The Curious Robot – Structuring Interactive Robot Learning

Ingo Lütkebohle, Julia Peltason,
Lars Schillingmann, Britta Wrede and Sven Wachsmuth
Applied Informatics Group
Bielefeld University, Germany
{iluetkeb,jpeltaso,lschilli,bwrede,swachsmu}@techfak.uni-bielefeld.de

Christof Elbrechter and Robert Haschke
Neuroinformatics Group
Bielefeld University, Germany
{celbrech,rhaschke}@techfak.uni-bielefeld.de

***Abstract*— If robots are to succeed in novel tasks, they must be able to learn from humans. To improve such human-robot interaction, a system is presented that provides dialog structure and engages the human in an exploratory teaching scenario. Thereby, we specifically target untrained users, who are supported by mixed-initiative interaction using verbal and non-verbal modalities. We present the principles of dialog structuring based on an object learning and manipulation scenario. System development is following an interactive evaluation approach and we will present both an extensible, event-based interaction architecture to realize mixed-initiative and evaluation results based on a video-study of the system. We show that users benefit from the provided dialog structure to result in predictable and successful human-robot interaction.**

I. INTRODUCTION

In the last years, robotic platforms have made significant progress towards increasing autonomy in constrained as well as increasingly open environments. Here, the ultimate goal of policy design is to increase the flexibility of accomplishing a dedicated task despite unforeseen events. The task specification itself is completely decoupled from its execution.

One of the most striking changes that service robotics has brought into view is the interaction between human and robots. While strict separation was common in industrial applications for a long time, service robots have to share their environment with humans and may even collaborate with them. Thus, the earliest works in service robotics already recognized both the difficulty of human-robot interaction, due to unstructured environments and tasks [1], as well as the promise: That human-robot collaboration can substantially increase success, especially in new or unclear situations [2].

A particular challenge for interaction has been found to be at the initial stage [3], with two main issues: Firstly, users require significant training to learn about the robot's interaction [4]. Secondly, human behavior is tremendously variable, which creates an as yet unsolved problem for automatic action recognition. Thus, is it not surprising that most existing work assumes expert users, e.g., in space or rescue robotics [5], [6].

In contrast, the present work proposes a task structuring strategy that allows *untrained* users to work with a robot using natural human modalities in a peer-to-peer fashion. Whereas in previous approaches it is the human who demonstrates an object, our approach reverses the roles, with the robot providing the initial task-structure. For instance, the robot can determine interesting visual areas, engage in pointing and grasping and ask questions about its environment. The robot's initiative thus gives the human partner explicit information about its focus and language capabilities. Having learned the interaction style, the human may take the initiative as well, as our dialog system supports a mixed-initiative paradigm [7].

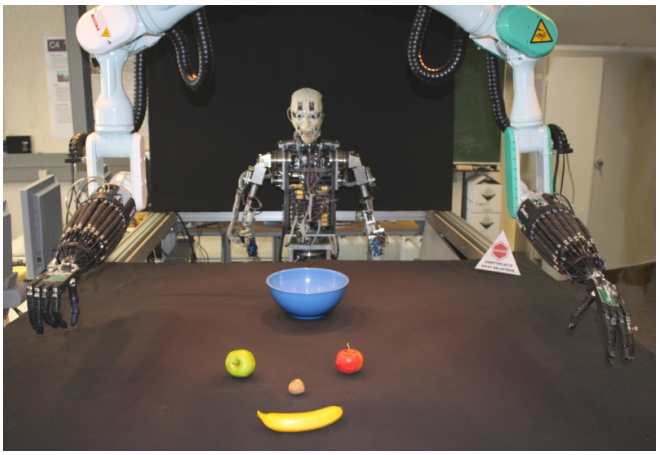


Fig. 1. The current Curious Robot Interaction Scenario. Two Mitsubishi PA-10 robot arms are fixed to the ceiling, with a left and right Shadow robot hand attached. In the background, an anthropomorphic robot torso is present. Sensors not visible are an overhead camera and a headset microphone.

From linguistic studies, it is known that humans align their behaviors to achieve more efficient dialogs [8]. A robot taking initiative can similarly influence the human's reactions, making them more predictable, particularly as the interaction target is already known. While interaction using natural modalities such as speech and gesture is often brittle, due to the difficulties of automatic pattern recognition, these constraints simplify the situation and increase robustness.

As it is by no means clear how to structure human-robot collaboration most effectively, the present work combines system development and interactive evaluation, following the general approach proposed by Hanheide et al [3]. In our scenario, the robot guides a human in an object learning and manipulation task, learning labels and grips. This task is a pre-requisite for many other applications and provides

This work was partially funded as part of the research project DESIRE by the German Federal Ministry of Education and Research (BMBF) under grant no. 01IME01N and partially supported by the German Research Council (DFG) as part of SRC 673.

a good learning environment for the user. The resulting system has been evaluated by performing video studies with inexperienced subjects, demonstrating the effectiveness of the proposed strategy.

A. Related Work

Interactive robot learning with mixed-initiative has been described by Hanheide et al for the so-called "home tour" scenario [3]. There, robot initiative provides feedback on internal models to solicit corrections by the human. This aspect has been picked in the current work, which uses the same dialog software. However, we extend it by also initiating at the start of the dialog and target learning for object manipulation instead of navigation.

Steels et al have described an interactive object-labeling scenario with the robot AIBO [9]. They show that social guidance improves learning because it focuses the robots attention. We follow their approach for social learning but add robot initiative to the interaction.

A substantial literature on the social mechanisms of human-robot interaction exists and has been surveyed in [10]. Most work addresses imitation learning or learning from demonstration in isolation. In contrast, we provide a dialog structuring strategy that can embed such methods and enable them to be used without instruction.

Object learning, e.g. for grasping or object detection, can also be performed without explicit human instruction [11], [12]. Generally, such methods require many training samples and are most suitable for acquisition of basic motor primitives. For interaction, they lack human-understandable descriptions.

Explorative behaviors based on multi-modal salience have recently been explored by Ruesch et al to control the gaze of the iCub robot [13]. The resulting behavior appears well interpretable by human observers and might be the basis for starting an interaction. At the moment, however, no further activity is created by their system. In contrast, our system uses salience just to initiate a dialog that can then acquire more information.

II. MIXED-INITIATIVE LEARNING SCENARIO

The task in our learning scenario is to acquire human-understandable labels for novel objects, learn how to pick them up and place them in a container on the working surface. Several grasping primitives are known to the system but their association to the objects is not. Through a dialog system and speech recognition, the human partner is collaborating with the robot in a peer-to-peer fashion to provide missing knowledge and error correction.

A. Dialog Shaping Strategy

As outlined before, we would like the robot to guide the user, particularly at the beginning of an interaction. Therefore, we have chosen a bottom-up strategy to drive the robots interest, as this requires no interaction history. Many potential bottom-up features exist and we have architected the system to be extensible in this respect.

The first implementation is based on visual salience, a well-established feature to determine interesting objects in the robot's visual field [14]. It provides a ranking (cf section II-D) of interaction targets, which the robot may ask about to start the interaction.

To disambiguate its focus of interest, the robot produces appropriate gesture (such as pointing), when asking for an object label. This allows us to bootstrap the dialog without knowledge of object attributes by using the robots embodiment. Last, but not least, the robot provides verbal feedback about its goal during motor activities.

For interaction with inexperienced users, we consider the structure provided by the robot to be the most important factor. However, the human tutor often has helpful comments or may detect errors earlier than the robot. For these cases, the support for mixed-initiative allows the user to actively engage in the robot's action at any time.

B. System Description

The hardware used for the interaction scenario is shown in figure 1. It is a bi-manual, fixed setup that affords complex grasping and manipulation capabilities. To achieve a robot capable of interacting with a human in a natural way, a number of perception and dialog components are needed in addition to the robot control software. An overview of the components present is given in figure 3 and the activity diagram showing their interaction is shown in figure 2. We will first give an overview of the whole system, before describing some components in detail. The system is built using the XCF middleware toolkit [3].

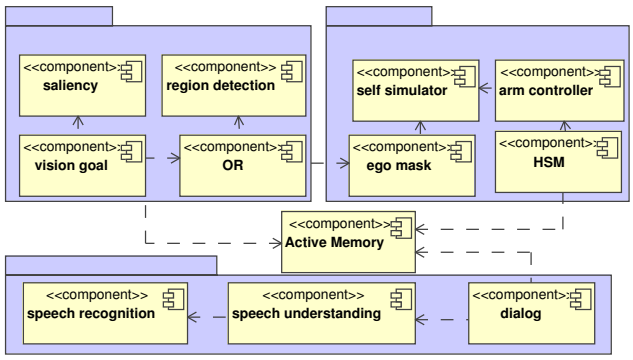


Fig. 3. System Components.

The system is composed of three major parts: Perceptual analysis, task generation ("initiative") and dialog-oriented task execution. These three parts communicate exclusively through events, where sink components register for event types they are capable of handling. Their interaction is shown in figure 2, and described in the following.

The dialog shaping strategy occurs in the perceptual and initiative parts of the "system-level" lane: Visual analysis creates events describing interesting regions ("interest items") which are then ranked and proposed as new dialog actions. See section II-D for details.

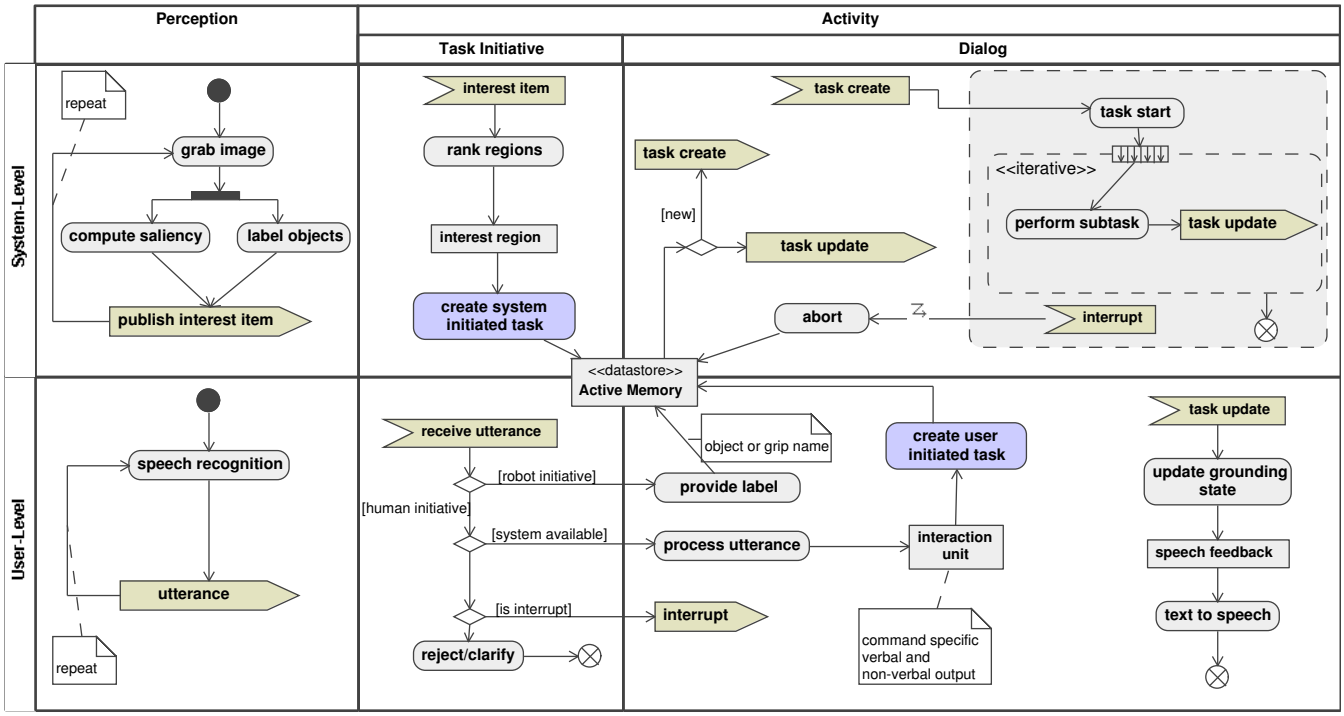


Fig. 2. UML 2.0 System Activity Diagram. Note that components execute in parallel and communicate using event signals, facilitating extensibility in the proposed system. Different input modalities are mapped to different task types to realize mixed-initiative.

Parallel to that, user input is always possible and handled in the "user-level" lane. It is important to note that user utterances may serve different purposes: For example, they may be replies to robot questions or commands. See section II-E for more information.

Task execution and coordination is the main responsibility of the "dialog" part. Activity in this part occurs both verbally (replies, questions, progress feedback) and non-verbally (pointing and grasping). The main point here is that coordination between various components and progress in subtasks is coordinated through the Active Memory [15], which stores task descriptions and notifies participating components when they are updated during execution. Thereby, the various components do not have to directly know each other but simply provide and receive information items.

Objects are grasped using one out of three basic grasp prototypes, as shown in figure 4, created from a previously developed algorithm [16]. Pick-and-Place operations are coordinated using hierarchical state machines, which parameterize appropriate low-level robot controllers [17].

C. Perceptual Analysis

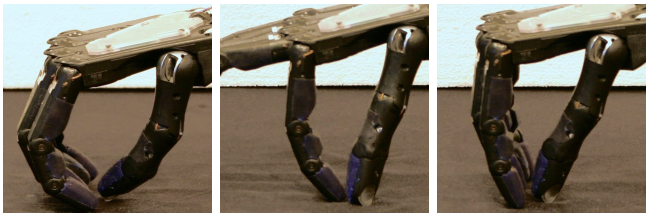
Perceptual analysis is multi-modal, including speech, vision and proprioception. Speech is recognized by the ESMERALDA speech recognizer [18], with speaker-independent acoustic model, and a situated speech-understanding component [19].

Visual analysis employs standard components for saliency, blob detection and object recognition. Please note that initially, object recognition is untrained and thus only salience and blob detection will produce events. Saliency computation is based on a psychologically motivated algorithm [14], which would also work with much more cluttered scenes. Proprioception is used to estimate the robots own position in the visual field, so that we may ignore self-generated events.

D. Saliency-Driven Task Proposal

As previously mentioned, the robot should help structure interaction by pointing out what it is interested in. In our current scenario that is "grasping of visually interesting regions in the robot's immediate vicinity". Starting point for the task selection process is the ranking of visual regions, to select an interaction target. Besides its saliency value S_i, each region may be associated to additional context information, i.e. the object label and required grip prototype.

The exact formula for the ranking function is extensible and should depend on which tasks the system supports. At the moment, we fuse bottom-up (salience) and top-down (object/grip label) information using the formers numerical



(a) power grasp (b) two finger precision (c) all finger precision

Fig. 4. Basic Grasp Primitives

value and a binary indicator variable for the latter: With salience S_i of the i'th object in $[0, 1]$ and $I_{ij} = 1$ if the j'th piece of information is available, 0 otherwise, the top region is given by $\text{argmax}_i \left(S_i + \sum_j I_{ij} \right)$.

To acquire information through the dialog, three different task types exist: "acquire label", "acquire grip type" and "grasp". In the beginning the robot only has salience information available, so it simply selects the region with highest salience as its focus and emits an "acquire label" task. Having received a label, more components become active and their information is fused based on the spatial overlap of their corresponding regions. The task initiative component then sequentially requests the information that is still missing by emitting the appropriate tasks. See figure 5 for an illustration.

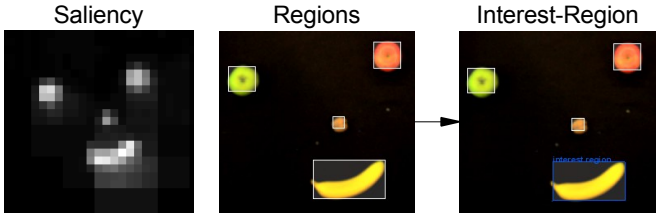


Fig. 5. Example illustrating the fusion of the object detector's and salience module's outputs. The top ranked "Interest-Region" is highlighted.

E. Interactive Learning Framework

The interactive learning framework is realized by a multi-modal dialog system based on grounding [20]. Its extensible architecture can use both human and system generated task initiative proposals for mixed-initiative interaction. Dialog examples of the current system are given in table I.

Initiative	Interaction goal	Example subdialog
Robot	Acquire label	R: What is this? ⟨pointing⟩ H: That is a banana.
	Acquire grip	R: How can I grasp the banana? H: With the power grasp.
	Grasp	R: I am going to grasp the banana. R: I start grasping now. R: ⟨grasping⟩ R: OK!
Human	Command grasping	H: Grasp the apple! R: OK. I start grasping now. R: ⟨grasping⟩ R: OK!
	Interrupt system	H: Stop! R: OK, I'll stop. ⟨stops grasping⟩ R: OK!

TABLE I

EXAMPLE DIALOGS FOR BOTH INITIATIVE TYPES.

For effective interactive learning, a framework has to fulfill two objectives: Constrain what is to be learned and focus the attention of the learner [9]. While usually the human provides structure, we achieve it by using robot initiative, with the benefits outlined in the introduction. For example, the learning task (label or grip) is constrained through the robot's question and the focus of attention is given initially through deictic gesture and later, after learning, also by verbal reference.

One consequence of reversing the roles is that the robot becomes more autonomous, which naturally has implications for interaction. To let the user know what is happening, the autonomous activities of the robot must be made transparent [21]. We address this by providing verbal feedback during longer motor activities. For example, during grasping, we announce the goal before moving the arm, the beginning of the grasp when in position and acknowledge when the object has been placed (compare table II, lines 9-15).

Communication of the robot's hypotheses is provided to enable the user to control the learning process. This is primarily done by repeating the information learned (see table II, lines 5 and 8).

Another important issue is communication of the robot's expectations. Of course, the best solution would be to design the interaction in a way that the human intuitively meets the expectations. As the experiment showed, this is achieved with the robot's question for the object label (see section IV-B). However, if the robot has expectations that are not that obvious, they are communicated explicitly. For example, the robot can pass the bowl to the human and it needs to know when the human is ready to accept it, which is requested verbally during the activity.

Last, but not least, human-initiative is handled by the learning framework depending on the current interaction state. We distinguish replies, commands and interruptions. Replies are based on robot initiative and may only occur after robot initiative. Commands propose tasks such as grasping, and are accepted when the system is idle. Otherwise, the dialog provides helpful feedback on when it will become possible again. Interruptions, such as "Stop" pertain to the currently executed task and are immediately applied, e.g. by interrupting motor control.

III. SYSTEM EVALUATION

From the issues identified in the introduction we have two main items to evaluate. Firstly, the questions posed by the system should give users a better idea of what to do next, particularly at the beginning. Secondly, the guidance provided by the system should be well understood by the human partner and trigger more predictable answers. As an aside, we were interested in how humans combine modalities when demonstrating grasps to the robot.

The evaluation has been performed as a video-study, where an interaction with an experienced test person was recorded and shown to the test subjects. During the video, questions are posed to the subjects, as shown in table III. The questions are asked after the robot has acted, but prior to the moment where the recorded person answers, to guarantee an unbiased answer. We can then compare user's reactions in the varying interaction situations. The difference in the responses for the various situations can give us insight on the general effectiveness of the guidance (our first item), and the variability in the responses indicates whether the constraints increase predictability (our second item).

The advantage of a video study like this one is that diverse interactions may be explored without frustrating the

	Initiative	Speaker	Intention	Verbal Unit	Non-verbal Unit	Comment
1	Human	Human	-	Hello, robot.	-	
2		Robot	Greet	Hello!	-	
3	Robot	Robot	Learn label	What is this?	Point at object	
4		Human	- .	This is a banana.	-	
5		Robot	Confirm hypothesis.	Banana. OK.	-	
6	Robot	Robot	Learn grip	How can I grasp the banana?	-	
7		Human	- .	With power grasp.	-	
8		Robot	Confirm hypothesis.	Power grasp. OK.	-	
9	Robot	Robot	Explore grip	I am going to grasp the banana.	-	
10		Robot	Confirm	OK, I start grasping now.	Grasp	Grasp will fail
11	Human	Human	-	Stop!	Release	
12		Robot	Abort action	OK, I stop.	-	
13	Human	Human	-	Grasp the banana!	-	
14		Robot	Confirm start	OK, I start grasping now.	Grasp	
15		Robot	Confirm end	OK!		Grasp successful
16	Human	Human	-	Good bye!	-	
17		Robot	Say goodbye	Good bye.	-	

TABLE II
EXAMPLE DIALOG

Time (mm:ss)	Situation	Question
00:07	Scenario shown	What do you think could this robot do? How would you instruct this robot?
00:29	"What is that?"	What would you do now?
00:47	"How can I grasp that?"	What would you do now?
00:51	"Beg your pardon?"	How would you correct?
03:40	Failed to grasp apple.	What would you do now?
06:33	Points at empty position.	What is happening?

TABLE III
STUDY PLAN

subjects, because they can show their intuitive behavior first, which may or may not be supported by the system, yet, and then observe continue further interactions based on the behavior the experienced test subject demonstrates. The obvious disadvantage is that results may not directly generalize to direct interaction. However, video studies have been shown to generalize well when correctly analyzed [22]. Therefore, we consider the benefits of early feedback to outweigh the potential drawbacks and use video studies as one tool for experiment-driven system design.

A. Experimental Setup

In the experiment, the user and the robot collaboratively identify objects lying on the table, coordinate how to grasp an object and then the robot places them in a bowl (see figure 6). Ten test subjects were recruited from visitors to a university wide outreach event and thus had varying backgrounds and ages. They did not receive any instruction whatsoever but were told that we intend to broaden the interaction capabilities of the robot and that any action they would like to take was acceptable and of interest to us.

The video shown includes several dialog acts with varying initiative, changes to the scenario and several instances of successful error recovery. The duration of the interaction as shown to the subjects was seven minutes. We videotaped the subjects during the experiment and had them take a short questionnaire at the end. A single run, including the questionnaire, took from 20 to 30 minutes. The study plan, with timing information is shown in table III.

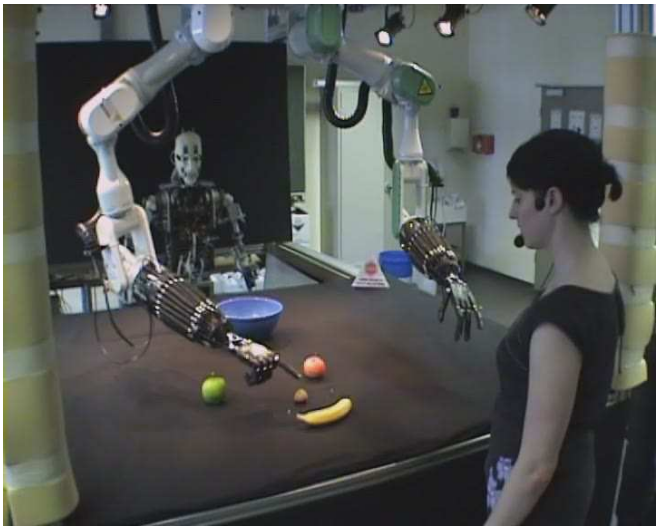


Fig. 6. Situation for "What is that?", as shown in the experiment video. The robot is pointing at the banana. The camera angle is slightly different from the real viewpoint but we did not see complications due to that.

IV. RESULTS

This section presents our findings on the effectiveness of dialog structuring and the implications for the design of robotic systems that learn from humans.

A. Initial System Description

The first situation analyzed was a static image of the scenario (similar to figure 1), where subjects were asked to speculate on the systems interaction capabilities by appearance alone. All subjects could deduce the task to be "placing objects into the bowl". They also agreed in that the system was capable of vision, grasping and speech recognition, even though no direct indication of that was given.

After that, however, the descriptions of how they might attempt to interact with the system varied widely and no clear pattern emerged. For example, one person said "Take the green apple and put it in the blue bowl" while another provided "I would explain that it should combine the four things" and a third said "Make a fruit-salad!". A summary

of the variations is shown in table IV. Apart from variations in terminology and concepts, we consider it particularly interesting that half the subjects only used meta-commentary, such as in the second example above, and did not provide any concrete command, even though the experimenters prompted them multiple times. This may have been due to the study setup, but as we can see in later parts, subjects did produce concrete example commands when it was clear to them what they could say.

Label Domain	fruit name 80%	"object" 20%		
Container Label	"bowl" 40%	"dish" 40%	none 20%	
Attributes Used	none 50%	Shape 40%	Color 30%	Size 10%
Subtask	none 70%	sorting 30%		
Commands Given	none 50%	"put *a* in *b* 20%	"put all..." 20%	"sort" 10%

TABLE IV

PERCENT OF SUBJECTS USING A PARTICULAR CONCEPT

B. Reactions to System Guidance

In contrast, answers to the "What is that?" question by the robot were considerably more consistent, as shown in table V. Only three constructions were used in total and they are all slight variations of a single sentence. The subjects apparently found it easy to answer this question, as they needed only an average five seconds to answer (measured from end of question to end of answer). Only one subject required clarification.

We also looked at an error condition, where the system pointed at an empty spot, and here two variations occur, roughly in equal proportion: Asking for clarification and giving the name of the closest object. The latter were always accompanied by comments expressing that an error occurred and thus recognizably different from regular replies.

Situation	Answer	Percent of Subjects
"What is that?"	"That is a..."	70%
	"a ..."	20%
	"a yellow ..."	10%
empty pointing	"What do you mean?"	50%
	(pointing wrong) "That is a ..."	40%
	"nothing"	10%

TABLE V

REPLIES AFTER SYSTEM INITIATIVE

C. Description of Grasping

One of the questions used during the trial was "How do I grasp the 'object'?". The robot did not provide any indication on which aspect of grasping it wants described, hence this question is considerably more open than the others. The motivation underlying this question is twofold: Firstly, we wanted to see how subjects react to unclear guidance and secondly, we wanted to get uninfluenced results on how subjects naturally describe grasping. Table VI shows the aspects used (sometimes several aspects were given). Results were very clear: Subjects took an average of 19 seconds to answer, compared to just 5 seconds for the label question.

Aspect Described	Percent of Subjects
Effector position relative to object	30%
Trajectory of effector	20%
Fingers to Use	40%
Force to Use	30%
Grasp point on object	20%

TABLE VI

ASPECT OF GRASPING DESCRIBED.

D. User Initiative

An example of user initiative can be observed in a situation where the robot fails to grasp the object. These utterances are syntactically more varied, particularly when users provide corrections, see table VII. However, they are conceptually much more straightforward than the initial descriptions and we consider it promising that users do provide verbal commentary relating to grasp parameters, such as "rounder" or "softer", which are complementary to visual demonstration.

Answer	% of Subjects
"Try again"	40%
"Grasp the ..."	20%
Grasp corrections ("rounder", "both hands", "softer")	40%

TABLE VII

USER COMMANDS AFTER FAILED GRASP

E. Discussion

Speculation behavior. From the initial speculations of the users, we can see that subjects tend to make judgments of the sort "because multiple colors appear, the system can differentiate colors", thus assuming capabilities that the system may not actually support. In our case, they assumed object labels to be known, which was not the case and would have been a problem if not for the system's guidance. This illustrates the (sometimes accidental) influence of appearances, and a dialog system should be prepared to address such preconceptions.

Detecting subject uncertainty. It was notable that subjects sometimes used meta-commentary ("I would have...") and sometimes gave very explicit answers, despite the same amount of prompting by the experimenters. We surmise that when the subjects used meta-commentary, they would have been unsure of what to do in a real situation.

In contrast, responses after guidance by the system were extremely consistent, almost to the point of being exact repetitions. Even reactions to errors were surprisingly consistent and corrections were provided without hesitation. We expect that these results will generalize due to the great consistency between subjects, even though the test group comprised just ten subjects.

From this we can conclude that task-structuring by the robot is necessary and should include not just verbal help but also contextual constraints. Our results indicate that the proposed method achieves this for object reference but that grasp descriptions need more guidance.

Discourse structuring Another result from the responses is that a dialog system is required and simple "question-reply" not sufficient: Requests for clarification occur fre-

quently and user initiative plays an important role for error detection. Additionally, even though utterances are relatively consistent conceptually, there are still considerable syntactical variations present.

The responses by the test subjects also show that the interaction as currently implemented would not be their preferred mode of interaction in some cases. The preferred alternatives were relatively few and consistent, so that they can be implemented in the next iteration of the system.

An aspect that remains open is how to let users know when they may interrupt the system, with additional commentary or error feedback. The study design prompted them, but in a real situation, other cues are necessary. This is basically a social interaction issue and it would thus be interesting to add more social feedback mechanisms to the interaction.

V. CONCLUSION

We have presented an interactive robot-learning scenario that supports inexperienced users through task-structuring and proposed a structuring strategy based on saliency and the dialog history. Results indicate that our system creates interactions consistent between users while keeping the ability for user initiative.

The resulting interaction is also much closer to the technical capabilities of the system than an unstructured dialog, without incurring the constraints of traditional system-initiative approaches. A mixed-initiative dialog system can thus provide valuable clues to the user for interacting with the system and make its capabilities more transparent.

Very promising results have been seen regarding verbal commentary during demonstration of gesture and during error feedback. The provided input is complementary to visually available information and thus provides a valuable additional clue for learning. We plan to explore this avenue in future work, to tightly integrate dialog with the learning of manipulative actions and regarding error feedback, based on the results presented.

To summarize, we have shown that a bottom-up initiative can provide dialog structure to guide users during interaction with the robot and significantly improve interaction success, even without additional instruction. Thereby, we have significantly lowered the bar for interaction with the robot system.

VI. ACKNOWLEDGMENTS

We are indebted to helpful discussions with Manja Lohse about the study design and to the participants of our study for their kind cooperation.

REFERENCES

[1] K. G. Engelhardt and R. A. Edwards, *Human-Robot Integration for Service Robotics*. Taylor & Francis Ltd, 1992.

[2] T. Fong, C. Thorpe, and C. Baur, "Collaboration, dialogue, human-robot interaction," in *Advances in Telerobotics*, 2003, pp. 255–266. [Online]. Available: http://dx.doi.org/10.1007/3-540-36460-9_17

[3] M. Hanheide and G. Sagerer, "Active memory-based interaction strategies for learning-enabling behaviors," *International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2008.

[4] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, "The vocabulary problem in human-system communication," *Commun. ACM*, vol. 30, no. 11, pp. 964–971, November 1987. [Online]. Available: http://dx.doi.org/10.1145/32206.32212

[5] T. Fong, C. Kunz, L. M. Hiatt, and M. Bugajska, "The human-robot interaction operating system," in *HRI '06: Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. New York, NY, USA: ACM, 2006, pp. 41–48. [Online]. Available: http://portal.acm.org/citation.cfm?id=1121241.1121251

[6] R. R. Murphy, "Human-robot interaction in rescue robotics," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 34, no. 2, pp. 138–153, 2004. [Online]. Available: http://dx.doi.org/10.1109/TSMCC.2004.826267

[7] J. F. Allen, "Mixed-initiative interaction," *IEEE Intelligent Systems*, vol. 14, no. 5, pp. 14–23, 1999.

[8] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and Brain Sciences*, vol. 27, pp. 169–226, 2004.

[9] L. Steels and F. Kaplan, "Aibo's first words: The social learning of language and meaning," *Evolution of Communication*, vol. 4, no. 1, pp. 3–32, 2001.

[10] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonomous Systems*, vol. 42, no. 3-4, pp. 143–166, March 2003. [Online]. Available: http://dx.doi.org/10.1016/S0921-8890(02)00372-X

[11] L. Natale, F. Orabona, G. Metta, and G. Sandini, "Exploring the world through grasping: a developmental approach," in *Proc. of Computational Intelligence in Robotics and Automation*. IEEE, June 2005, pp. 559–565.

[12] P. Fitzpatrick, G. Metta, L. Natal, S. Rao, and G. Sandini, "Learning about objects through action - initial steps towards artificial cognition," in *Proc. IEEE Int. Conf. on Robotics and Automation*. IEEE, 2003.

[13] J. Ruesch, M. Lopes, A. Bernardino, J. Hornstein, J. Santos-Victor, and R. Pfeifer, "Multimodal saliency-based bottom-up attention a framework for the humanoid robot icub," in *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, 2008, pp. 962–967. [Online]. Available: http://dx.doi.org/10.1109/ROBOT.2008.4543329

[14] Y. Nagai, K. Hosada, A. Morita, and M. Asada, "A constructive model for the development of joint attention," *Connection Science*, vol. 15, no. 4, pp. 211–229, December 2003. [Online]. Available: http://dx.doi.org/10.1080/09540090310001655101

[15] J. Fritsch and S. Wrede, *An Integration Framework for Developing Interactive Robots*, ser. Springer Tracts in Advanced Robotics. Berlin: Springer, 2007, vol. 30, pp. 291–305.

[16] F. Rothling, R. Haschke, J. J. Steil, and H. Ritter, "Platform portable anthropomorphic grasping with the bielefeld 20-dof shadow and 9-dof tum hand," in *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on*, 2007, pp. 2951–2956. [Online]. Available: http://dx.doi.org/10.1109/IROS.2007.4398963

[17] H. Ritter, R. Haschke, and J. Steil, "A dual interaction perspective for robot cognition: Grasping as a "rosetta stone"," 2007, pp. 159–178. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-73954-8_7

[18] G. A. Fink, "Developing HMM-based recognizers with ESMERALDA," in *Lecture Notes in Artificial Intelligence*, V. Matoušek, P. Mautner, J. Ocelíková, and P. Sojka, Eds., vol. 1692. Berlin Heidelberg: Springer, 1999, pp. 229–234.

[19] S. Hüwel, B. Wrede, and G. Sagerer, "Robust speech understanding for multi-modal human-robot communication," IEEE Press. IEEE Press, 2006, inproceedings, pp. 45–50. [Online]. Available: files/papers/Huewel2006-RSU.pdf

[20] S. Li, B. Wrede, and G. Sagerer, "A computational model of multi-modal grounding," ACL Press. ACL Press, 2006, inproceedings, pp. 153–160. [Online]. Available: files/papers/Li2006-ACM.pdf

[21] T. Kim and P. Hinds, "Who should i blame? effects of autonomy and transparency on attributions in human-robot interaction," *The 15th IEEE Interantional Symposium on Robot and Human Interactive Communication (RO-MAN06)*, pp. 80–85, September 2006.

[22] S. N. Woods, M. L. Walters, K. L. Koay, and K. Dautenhahn, "Methodological issues in HRI: A comparison of live and video-based methods in robot to human approach direction trials," in *Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2006, pp. 51–58.