If the resulting performance function can be shown to predict user satisfaction as a function of other, more easily measured system properties, it will be widely applicable: in addition to making it possible to evaluate systems based on automatically available data from log files without the need for extensive experiments with users, for example, such a performance function can be used in an online, incremental manner to adapt system behaviour to avoid entering a state that is likely to reduce user satisfaction, or can be used as a reward function in a reinforcement-learning scenario (Walker, 2000).

Automated evaluation metrics that rate system behaviour based on automatically computable properties have been developed in a number of other fields: widely used measures include BLEU (Papineni et al., 2002) for machine translation and ROUGE (Lin, 2004) for summarisation, for example. When employing any such metric, it is crucial to verify that the predictions of the automated evaluation process agree with human judgements of the important aspects of the system output. If not, the risk arises that the automated measures do not capture the behaviour that is actually relevant for the human users of a system. For example, Callison-Burch et al. (2006) presented a number of
counter-examples to the claim that BLEU agrees with human judgements. Also, Foster (2008) examined a range of automated metrics for evaluation generated multimodal output and found that few agreed with the preferences expressed by human judges.

In this paper, we apply a PARADISE-style process to the results of a user study of a human-robot dialogue system. We build models to predict the results on a set of subjective user-satisfaction measures, based on objective measures that were either gathered automatically from the system logs or derived from the video recordings of the interactions. The results indicate that the most significant contributors to user satisfaction were the number of system turns in the dialogues, the users’ ability to recall the instructions given by the robot, and the number of times that the user had to ask for instructions to be repeated. The former two measures were positively correlated with user satisfaction, while the latter had a negative impact on user satisfaction; however the correlation in all cases was relatively low. At the end of the paper, we discuss possible reasons for these results and propose other measures that might have a larger effect on users’ judgements.

2 Task-Based Human-Robot Dialogue

This study makes use of the JAST human-robot dialogue system (Rickert et al., 2007) which supports multimodal human-robot collaboration on a joint construction task. The user and the robot work together to assemble wooden construction toys on a common workspace, coordinating their actions through speech, gestures, and facial displays. The robot (Figure 1) consists of a pair of manipulator arms with grippers, mounted in a position to resemble human arms, and an animatronic talking head (van Breemen, 2005) capable of producing facial expressions, rigid head motion, and lip-synchronised synthesised speech. The system can interact in English or German.

The robot is able to manipulate objects in the workspace and to perform simple assembly tasks. In the system that was used in the current study, the robot instructs the user on building a particular compound object, explaining the necessary assembly steps and retrieving pieces as required, with the user performing the actual assembly actions. To make joint action necessary for success in the assembly task, the workspace is divided into

Figure 1: The JAST dialogue robot

| SYSTEM | First we will build a windmill. Okay? |
| USER | Okay. |
| SYSTEM | To make a windmill, we must make a snowman. |
| SYSTEM | [picking up and holding out red cube] To make a snowman, insert the green bolt through the end of this red cube and screw it into the blue cube. |
| USER | [takes cube, performs action] Okay. |
| SYSTEM | [picking up and holding out a small slat] To make a windmill, insert the yellow bolt through the middle of this short slat and the middle of another short slat and screw it into the snowman. |
| USER | [takes slat, performs action] Okay. |
| SYSTEM | Very good! |

Figure 2: Sample human-robot dialogue
two areas—one belonging to the robot and one to
the user—so that the robot must hand over some
pieces to the user. Figure 2 shows a sample dia-
logue in which the system explains to the user how
to build an object called a ‘windmill’, which has a
sub-component called a ‘snowman’.

3 Experiment Design

The human-robot system was evaluated via a user
study in which subjects interacted with the com-
plete system; all interactions were in German. As
a between-subjects factor, we manipulated two as-
pects of the generated output: the strategy used by
the dialogue manager to explain a plan to the user,
and the type of referring expressions produced by
the system. Foster et al. (2009) give the details
of these factors and describes the effects of each
individual manipulation. In this paper, we concen-
trate on the relationships among the different fac-
tors that were measured during the study: the effi-
ciency and quality of the dialogues, the users’ suc-
cess at building the required objects and at learn-
ing the construction plans for new objects, and the
users’ subjective reactions to the system.

3.1 Subjects

43 subjects (27 male) took part in this experi-
ment; the results of one additional subject were
discarded due to technical problems with the sys-
tem. The mean age of the subjects was 24.5, with a
minimum of 14 and a maximum of 55. Of the sub-
jects who indicated an area of study, the two most
common areas were Informatics (12 subjects) and
Mathematics (10). On a scale of 1–5, subjects
gave a mean assessment of their knowledge of
computers at 3.4, of speech-recognition systems
at 2.3, and of human-robot systems at 2.0. The
subjects were compensated for their participation
in the experiment.

3.2 Scenario

In this experiment, each subject built the same
three objects in collaboration with the system,
always in the same order. The first target
was a ‘windmill’ (Figure 3a), which has a sub-
component called a ‘snowman’ (Figure 3b). Once
the windmill was completed, the system then
walked the user through building an ‘L shape’
(Figure 3c). Finally, the robot instructed the user
to build a ‘railway signal’ (Figure 3d), which com-
bines an L shape with a snowman. During the con-
struction of the railway signal, the system asked
the user if they remembered how to build a snow-
man and an L shape. If the user did not remember,
the system explained the building process again; if
they did remember, the system simply told them to
build another one.

3.3 Dependent Variables

We gathered a wide range of dependent measures:
objective measures derived from the system logs
and video recordings, as well as subjective mea-
sures based on the users’ own ratings of their ex-
perience interacting with the system.

3.3.1 Objective Measures

We collected a range of objective measures from
the log files and videos of the interactions. Like
Litman and Pan (2002), we divided our objective
measures into three categories based on those used
in the PARADISE framework: dialogue efficiency,
dialogue quality, and task success.

The dialogue efficiency measures concentrated
on the timing of the interaction: the time taken to
complete the three construction tasks, the number
of system turns required for the complete interac-
tion, and the mean time taken by the system to re-
spond to the user’s requests.

We considered four measures of dialogue qual-
ity. The first two measures looked specifically for
signs of problems in the interaction, using data au-
automatically extracted from the logs: the number of times that the user asked the system to repeat its instructions, and the number of times that the user failed to take an object that the robot attempted to hand over. The other two dialogue quality measures were computed based on the video recordings: the number of times that the user looked at the robot, and the percentage of the total interaction that they spent looking at the robot. We considered these gaze-based measures to be measures of dialogue quality since it has previously been shown that, in this sort of task-based interaction where there is a visually salient object, participants tend to look at their partner more often when there is a problem in the interaction (e.g., Argyle and Graham, 1976).

The task success measures addressed user success in the two main tasks undertaken in these interactions: assembling the target objects following the robot’s instructions, and learning and remembering to make a snowman and an L shape. We measured task success in two ways, corresponding to these two main tasks. The user’s success in the overall assembly task was assessed by counting the proportion of target objects that were assembled as intended (i.e., as in Figure 3), which was judged based on the video recordings. To test whether the subjects had learned how to build the sub-components that were required more than once (the snowman and the L shape), we recorded whether they said yes or no when they were asked if they remembered each of these components during the construction of the railway signal.

### 3.3.2 Subjective Measures

In addition to the above objective measures, we also gathered a range of subjective measures. Before the interaction, we asked subjects to rate their current level on a set of 22 emotions (Ortony et al., 1988) on a scale from 1 to 4; the subjects then rated their level on the same emotional scales again after the interaction. After the interaction, the subjects also filled out a user-satisfaction questionnaire, which was based on that used in the user evaluation of the COMIC dialogue system (White et al., 2005), with modifications to address specific aspects of the current dialogue system and the experimental manipulations in this study. There were 47 items in total, each of which requested that the user choose their level of agreement with a given statement on a five-point Likert scale. The items were divided into the following categories:

| Category                          | Description                                                                 |
|-----------------------------------|-----------------------------------------------------------------------------|
| Opinion of the robot as a partner | 21 items addressing the ease with which subjects were able to interact with the robot |
| Instruction quality               | 6 items specifically addressing the quality of the assembly instructions given by the robot |
| Task success                      | 11 items asking the user to rate how well they felt they performed on the various assembly tasks |
| Feelings of the user              | 9 items asking users to rate their feelings while using the system           |

At the end of the questionnaire, subjects were also invited to give free-form comments.

### 4 Results

In this section, we present the results of each of the individual dependent measures; in the following section, we examine the relationship among the different types of measures. These results are based on the data from 40 subjects: we excluded results from two subjects for whom the video data was not clear, and from one additional subject who appeared to be ‘testing’ the system rather than making a serious effort to interact with it.

#### 4.1 Objective Measures

**Dialogue efficiency**  The results on the dialogue efficiency measures are shown in Table 1. The average subject took 305.1 seconds—that is, just over five minutes—to build all three of the objects, and an average dialogue took 13 system turns to complete. When a user made a request, the mean delay before the beginning of the system response was about three seconds, although for one user this time was more than twice as long. This response delay resulted from two factors. First, preparing long system utterances with several referring expressions (such as the third and fourth system turns in Figure 2) takes some time; second, if a user made a request during a system turn (i.e., a ‘barge-in’ attempt), the system was not able to respond until the current turn was completed.

| Measure                | Mean (Stdev) | Min | Max |
|------------------------|--------------|-----|-----|
| Length (sec)           | 305.1 (54.0) | 195.2 | 488.4 |
| System turns           | 13.4 (1.73)  | 11  | 18  |
| Response time (sec)    | 2.79 (1.13)  | 1.27 | 7.21 |

Table 1: Dialogue efficiency results
These three measures of efficiency were correlated with each other: the correlation between length and turns was 0.38; between length and response time 0.47; and between turns and response time 0.19 (all $p < 0.0001$).

**Dialogue quality** Table 2 shows the results for the dialogue quality measures: the two indications of problems, and the two measures of the frequency with which the subjects looked at the robot’s head. On average, a subject asked for an instruction to be repeated nearly two times per interaction, while failed hand-overs occurred just over once per interaction; however, as can be seen from the standard-deviation values, these measures varied widely across the data. In fact, 18 subjects never failed to take an object from the robot when it was offered, while one subject did so five times and one six times. Similarly, 11 subjects never asked for any repetitions, while five subjects asked for repetitions five or more times. On average, the subjects in this study spent about a quarter of the interaction looking at the robot head, and changed their gaze to the robot 23.5 times over the course of the interaction. Again, there was a wide range of results for both of these measures: 15 subjects looked at the robot fewer than 20 times during the interaction, 20 subjects looked at the robot between 20 to 30 times, while 5 subjects looked at the robot more than 30 times.

The two measures that count problems were mildly correlated with each other ($R^2 = 0.26, p < 0.001$), as were the two measures of looking at the robot ($R^2 = 0.13, p < 0.05$); there was no correlation between the two classes of measures.

**Task success** Table 3 shows the success rate for assembling each object in the sequence. Objects in italics represent sub-components, as follows: the first snowman was constructed as part of the windmill, while the second formed part of the railway signal; the first L-shape was a goal in itself, while the second was also part of the process of building the railway signal. The Rate column indicates subjects’ overall success at building the relevant component—for example, 55% of the subjects built the windmill correctly, while both of the L-shapes were built with 90% accuracy. For the second occurrence of the snowman and the L-shape, the Memory column indicates the percentage of subjects who claimed to remember how to build it when asked. The Overall row at the bottom indicates subjects’ overall success rate at building the three main target objects (windmill, L shape, railway signal): on average, a subject built about two of the three objects correctly.

The overall correct-assembly rate was correlated with the overall rate of remembering objects: $R^2 = 0.20, p < 0.005$. However, subjects who said that they did remember how to build a snowman or an L shape the second time around were no more likely to do it correctly than those who said that they did not remember.

### 4.2 Subjective Measures

Two types of subjective measures were gathered during this study: responses on the user-satisfaction questionnaire, and self-assessment of emotions. Table 4 shows the mean results for each category from the user-satisfaction questionnaire across all of the subjects, in all cases on a 5-point Likert scale. The subjects in this study gave a generally positive assessment of their interactions with the system—with a mean overall satisfaction score of 3.75—and rated their perceived task success particularly highly, with a mean score of 4.1.

To analyse the emotional data, we averaged all of the subjects’ emotional self-ratings before and after the experiment, counting negative emotions on an inverse scale, and then computed the difference between the two means. Table 5 shows the results from this analysis; note that this value was assessed on a 1–4 scale. While the mean emotional
Table 4: User-satisfaction questionnaire results

| Question category   | Mean (Stdev) |
|---------------------|--------------|
| Robot as partner    | 3.63 (0.65)  |
| Instruction quality | 3.69 (0.71)  |
| Task success        | 4.10 (0.68)  |
| Feelings            | 3.66 (0.61)  |
| Overall             | 3.75 (0.57)  |

Table 5: Mean emotional assessments

| Mean (Stdev) | Min | Max |
|--------------|-----|-----|
| Before the study | 2.99 (0.32) | 2.32 | 3.68 |
| After the study  | 3.05 (0.32) | 2.32 | 3.73 |
| Change         | +0.06 (0.24) | -0.55 | +0.45 |

The factors included in Table 6 were the most significant contributors to user satisfaction. If a predictor does not contribute significantly, its $w_i$ value is zero after the stepwise process.

Using stepwise linear regression, we computed a predictor function for each of the subjective measures that we gathered during our study: the mean score for each of the individual user-satisfaction categories (Table 4), the mean score across the whole questionnaire (the last line of Table 4), as well as the difference between the users’ emotional states before and after the study (the last line of Table 5). We included all of the objective measures from Section 4.1 as initial predictors.

The resulting predictor functions are shown in Table 6. The following abbreviations are used for the factors that occur in the table: $Rep$ for the number of repetition requests, $Turns$ for the number of system turns, $Len$ for the length of the dialogue, and $Mem$ for the subjects’ memory for the components that were built twice. The $R^2$ column indicates the percentage of the variance that is explained by the performance function, while the $Significance$ column gives significance values for each term in the function.

Although the $R^2$ values for the predictor functions in Table 6 are generally quite low, indicating that the functions do not explain most of the variance in the data, the factors that remain after stepwise regression still provide an indication as to which of the objective measures had an effect on users’ opinions of the system. In general, users who had longer interactions with the system (in terms of system turns) and who said that they remembered the robot’s instructions tended to give the system higher scores, while users who asked for more instructions to be repeated tended to give it lower scores; for the robot-as-partner questions, the length of the dialogue in seconds also made a slight negative contribution. None of the other objective factors contributed significantly to any of the predictor functions.

6 Discussion

That the factors included in Table 6 were the most significant contributors to user satisfaction is not surprising. If a user asks for instructions to be re-
peated, this is a clear indication of a problem in
the dialogue; similarly, users who remembered
the system’s instructions were equally clearly having
a relatively successful interaction.

In the current study, increased dialogue length
had a positive contribution to user satisfaction; this
contrasts with results such as those of Litman and
Pan (2002), who found that increased dialogue
length was associated with decreased user satis-
faction. We propose two possible explanations for
this difference. First, the system analysed by Lit-
man and Pan (2002) was an information-seeking
dialogue system, in which efficient access to the
information is an important criterion. The current
system, on the other hand, has the goal of joint task
execution, and pure efficiency is a less compelling
measure of dialogue quality in this setting. Sec-
ond, it is possible that the sheer novelty factor of
interacting with a fully-embodied humanoid robot
affected people’s subjective responses to the sys-
tem, so that subjects who had longer interactions
also enjoyed the experience more. Support for this
explanation is provided by the fact that dialogue
length was only a significant factor in the more
‘subjective’ parts of the questionnaire, but did not
have a significant impact on the users’ judgements
about instruction quality or task success. Other
studies of human-robot dialogue systems have also
had similar results: for example, the subjects in the
study described by Sidner et al. (2005) who used
a robot that moved while talking reported higher
levels of engagement in the interaction, and also
tended to have longer conversations with the robot.

While the predictor functions give useful in-
sights into the relative contribution of the objective
measures to the subjective user satisfaction, the

| Measure                  | Function                                                                 | $R^2$ | Significance               |
|-------------------------|--------------------------------------------------------------------------|-------|---------------------------|
| Robot as partner        | $3.60 + 0.53 \cdot N(\text{Turns}) - 0.39 \cdot N(\text{Rep}) - 0.18 \cdot N(\text{Len})$ | 0.12  | Turns: $p < 0.01$, Rep: $p < 0.05$, Length: $p \approx 0.17$ |
| Instruction quality     | $3.66 - 0.22 \cdot N(\text{Rep})$                                        | 0.081 | Rep: $p < 0.05$           |
| Task success            | $4.07 + 0.20 \cdot N(\text{Mem})$                                        | 0.058 | Mem: $p \approx 0.07$     |
| Feelings                | $3.63 + 0.34 \cdot N(\text{Turns}) - 0.32 \cdot N(\text{Rep})$           | 0.044 | Turns: $p \approx 0.06$, Rep: $p \approx 0.08$         |
| Overall                 | $3.73 - 0.36 \cdot N(\text{Rep}) + 0.31 \cdot N(\text{Turns})$           | 0.062 | Rep: $p < 0.05$, Turns: $p \approx 0.06$                |
| Emotion change          | $0.07 + 0.14 \cdot N(\text{Turns}) + 0.11 \cdot N(\text{Mem}) - 0.090 \cdot N(\text{Rep})$ | 0.20  | Turns: $p < 0.05$, Mem: $p < 0.01$, Rep: $p \approx 0.17$ |

Table 6: Predictor functions
useful: more global measures such as how often the users look at the robot arms or at the objects on the table, as well as more targeted measures examining factors such as the user’s gaze and other behaviour during and after different types of system outputs. In future studies, we will also gather data on these additional non-verbal behaviours, and we expect to find higher correlations with subjective judgements.

7 Conclusions and Future Work

We have presented the JAST human-robot dialogue system and described a user study in which the system instructed users to build a series of target objects out of wooden construction toys. This study resulted in a range of objective and subjective measures, which were used to derive performance functions in the style of the PARADISE evaluation framework. Three main factors were found to affect the users’ subjective ratings: longer dialogues and higher recall performance were associated with increased user satisfaction, while dialogues with more repetition requests tended to be associated with lower satisfaction scores. The explained variance of the performance functions was generally low, suggesting that factors other than those measured in this study contributed to the user satisfaction scores; we have suggested several such factors.

The finding that longer dialogues were associated with higher user satisfaction disagrees with the results of many previous PARADISE-style evaluation studies. However, it does confirm and extend the results of previous studies specifically addressing interactions between users and embodied agents: as in the previous studies, the users in this case seem to view the agent as a social entity with whom they enjoy having a conversation.

A newer version of the JAST system is currently under development and will shortly undergo a user evaluation. This new system will support an extended set of interactions where both agents know the target assembly plan, and will also incorporate enhanced components for vision, object recognition, and goal inference. When evaluating this new system, we will include similar measures to those described here to enable the evaluations of the two systems to be compared. We will also gather additional objective measures in order to measure their influence on the subjective results. These additional measures will include those mentioned at the end of the preceding section, as well as measures targeted at the revised scenario and the updated system capabilities—for example, an additional dialogue quality measure will assess how often the goal-inference system was able to detect and correctly respond to an error by the user.

Acknowledgements

This research was supported by the European Commission through the JAST² (IST-FP6-003747-IP) and INDIGO³ (IST-FP6-045388) projects. Thanks to Pawel Dacka for his help in running the experiment and analysing the data.

References

M. Argyle and J. A. Graham. 1976. The Central Europe experiment: Looking at persons and looking at objects. Environmental Psychology and Nonverbal Behavior, 1(1):6–16. doi:10.1007/BF01115461.

A. J. N. van Breemen. 2005. iCat: Experimenting with animabotics. In Proceedings of the AISB 2005 Creative Robotics Symposium.

C. Callison-Burch, M. Osborne, and P. Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In Proceedings of EACL 2006. ACL Anthology E06-1032.

M. E. Foster. 2008. Automated metrics that agree with human judgements on generated output for an embodied conversational agent. In Proceedings of INLG 2008. ACL Anthology W08-1113.

M. E. Foster, M. Giuliani, A. Isard, C. Matheson, J. Oberlander, and A. Knoll. 2009. Evaluating description and reference strategies in a cooperative human-robot dialogue system. In Proceedings of IJCAI 2009.

M. Huber, M. Rickert, A. Knoll, T. Brandt, and S. Glasauer. 2008. Human-robot interaction in handing-over tasks. In Proceedings of IEEE RO-MAN 2008. doi:10.1109/ROMAN.2008.4600651.

C. Y. Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In Proceedings of the ACL 2004 Workshop on Text Summarization. ACL Anthology W04-1013.

²http://www.euprojects-jast.net/
³http://www.ics.forth.gr/indigo/
D. J. Litman and S. Pan. 2002. Designing and evaluating an adaptive spoken dialogue system. *User Modeling and User-Adapted Interaction*, 12(2–3):111–137. doi:10.1023/A:1015036910358.

S. Möller, K.-P. Engelbrecht, and R. Schleicher. 2008. Predicting the quality and usability of spoken dialogue systems. *Speech Communication*, 50:730–744. doi:10.1016/j.specom.2008.03.001.

D. G. Novick. 1997. What is effectiveness? In *Working notes, CHI ’97 Workshop on HCI Research and Practice Agenda Based on Human Needs and Social Responsibility*. http://www.cs.utep.edu/novick/papers/eff.chi.html.

A. Ortony, G. L. Clore, and A. Collins. 1988. *The Cognitive Structure of Emotions*. Cambridge University Press.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*. ACL Anthology P02-1040.

M. Rickert, M. E. Foster, M. Giuliani, T. By, G. Panin, and A. Knoll. 2007. Integrating language, vision and action for human robot dialog systems. In *Proceedings of HCI International 2007*. doi:10.1007/978-3-540-73281-5_108.

C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich. 2005. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1–2):140–164. doi:10.1016/j.artint.2005.03.005.

M. Walker, C. Kamm, and D. Litman. 2000. Towards developing general models of usability with PARADISE. *Natural Language Engineering*, 6(3–4):363–377.

M. A. Walker. 2000. An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. *Journal of Artificial Intelligence Research*, 12:387–416.

M. A. Walker, J. Fromer, G. D. Fabbrizio, C. Mes tel, and D. Hindle. 1998. What can I say?: Evaluating a spoken language interface to email. In *Proceedings of CHI 1998*. doi:10.1145/274644.274722.

M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *Proceedings of ACL/EACL 1997*. ACL Anthology P97-1035.

M. White, M. E. Foster, J. Oberlander, and A. Brown. 2005. Using facial feedback to enhance turn-taking in a multimodal dialogue system. In *Proceedings of HCI International 2005*.