Talking with Robots about Objects: 
A System-Level Evaluation in HRI

Julia Peltason, Nina Riether, Britta Wrede and Ingo Lütkebohle
Applied Informatics Group
Bielefeld University, Germany
{jpeltaso|nriether|bwrede|iluetkeb}@techfak.uni-bielefeld.de

ABSTRACT
We present the design process, realization and evaluation of a robot system for interactive object learning. The system-oriented evaluation, in particular, addresses an open problem for the evaluation of systems, where overall user satisfaction depends not only on the performance of the parts, but also on their combination, and on user behavior. Based on the PARADISE method known from spoken dialog systems, we have defined and applied internal and external metrics for fine-grained and largely automatable identification of such relationships. Through evaluation with n=28 subjects, indicator functions explaining up to 55% of variation in several satisfaction metrics were found. Furthermore, we demonstrate that the system's interaction style reduces the need for instruction and successfully recovers partial failures.

Categories and Subject Descriptors
C.4 [Performance of Systems]: Measurement techniques; I.2.7 [Artificial Intelligence]: Robotics; I.2.9 [Artificial Intelligence]: Natural Language Processing

General Terms
Measurement, Performance, Human Factors

Keywords
Human-robot interaction, object learning, social robotics

1. INTRODUCTION
Interaction provides many benefits for (semi-)autonomous robotic systems, e.g., to set goals, to efficiently enlist human help, and to explain the system's capabilities and limitations, for better human-robot-cooperation [2, 4]. It is well accepted that a combination of modalities and capabilities improves such interactions. However, understanding the effects of dynamic, interactive combinations of capabilities remains very challenging.

One way to evaluate this is to systematically vary each capability, but while this is certainly a powerful approach, it is infeasible to conduct it for large numbers of capabilities. Furthermore, it is difficult in such an approach to determine the relative importance of each capability.

Therefore, in this contribution, we will, firstly, present the design and realization of a full robotic system, and, secondly, its system-level evaluation. The system realizes an interactive, social object learning scenario, where object labels are learned during spoken language interaction.

As a concrete example of the benefit of such evaluations, consider the following: For object learning, object recognition performance depends both on the algorithm and on training image quality. The latter, then, depends on both the moment when the training image is taken, as well as demonstration behavior (i.e., hands in front of the object or similar), both of which are influenced by the dialog strategy. So, which of these is most important in the end?

Similar explorative questions arise in many larger systems, and we believe that this constitutes one of the major issues for robotics research today. One possible way to address it is through regression analysis, which relates user's subjective judgements with metrics that describe system behavior. Specifically, we have chosen a method widely used for speech-based dialog systems called PARADISE [15].

An important prerequisite to use PARADISE, however, is the definition, and efficient capture, of insightful metrics. In particular, automatic capture of metrics would greatly increase practicability. Therefore, we propose to analyze system message logs in combination with two general action models, one for interaction episodes (Interaction Patterns [12]) and one for system activities (the Task-State-Pattern [10]).

Related Work. Despite structural similarities to spoken dialog systems, the PARADISE approach is currently little used for HRI. One of the few studies is by Foster et al. [5], who explored the predictive power of interaction effort measures. These are found to be significant, but only explain a relatively small amount of variation (between 4 and 12%). In comparison, typical dialog system evaluations find values between 30 and 70% [5], and we found between 17 and 55%.

A later study by Dethlefs et al. [3] compares a binary task success metric with a scoring scheme that takes difficulties into account. They find that the binary metric is not significantly related to user satisfaction, while the more detailed scores are. In combination with an interaction-turn measure, they account for 62% of variation. This suggests that the process is of crucial importance for user satisfaction.

Both of these studies take an external view, with relatively
few (three to five) metrics that are manually annotated. In contrast, our approach looks \textit{inside} the system, by directly analyzing message logs. This inside view allows us to determine not only how the system performs, but also why.

2. \textbf{SYSTEM DESCRIPTION}

In our interactive learning scenario, a number of previously unknown objects are present on a table (see figure 1(b)) and participants can show them to the robot, label them, and verify what has been learned. Moreover, the robot can also take initiative and ask for object labels itself, in a mixed-initiative setting. Most parts of the system are autonomous, which includes real object and speech recognition to create realistic challenges. Two operator-assisted (Wizard-of-Oz) components are used for reference resolution and region selection during robot initiative (cf. section 2.3). In the remainder of this section, we present the interaction strategy design, architecture and capabilities, as well as the dialog principles.

(a) The Flobi head [7] (b) Demonstrating a fruit

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{scenario_overview.png}
\caption{Scenario overview}
\end{figure}

\subsection{2.1 Iterative Interaction Design}

For the design of dialog systems, it is important to know which capabilities users expect in a given situation. While the variability of human behavior means that not everything will (or has to) be supported, missing important aspects in this stage could lead to errors that overshadow other evaluation results. We handled this using pre-studies which we will describe in the following. In addition to these, we also conducted an earlier study on the robot’s initiative, to validate the mixed-initiative setting [8].

Initially, we focused on speech recognition, as it has often been reported to be a major contributor to user satisfaction (e.g., [14]). In an explorative user study with 10 participants, from a total of 392 user utterances 27\% were not understood correctly. Among these, out-of-capability utterances (i.e., the utterance is beyond the system capabilities) represented the by far largest portion (54\%, i.e., 15\% of all utterances). More than half of these were caused by the user attempting to reverse the roles after the initial steps, proceeding to demonstrate objects on their own, which was not possible in that system. As a consequence, this capability has now been added.

Based on re-examination of an earlier Wizard-of-Oz (WOz) study on object teaching by our colleagues, Lang et al. [6] (targeting facial expressions), with 11 participants, we determined the dialog strategies used. All interactions had a very similar structure, consisting of an opening part, a task-related part and a closing part. 36\% of the interactions additionally feature transitional phrases that introduce the task-related part. In the opening phase, introducing each other (82\%) and exchanging pleasantries (18\%) are frequent.

Aside from object demonstrations, the task-related phase consists of checking learned objects (45\%) and transitional phrases between the objects (36\%). Praising the robot for correctly learned objects turned out to be universal (100\%). The task-related part may include closing remarks (36\%).

Based on this information, we designed the interaction strategy of the present scenario. In general, most of the observed strategies could be realized. Some particularly rich descriptions went beyond the current system capabilities and were ignored, however. For example, the object descriptions sometimes included functional description of objects that are not yet supported. Table 1 lists the resulting interaction capabilities of the system.

Beside the out-of-capability utterances, out-of-vocabulary utterances (i.e., the meaning is within capabilities, but the particular wording is not in the configured set) constitute a significant portion of non-understandings (12\%). Thus, enhancing the speech recognition grammar became another focus of the current iteration. Here, a balance had to be found between broad grammar coverage and acceptable recognition performance. We opted for an inclusive grammar, so some errors will inevitably remain. In order to cope with these, robust key-word matching techniques were applied for interpretation of recognition results.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|}
\hline
\textbf{Purpose} & \textbf{Example interaction} \\
\hline
Greeting & User: Hello, Flobi \\
& Flobi: Hi, hello. \\
Introducing & User: My name is Anneliese. \\
& Flobi: Hello, Anneliese. Nice to meet you. \\
Exchanging pleasantries & User: How are you? \\
& Flobi: Fine, thank you! \\
Task transition & User: I am here to show you some objects. \\
& Flobi: Alright, let’s start! \\
Attracting attention & User: Flobi, look here! \\
& Flobi: Yes, I’m looking. \\
Object demonstration & User: This is a yellow lemon. \\
& Flobi: Alright, a yellow melon. Is that correct? \\
& User: No, it’s a lemon. \\
& Flobi: Alright, a lemon. Is that correct? \\
& User: Yes, that’s correct. \\
& Flobi: Okay, I am going to learn the lemon. \\
& Flobi: I have learned the lemon. \\
Object query & Flobi: What is that? \\
& User: This is an apple. \\
& (remainder similar to object demonstration) \\
Listing learned objects & User: What have you learned so far? \\
& Flobi: I have learned a lemon and an apple. \\
Checking & User: Flobi, do you know this one? \\
& Flobi: I will have a look at it. \\
& Flobi: This is an apple. \\
Praising & User: Well done, Flobi. \\
& Flobi: Thank you. \\
Restart & User: Restart. \\
& Flobi: Okay, let’s start over again. \\
Transitional phrases & User: Would you like to learn another object? \\
& Flobi: Yes, I want to learn them all. \\
Closing task & User: Okay, that’s it. \\
& Flobi: Are you leaving already? \\
Parting & User: Good bye, Flobi. Take care! \\
& Flobi: Bye, see you later. \\
\hline
\end{tabular}
\caption{Interaction capabilities of the system. The blocks are: i) Opening, ii) task-related, iii) closing.}
\end{table}

\subsection{2.2 Architectural Concepts}

Our system uses a component-based, layered architecture with event-based, asynchronous communication, whose main ideas have previously been reported in [8, 9]. The components of the system are shown in figure 2.
module with current dialog activity. Of ongoing tasks is also used to coordinate the gaze feedback (i.e., when the task begins and ends). This detailed reporting gives the user a consistent fashion (e.g., when object recognition or learning begins). There are also transitions for updating the goal or aborting tasks. Through the corresponding states, we can specify the goal for the task-success metrics (cf. section 4.3).

A central concept, with particular relevance for this paper, is the Task-State-Pattern model for representing system activities [10]. It represents all actions of the system as a pair of i) state and ii) goal specification. The set of possible states is the same for all tasks of the system. The goal specification, on the other hand, is specific to the task.

Communication in the system occurs through event notifications on every transition. Example transitions include initiated for starting it, completed for success and failed for failure. There are also transitions for updating the goal or aborting tasks. Through the corresponding states, we can track all tasks in a generic manner, which forms the basis for the task-success metrics (cf. section 4.3).

Moreover, the dialog manager uses this general model to trigger activities and inform the user about their progress in a consistent fashion (e.g., when object recognition or learning begins and ends). This detailed reporting gives the user clear information on the system’s state. Lastly, observation of ongoing tasks is also used to coordinate the gaze feedback module with current dialog activity.

2.3 System Capabilities

The hardware used includes the Flobi anthropomorphic robot head (cf. figure 1(a)), a close-talking microphone, and a flat-screen display, used to show the selected object in robot-initiative (but not with user-initiative).

Apart from basic social talking skills (greeting, complimenting, etc.), the system has three main capabilities: i) Having the user teach an object, ii) asking for an object on its own initiative, and iii) answering requests about object labels and known objects. All of these can be triggered and answered through a variety of verbal utterances (cf. table 1). Furthermore, the system also orients its gaze at the current most (visually) salient point or the user, as appropriate for the task.

Compared to the previous scenario [8], an important difference is the operator-assisted reference resolution and region of interest selection. Despite our general goal of using only autonomous behavior, we found through the WoZ study on object teaching that referencing behavior varies considerably, and this observation was confirmed by the present user study, where 8 different referencing strategies could be identified, as listed in table 2. Not all of these could be acceptably automated for now. Hence, the system detects references on its own, but for the moment resolving is done by an operator selecting the appropriate object region. The robot will then gaze at the selected region. Requests may also be rejected, e.g., when no object is visible in the robot’s cameras.

For the region of interest selection, we previously used a saliency model [8], which was found to be not predictable enough for repeatable study results, however. Therefore, in the present system, a human operator chose one of the objects visible when the system had initiative.

The remainder of the system is autonomous, however, and for all of the functions, vision and speech interact in a tightly integrated manner, with each step modeled as a task, and integration based on event notifications.

2.4 Interaction Patterns

The mixed-initiative interaction strategy was realized with the PaMini (for Pattern-Based Mixed-Initiative HRI) dialog management approach that we have suggested recently [12]. This approach relies on generic Interaction Patterns that model recurring conversational structures, such as making a suggestion or negotiating information. Interaction Patterns are defined at an abstract level, but can be tailored with an application-specific configuration. They can be initiated either by the human (e.g., Human Object Test) or by the robot (e.g., Robot Information Request).

Within a pattern, the above system tasks may be initiated, and in turn the patterns react to task state changes. As an example, figure 3 depicts the Human Object Test pattern, which was used to model the user’s object recognition requests (“Flobi, what is that?”). Within this pattern, two sequential tasks are initiated (first a reference resolution task, then a lookup task), and the pattern combines system task states (such as task failed) with appropriate robot dialog acts (such as R.apologize), thus relating the domain level with the conversation level. Similarly, the pattern used for the user’s object demonstration (“This is an apple”) initiates a resolve reference task first, then a learn object task.

For each interaction purpose in table 1, a corresponding Interaction Pattern was configured, plus several additional patterns as variations. Overall, the dialog consisted of 20 Interaction Patterns, each of which is capable to respond to numerous different wordings for a specific dialog act. At run-time, multiple patterns can be active at the same time, and can be interleaved to achieve a more flexible interaction style. The information about the interaction patterns used during an interaction, and the sequence of states they took, forms the basis for the dialog efficiency and quality metrics (see section 3).

Table 2: Ratios for referencing strategies (rounded).

<table>
<thead>
<tr>
<th>Referencing strategy</th>
<th>% of total</th>
<th>% of user references</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robot</td>
<td>25%</td>
<td>n/a</td>
</tr>
<tr>
<td>Lifting</td>
<td>19%</td>
<td>30%</td>
</tr>
<tr>
<td>Isolated object</td>
<td>14%</td>
<td>22%</td>
</tr>
<tr>
<td>Pointing</td>
<td>13%</td>
<td>22%</td>
</tr>
<tr>
<td>False positive</td>
<td>11%</td>
<td>n/a</td>
</tr>
<tr>
<td>Touching</td>
<td>10%</td>
<td>16%</td>
</tr>
<tr>
<td>Non-visual</td>
<td>5%</td>
<td>7%</td>
</tr>
<tr>
<td>Isolated-object + pointing</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>Isolated-object + touching</td>
<td>1%</td>
<td>1%</td>
</tr>
</tbody>
</table>
3. PARADISE EVALUATION FOR HRI

The PARADISE evaluation approach has originally been proposed for spoken dialog systems [15], where it has become a de-facto standard. Its goal is to explore the relative importance for, and contribution of, the functions of a system for user satisfaction. The system is characterized using objective metrics, and user satisfaction is determined through subjective judgments. The relationship between these is then established using multiple linear regression (cf. section 4.5). A critical aspect to achieve a meaningful PARADISE analysis is the choice of metrics. Ideally metrics should be standardized, but this is not yet the case. We hope to make one step in this direction with the description of our procedure.

Choosing Subjective Measures: Subjective measures are usually determined through a questionnaire and here, a number of good proposals for HRI and dialog are available. We recommend using a combination of these to capture various aspects. For general impression of the system, we have used Bartneck et al.’s “godspeed” proposal [1], aggregated by category. For interaction-related metrics, we have adapted a questionnaire for evaluation of the COMIC system [16]. See section 4.4 for details on the metrics used.

Choosing Objective Measures: The objective measures need to capture what actually happened, i.e., what the system and the user did. They are assigned to three categories: task success, dialog quality and dialog efficiency.

Measures are based on system logs, but need to reflect system performance, so we cannot simply count messages. Instead, we have based our metrics on two generalizations: Firstly, user and system utterances are assigned to Interaction Patterns (cf. 2.4), which describe a single interaction episode. From this, we can compute the length and the number of steps for each episode. Secondly, messages about system activities are assigned to the task they pertain to, using the Task-State-Pattern (cf. 2.2), which describes the outcome (success/failure) of all tasks, grouped by type. Thus, we exploit the Interaction Patterns to gain information about interactional aspects of system performance (i.e., about dialog quality and efficiency), whereas the Task State Pattern provides information at task level (i.e., about the task success).

Interaction Patterns and Tasks are a form of system model which we also use at runtime. However, this is not a prerequisite for defining metrics, it just simplifies computation.

4. USER STUDY

The robot system was evaluated via a user study in which participants interacted with the system. It focuses on the relationships between the various objective, system-based measures contributing to dialog efficiency, dialog quality and task success (see 4.3), and the subjective variables (see 4.4). Thus, our evaluation was aimed to detect general connections among system based measures and subjective reactions to the system (see 5.3).

4.1 Participants

From 32 participants who took part, we used 28 recordings (14 male, 14 female), with the remaining four excluded due to technical problems. Most of the participants had been recruited at a university event for the general public and thus represented a wide age range, with mean age at 33.5 years, minimum 21 and maximum 79. On a scale from 1 (none) to 6 (lots of), the average rating for knowledge of computers was at 5.07, of speech systems at 2.52, of robot systems at 1.96 and of programming experience at 2.26. They were compensated for their participation in the experiment.

4.2 Instruction

In order to study natural demonstration behavior, participants received as little instruction as possible. They received written instructions, specifying that they were to engage in interaction with the robot Flobi, and that Flobi was supposed to learn object labels during interaction. They were also advised to check that the robot had actually learned the labels. It was, however, not specified how they should present and check objects. They were told that they should interact with the robot as long as they wish, with 5-10 minutes recommended as a guideline. Also, they were informed that they could begin the interaction by greeting the robot, and end the interaction by saying goodbye. In addition, participants were advised not to be discouraged by speech recognition problems, and that they could repeat or rephrase their utterance in such cases. Lastly, an emergency phrase (“Restart”) was provided. The interactions were in German.

4.3 Objective Measures

A wide range of objective measures has been collected, most of which were derived from system message logs. For each component, the relevant messages were logged, such as speech recognition results, text-to-speech output, Interaction Pattern state changes as well as object recognition and reference resolution tasks. With these log data, a detailed reconstruction of the interaction can be achieved. A few measures have been annotated manually based on the video material, such as inappropriate robot utterances, or the correctness of the robot’s answer on a test question. In total, we used 22 of these for evaluation (see table 3).

The dialog efficiency measures capture the rapidity of the interaction and include for example the duration of interaction, the number of user and robot utterances within a certain time unit, the mean length of user utterances, or the number of objects learning episodes within a certain time

Figure 3: The Human Object Test pattern, which is used to model the user’s object recognition requests.
unit. The dialog quality measures address the smoothness of the interaction. We considered for example gaps, overlaps, repairs and label corrections. The task success measures concentrate on the outcome of the interaction with respect to object learning. Among others, we measured the proportion of successful reference resolution and object learning tasks, the proportion of correct robot answers on test questions, and the user’s out-of-capability utterances.

4.4 Subjective Measures

In addition to the objective measures described above, we collected subjective measures based on a questionnaire the participants were asked to complete after the interaction with Flobi. The questionnaire consisted of 50 items, that we aggregated into seven category measures. The first four categories, dialog efficiency, robustness, cooperativeness and usability, refer to the interaction itself. These interaction-oriented items are roughly based on the evaluation of the COMIC dialog system [16], which we adapted for our specific scenario. The remaining three categories, likeability, perceived intelligence and animacy address the participants’ impression of the robot. They were adopted from the standardized Godspeed questionnaire¹ [1]. In addition, the questionnaire included five single (summarizing) Likert-scale questions, targeting the overall impression of ease, efficiency, clarity, pleasantness and understandability of the interaction.

4.5 Computing Performance Functions

As PARADISE is not yet widely used in HRI, we will shortly summarize the procedure for performing it here, and place it in the appropriate context. For more background, see [15].

The model uses stepwise multiple linear regression to make predictions about subjective measures, like user satisfaction, based on several objective performance dimensions, like task success, dialogue quality, and dialogue efficiency. Multiple linear regression analysis is widely used to assess the degree to which an outcome or dependent variable is related to a set of predictor or independent variables. It specifies a model in which a dependent variable Y represents a linear combination of several parameters, including multiple independent variables, of the form

\[ Y_i = (b_0 + b_1x_{1i} + b_2x_{2i} + \ldots + b_nx_{ni}) + \epsilon_i \]

The \( b_0 \) term represents a constant, whereas \( b_1 \) is the coefficient, or weight, of predictor \( x_1 \) and \( \epsilon_i \) stands for the error term, or residual, of the equation.

In a stepwise approach, the best-fit predictors from a number of possible predictors are selected, by excluding predictors that do not contribute significantly to the outcome variable (Y). Taken together, the linear combination of the selected predictor variables and their computed coefficients can then be applied to make predictions of the dependent variable [13]. This technique also allows assessment of the relative contribution of each of the predictors towards the outcome variable (the \( \beta \)-value).

The multiple linear regression model can be referred to as general linear model with only one outcome or dependent variable. Along with path analysis, factor analysis and canonical correlation, it is a special case or subform of the more complex structural equation modeling [13].

5. RESULTS

In this section we present results of the objective and subjective measures, and on the relationship between them, described by several performance functions.

5.1 Objective Measures

Table 3 shows a sample of the results of the objective measures. On average, interactions lasted 11.07 minutes, during which 9.43 objects were taught successfully (0.91 per minute). In the mean, there were 3.55 user utterances and 10.91 robot utterances per minute. The mean value of gaps between utterances was 1.5 seconds, whereas utterances overlapped during 20% of the speaking time. 14% of the robot’s utterances were repair utterances, and 6% were inappropriate, i.e. the robot misunderstood a “valid” user utterance. When participants demonstrated objects to the robot, reference resolution succeeded in 79% of all cases. From these, 78% could be learned, i.e. 62% of the overall demonstrated objects. The robot could answer 53% of the test questions, 55% of them correctly, i.e. 29% of the overall test questions. It must be admitted that while the participants were generally very accepting, this is not the final task performance we had expected from a state-of-the-art recognition algorithm as used here. While it is obviously the combination of all errors, it is still puzzling. We surmise that very variable lighting conditions may play a role, as well as a sub-optimal choice of training images at a time when the user’s hand is still visible. We will have to investigate this further.

That said, we were pleased to see that the out-of-capability utterances could be reduced considerably in comparison with the previous iteration of the scenario (from 15% to 6%). This means that with our preparatory analysis of the WoZ study and the resulting extensions of the system, the system capabilities were successfully adapted to the users’ expectations. Table 4 breaks down the remaining out-of-capability utterances into subcategories. Attempts to correct the robot’s answer on object check questions were a common error cause (which points out a deficiency in the interaction strategy), followed by meta comments on object organization (“let me put this one here”). Trying to demonstrate more than one object at a time was also common. Also, a significant portion of out-of-capability utterances referred to the robot’s gaze direction (“where are you looking?”). Other categories include canceling an on-going object learning action, teaching object categories (“these are fruits”), asking the robot to repeat its utterances, asking it to proceed to the next label, or teaching colors.

Finally, we’d like to note the remarkably high standard deviation, indicating significant differences between the single interactions.

5.2 Subjective Measures

Table 5 shows the results for the subjective measures. All aspects were generally rated positively, with ratings of more than 4 (on a scale from 1 to 6) for ease of use, clarity, pleas-

¹ However, we skipped the categories anthropomorphism and perceived safety, as we considered them irrelevant for the scenario at hand.
cross-validated using data splitting. The values of $R^2$ are rounded to full integers.

Table 3: Objective measures. Blocks are: i) dialog efficiency, ii) dialog quality, and iii) task success. Percentages are rounded to full integers.

Table 4: Ratios for out-of-capability (rounded).

Table 5: Results of subjective measures. Blocks are: i) Single-item measures, ii) aggregated measures.

5.3 Performance Functions

As outlined before (cf. section 4.5), we computed predictor functions for the various subjective measures. The objective measures from section 4.3 were included as potential predictors.

5.3.1 Presentation

Table 6 presents the resulting performance functions for our subjective outcome variables. The obtained results were cross-validated using data splitting. The values of $R^2$ describe the amount of variance of the outcome variable that is accounted for by the model. These scores range from .171 (perceived intelligence) to .553 (usability) for our data, meaning for instance that 55.3% of the variance of the subjective measure “usability” is explained by the respective predictor function.

The factors in the predictor functions show which objective measures affected the subjective ratings, and their weights specify how much. However, the absolute numbers for the weights are not easily interpretable, because they depend on the range and unit of the factor in question. Therefore, the next-column (“$\beta$ values”) gives normalized weights, which represent the standardized b-coefficient for each predictor, which is independent from the range and unit of measurement of that specific objective predictor. Hence, they are directly comparable to each other, giving a better insight into the “true” weight or importance of the respective predictor. Positive coefficients indicate a positive relationship between predictor and outcome, negative coefficients stand for a negative relationship.

The significance column states whether each predictor of the particular function made a significant contribution to the subjective measure. Note that for two subjective measures no predictor function could be obtained, meaning that no single predictor or combination of multiple predictors could make a significant contribution to the outcome variable.

5.3.2 Interpretation

Most of the performance functions resulting from the PARADISE evaluation (shown in table 6) appear plausible, but some of them point out unexpected relationships between measures that help to identify deficiencies of the system. The performance functions generally exhibit high $R^2$ values, indicating that they do explain much of the variance in the data.

For instance, the number of objects learned per minute is a strong predictor for the ease of use, together with the overall number of demonstrated objects and the duration of interaction, suggesting that users who found interaction easy and were successful in teaching objects tended to have longer interactions with the system.

In contrast, usability as an aggregated measure from
questions on control, predictability, concentration, clarity of when to speak is explained by a different function. Here, not only the number of objects learned per minute has an impact, but also how many interaction steps were required for learning, and the number of completed interaction patterns (which can be interpreted as a measure for the general interaction success, taking into account not only the task-related but also the social aspects). Somewhat surprisingly, the user utterances per minute had a negative impact on usability, and also on the perceived intelligence. Qualitative analysis revealed that for users who tend to talk rather fast and keep talking even during the robot’s utterances, there is a risk of cumulating delays in the robot responses, as the user utterances are simply queued and processed one after another. This indicates the need for a more flexible turn-taking behavior that enables the robot to suspend or abort its own utterances, or to ignore user utterances if there is a newer one present.

Asking for efficiency directly did not yield a significant model. However, estimating efficiency as an aggregated measure (from questions on time required for object learning, clarity of referred objects, and general functionality of the system) suggests as predictors the number of repair utterances, interaction steps required for object demonstrations and object requests, as well as the out-of-capability utterances. The interaction steps for object demonstrations contribute positively, which may seem surprising at first sight. Looking at the interaction pattern used for object demonstrations, we realized that a high number of steps indicates multiple object corrections, while a low number of steps indicates that already the reference resolution fails. Thus, an explanation for the positive contribution of interaction steps might be that failures of reference resolution have a stronger negative effect on efficiency than correcting a misunderstood label, and are even more frustrating.

The factors that influence clarity of interaction are the number of repair utterances and the number of objects learned per minute, with the former being a major measure for interaction success and the latter for task success. The pleasantness is predicted (not very strongly, though) only by the overall number of objects learned, but our measures possibly miss some factors that are relevant for pleasantness.

Interestingly, we note a significant impact of reference resolution failures on the understandability score. We attribute this to the fact that for such failures, the robot only reports that it could not determine what the user referred to, not why. This provides very little information towards good error recovery, leaving users guessing. Qualitative analysis confirms this interpretation: several users explicitly asked the robot why it was not able to determine the referred object. This result points out again the frustrating effect of reference resolution failures, and the importance of more informative feedback in error conditions.

The robustness of the system as an aggregated measure (from questions on reliability and robustness of the system) is predicted by the proportion of repair utterances, but not by measures that specifically refer to object learning and recognition. This demonstrates that a robust interaction can alleviate deficiencies at task level.

Cooperativeness (aggregated from questions on the robot’s readiness to interact, interest, attentiveness and autonomy) is affected not only by objects learned per minute, but also by gaps between utterances. While this may seem less obvious at first sight, it might be another evidence for the crucial role of turn-taking: longer gaps between utterances indicate a smooth turn-taking, and thus better speech recognition performance. It is probably for the same reason that gaps contribute to perceived intelligence as well.

### Table 6: Performance functions. Blocks are: i) Single-item measures, ii) aggregated measures. See section 5.3.1 for an explanation of the terms.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Function</th>
<th>R²</th>
<th>β</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ease of use</td>
<td>-0.30 + 4.33(Obj/min) + 1.15(ObjDemo) + 0.12(Time)</td>
<td>.471</td>
<td>Obj</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ObjDmo</td>
<td>&lt;.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Time</td>
<td>&lt;.059</td>
</tr>
<tr>
<td>Efficiency</td>
<td>No significant model</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clarity</td>
<td>4.63 - 6.92(Repair) + 1.39(Obj/min)</td>
<td>.367</td>
<td>Repair</td>
<td>&lt;.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Obj</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>Pleasantsness</td>
<td>3.58 + 0.18(Learn)</td>
<td>.184</td>
<td>Learn</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>Understandability</td>
<td>3.78 + 0.14(Ref)</td>
<td>.196</td>
<td>Ref</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>Efficiency</td>
<td>1.42 - 5.09(Repair) + 0.61(Steps(ObjDemo) - 0.06(UU/min))</td>
<td>.441</td>
<td>Repair</td>
<td>&lt;.05</td>
</tr>
<tr>
<td></td>
<td>- 0.10(Steps(ObjReq))</td>
<td></td>
<td>StepObjDemo</td>
<td>&lt;.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>UU/min</td>
<td>&lt;.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Steps(Obj)</td>
<td>&lt;.081</td>
</tr>
<tr>
<td>Usability</td>
<td>4.03 - 0.22(Steps(ObjDemo) + 2.39(Obj/min) + 0.21(UU/min) + 0.03(NPattern)</td>
<td>.553</td>
<td>StepObjDemo</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Obj</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>UU/min</td>
<td>&lt;.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NPattern</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>Robustness</td>
<td>5.07 - 6.08(Repair)</td>
<td>.183</td>
<td>Repair</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>Likeability</td>
<td>No significant model</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cooperativeness</td>
<td>1.89 + 1.79(Obj/min ) + 0.001(Gaps)</td>
<td>.446</td>
<td>Objcor</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Perceived Intelligence</td>
<td>3.02 + 0.001(Gaps)</td>
<td>.171</td>
<td>Gaps</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>Animacy</td>
<td>3.67 + 1.09(Obj/min) - 0.118(UU/min)</td>
<td>.281</td>
<td>Obj</td>
<td>&lt;.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>UU/min</td>
<td>&lt;.074</td>
</tr>
</tbody>
</table>
the various system functions for the subjective judgements, we related them to objective measures of the interaction through a PARADISE-style regression analysis.

A novel approach for such an evaluation has been the automatic definition and computation of objective system metrics based on, firstly, Interaction Patterns, which describe self-contained interaction episodes, and secondly, Task-State-Machines, which describe all system activities. This automated approach has allowed us to compute over 20 metrics, to perform fine-grained analysis. The resulting performance functions can significantly explain up to 55% (avg 33%) of variation in user satisfaction, which demonstrates that the metrics are meaningful.

Apart from providing information about the relative importance of the system’s functionality, these functions also revealed unexpected relations between objective system measures and certain aspects of user satisfaction. For example, turn-taking was related to usability and perceived intelligence in unexpected ways. Moreover, it helped us to focus additional qualitative analysis on the most important of the many possible aspects. For example, it is a general tenet of error reporting to provide enough information for error recovery. However, this is not always achieved. In this case, the error reporting for resolution failures was apparently not helpful enough, which was clearly identified by the PARADISE analysis.

We also described our iterative strategy for designing the dialog strategy, and the user study confirmed intuitiveness of interaction, as indicated e.g., by a considerable decrease of out-of-capability utterances compared to previous iterations, despite little instruction. The PARADISE analysis also showed that error recovery in the dialog is generally seen as positive, if it leads to a successful completion. The remaining out-of-capability utterances, however, point out issues for improvement, e.g., the somewhat unclear gaze direction of the robot.

One aspect that we’d like to focus more on in future work are the large individual differences between users, both in task and in dialog success. There is no clear explanation in the data analyzed, but this certainly requires further investigation to assist users with (initial) difficulties. Similarly, while for some subjective metrics the performance functions explain about half the variation, which we consider a very good value, there are some metrics where only 20% of variation is explained, which is less helpful. A first step here could be to determine participants’ initial expectations and predispositions and include them in the analysis.

7. ACKNOWLEDGMENTS

We gratefully acknowledge partial support by the German Aerospace Center (DLR) with funds from the Federal Ministry of Economics and Technology (BMBF) due to resolution 50RA1023 of the German Bundestag. We’d also like to thank Annika Rothert for assistance with preparing and executing the user studies. Last, but not least, we’d like to thank the anonymous reviewers for their exceptionally helpful commentary.

8. REFERENCES


