What makes Human-Robot Dialogues struggle?

Petra Gieselmann
Interactive Systems Lab
Universität Karlsruhe
Germany
petra@ira.uka.de

Alex Waibel
Interactive Systems Lab
Carnegie Mellon University
Pittsburg, PA 15221
ahw@cs.cmu.edu

1 Introduction

During the last few years, humanoid robots became very popular in the robotic research community and some humanoid robots are already commercially available, such as Asimo from Honda or Qrio from Sony. Comparing the currently possible human-robot communication with the human-human communication we can see that in human-human communication we have efficient strategies to avoid errors and also to recover from them, such as for example *grounding* new information (Traum, 1999; Traum and Dillenbourg, 1998; Poesio and Traum, 1998). This is still one of the biggest challenges for human-robot communication to develop a system which can cope with real world situations and is error tolerant so that it can react in a reasonable way even when something has been misunderstood or not understood at all. Therefore, in this paper we want to evaluate problematic situations in human-robot-communication and how they can be resolved.

Our target scenario is a household situation, in which the user can ask the robot questions related to the kitchen, such as “What’s in the fridge ?”, ”How do I cook Spaghetti Napoli?”, ask the robot to set the table, to switch certain lights on or off, to bring some objects, such as cups, dishes, etc. (Gieselmann et al., 2003; Stiefelhagen et al., 2004). In this context which is specifically tailored for unexperienced and older users, it is important that the user can talk to the robot in the same way as to a human servant. This means that the communication should be as natural and as comfortable as possible for the user and therefore, errors should be avoided or at least easy to correct, if they cannot be avoided beforehand.

We can distinguish two kinds of errors: *Non-understanding* vs. *misunderstanding*. Non-understanding means that the dialogue manager cannot find any information in the user utterance. This can be due to the fact that the grammar does not cover the user utterance which cannot be parsed therefore. Also on the pragmatic level, non-understanding is possible, when the user utterance is inconsistent with the current discourse. Misunderstanding means that a user utterance can be parsed and the semantic interpretation is integrated in discourse, but does not correspond to the user’s intention. This is above all due to speech recognition errors which means that a word has been misrecognized. But also a semantic misunderstanding might be possible, if some information from the user utterance has been integrated wrongly in the existing discourse.

Therefore, in this paper we want to classify the different kinds of errors which occur in human-robot communication. Section two gives an overview of related work on errors in human-machine dialogues and error classifications. Section three deals with our dialogue system: The household robot, the dialogue manager, and the web-based interface for user tests of human-robot dialogues are described. Section four gives experimental details and results, and section five gives a conclusion and outlook.

2 Related Work

2.1 Errors in Dialogues

The problems caused by errors in spoken dialogue systems are well known and can result in user frustration and task failure. Most of the research dealing with errors only take speech recognition errors into account until now. For example, Xu et al. and also Gorrell (Xu and Rudnicky, 2000; Gorrell, 2003) use different methods for dialogue state adaptation to the language model to improve speech recognition. Also different stages and language models are used to reduce word error rates and perplexity in error dialogues: A general n-gram language model is used at the beginning and in underspecified situations and a specialized language model which can be an n-gram language model or a grammar-based one is used in specific situations based on the preceding system prompt (Fosler-Lusier and Kuo, 2001). In (Solsona et al., 2002), the state-independent n-gram language model is also combined with a state-dependent finite state grammar by comparing the acoustic confidence scores. Furthermore, work on hyperarticulation concludes that speakers change the way they are speaking when facing errors in principle so that the language model has to be adapted therefore (Stifelman, 1993; Hirschberg et al., 2004).

Choularton (Choularton and Dale, 2004) examines different repair strategies of the users and how these strategies can be generalized to be domain-independent. Also Stifelman explains the user reactions to errors and how repair utterances can be automatically detected on the acoustic side (Stifelman, 1993). Both of them are looking for general strategies on error recognition and repair to prepare the speech recognizer better to the special needs of error communication.

Our concern, however, is with slightly different analyses in order cope with errors more efficiently: We want to concentrate on semantic errors and how they can be classified. We avoided speech recognition errors by using an interface with keyboard input to our robot, as explained in section 3.3. We want to find out the reasons for errors in order to avoid them as far as possible. Furthermore, we want to have a look at repair dialogues in order to be able to perform efficient error handling strategies in the future so that it is easier for the user to correct errors which could not be avoided.

2.2 User Tests and Error Classification

At the moment, there exist only very few error classifications based on the semantics of user utterances. Most of the researchers use the Levenstein distance (Levenstein, 1996) which gives the cheapest way to transform one string into another one by combining the following steps:

- **Substitution** of one symbol by another one
- **Deletion** of one symbol by another one
- **Insertion** of a new symbol

But since this is not useful in our case to find out, why the dialogue failed, we made a new error classification which is based on the semantics of the user utterance and possible reasons why it cannot be understood by the system.

3 The Dialogue System

3.1 Our Household Robot

We developed a rapid prototype system with approximately 33 dialogue goals, 190 dialogue moves and more than 140 ontology concepts. Furthermore, we developed more than

650 grammar rules and the lexicon has now more than 250 entries. By means of this prototype we started user tests and interactively develop now new versions of the robot grammar and domain model.

The robot can accomplish different tasks in the household environment. The user can for example ask it to get something from somewhere, put something somewhere else, set the table, switch on or off different lamps, to give him information about some recipes, make a cup of coffee or tea, etc.

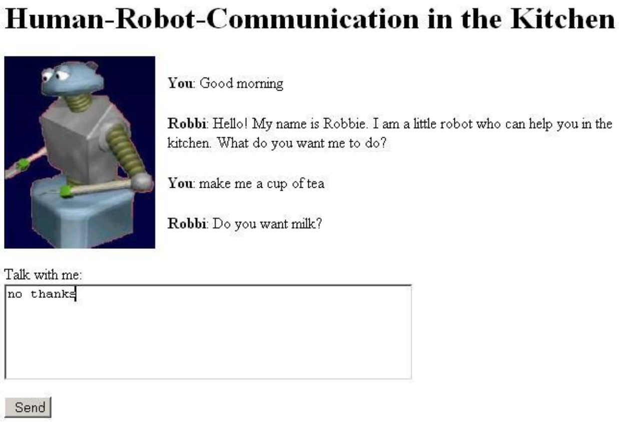
3.2 Dialogue Management

We use the TAPAS dialogue tools collection based on the approaches of the language and domain independent dialogue manager ARIADNE (Denecke, 2002) which is specifically tailored for rapid prototyping, so that can interactively develop new versions relying on the same base technology. We developed the domain and language dependent components, such as an ontology, a specification of the dialogue goals, a data base, a context-free grammar and generation templates.

The dialogue manager uses typed feature structures (Carpenter, 1992), to represent semantic input and discourse information. A context-free grammar enhanced by information from the ontology defining all the objects, tasks and properties about which the user can talk parses the user utterance. The parse tree is converted into a semantic representation and added to the current discourse. If all the information necessary to accomplish a goal is available in discourse, the dialogue system calls the corresponding service, such as ”getting the cup from the table to the user”. Otherwise, the dialogue manager generates clarification questions to the user by means of generation templates.

3.3 Web-based User Interface

An internet user test has the advantage that lots of users all over the world can participate whenever they like to so that the costs in time and money are lower than in other user studies (Schmidt, 1997). Also Reips explains these advantages of web-based experiments, such as ”speed, low cost, experimenting around the clock, and a high degree of automation” (Reips, 2002). Therefore, we made the rapid prototype accessible via the internet, as you can see in figure 1 and posted the link to different news groups and added it to some experimental portals in the web to get as much user data as possible.



Human-Robot-Communication in the Kitchen

You: Good morning

Robbi: Hello! My name is Robbie. I am a little robot who can help you in the kitchen. What do you want me to do?

You: make me a cup of tea

Robbi: Do you want milk?

Talk with me:

no thanks

Figure 1: The web-based Internet Interface for our Humanoid Robot

One drawback of web-based experiments is that users might dropout quite easily because there is no experimenter available who forces them to stay (Reips, 2002). But at the same time especially in our case this resembles much more the real world situation where the user has the robot in his own home and can decide whether he wants to use it or not. Therefore, we carefully evaluate all the situations when the users dropped out to avoid them in the future for a more comfortable use of the robot.

4 Experimental Details & Results

4.1 Details

The data are collected with about 70 test persons. All together, we have about 1000 turns;

```
1. Ask Robbi to make you a cup of tea with milk and sugar.
2. Ask Robbi to get you some water.
3. Ask Robbi to get you the blue cup.
4. You would like to cook Spaghetti Napoli.
Ask Robbi, how to do this.
5. You invited some friends for diner.
Ask Robbi to set the table for all of you.
6. Ask Robbi to make you a cup of coffee without milk, but with sugar.
7. Ask Robbi to get you some coke.
8. Ask Robbi to switch on the small lamp.
9. Imagine that you come home after work and are very hungry.
Now you want Robbi to cook something for you.
10. Imagine that you are sitting on your sofa thinking what you might cook
this evening. Since you are too lazy to go to the kitchen, you ask Robbi
to have a look at the fridge, what is still there.
```

Figure 2: Tasks for the User Test

on average, there are 15 turns per user. All the users talked to our robot via the webinterface and got the instruction to make the robot do five of the predefined tasks you can see in figure 2.

	on Average
Accomplished Tasks	2.65
New Tasks	0.81
New Objects	0.53
New Words	3.34
Overall Turns	14.48
	Rates (in %)
Parsability	74.62%
Turn Error Rate	56.2%
Finalized Goal Rate	25.3%
Dropout Rate	1.22%

Table 1: Detailed Results

4.2 Results

As you can see in table 1, the users managed to let the robot do more than half of the predefined tasks. The turn error rate was quite high because the system was only a prototype which did not cover all the utterances the users invented. Furthermore, some users did not read the instructions carefully and entered punctuation marks and digits which could not be parsed by the current dialogue manager because it expects input similar to the one from a speech recognizer. Therefore, we want to integrate a small component which can delete all the punctuation marks in the future.

Since the grammar was only a prototype, it did not cover all the user utterances, but some new concepts were used. In addition, we also found some new goals which were not covered by our dialogue manager. These new goals concern above all meta-communication, such as ”what can you cook?” or ”do you know the word coffee?”. Since the users got predefined tasks to accomplish, most of the other goals are already covered by the grammar. For the same reason, the users refer only to very few new objects, such as new recipes for example. They used some new words for known objects, such as ”cream” instead of ”milk”.

Since a conversation which consists of less than five turns means that the user talked to the robot less than a minute, we determined five turns as a limit for a conversation. Only very few users dropped out given this limit of five user turns, but most of the users seemed to have acquired a taste for the robot communication and went on talking with it for quite some time. All the users who dropped out did not manage to make the robot understand them at all during these first few turns which was most of the time due to the problem with punctuation marks mentioned above.

About three fourth of the user utterances can be parsed, but some of them cannot be transformed to the complete, correct semantic representation which explains the slightly higher turn error rate. We now want to have a closer look at all the utterances which cannot be understood correctly and results in errors. Therefore, we manually tagged all the utterances by means of the reasons why they failed, as you can see in table 2.

4.3 Error Analysis

We noticed that the main reason for errors were new ontological or grammatical concepts (cf. Table 2). Lots of new syntactical constructions were used, such as ”prepare a salad” instead of ”make a salad”, ” i want you to cook spaghetti for me” instead of ”please make spaghetti napoli”. Sometimes the participants used also new words for known objects, such as ”icebox” instead of ”fridge” or ”soda” instead of ”water”. This might be due to the fact that we only had a small prototype grammar. It is possible that a more complete grammar would result in lesser errors in this area. This could be explored in future studies.

Also some new goals were used by the participants, such as ”switch yourself off”, ”can you wash the dishes”. But above all most of the new goals can be defined as meta-communication and clarification questions from the user as already described in the previous section. When the robot did not understand the user, he tried to detect what went wrong by asking questions such as ”are you making the coffee?” or ”can you understand me?”. Therefore, we want to integrate a component in the future which can deal with all this kind of meta-communication and has access to the context model and the discourse to include the previous user utterances.

Very few new objects were used such as ”cupboard”, ”dustbin”. The small grammar seems to already cover most of them because we have such a fixed set of tasks the user should accomplish. It would be an interesting topic for future studies to see whether more complex task sets also require a bigger variability within the vocabulary.

Sometimes, the context to resolve an utterance is missing and also elliptical utterances and anaphora can be found quite often. As you can see in Figure 3 in the first example, where the users refers to the ”lamp” by saying ”the small one”, we need to include context management issues in future versions to resolve elliptical and anaphoric utterances.

On the other hand, we also have some utterances which are too complex and contain concatenated sentences which cannot be resolved at the moment, such as for example ”I need a cup of coffee that has about a quarter cup of milk in it”, ”I want you to cook spaghetti for me and a coke”, etc. In theses cases, we want to make sure that at least one part of the utterance can be understood so that the user can repeat the other one later.

In addition, lots of users entered punctuation marks and digits which could not be parsed by the current dialogue manager, such as ”set the table for 4 people”, ”How do I cook Spaghetti Napoli?”, ”Yes.”. All the digits and words with punctuation marks at the end are simply ignored by the parser and result in errors therefore.

And we have some grammatically wrong utterances, such as ”can you put the light on”, ”you bring coffee”, ”cook a salad for me”, ”i want drink”, etc. Since these seem to be due to the fact that some of the user were non-natives, we want to integrate therefore possibilities to parse at least parts of ungrammatical sentences so that the robot can ask clarifications questions to the user to understand what he wants to say.

4.4 User Behaviour with Respect to Problems and Errors

The behaviour of the users when facing problems is very interesting: Lots of users tend to

	Rate of All Utterances	Rate of not Understood Utterances
New Concepts		
New Formulation	3.34%	39.82%
Meta Communication	0.76%	9.12%
New Goal	0.41%	5.65%
New Object	0.53%	6.32%
Ellipsis		
Elliptical Utterances / Anaphora	0.63%	7.54%
Missing Context	0.22%	2.63%
Complexity		
Concatenated Utterances	0.48%	5.79%
Input Problems		
Punctuation	0.97%	11.58%
Digits	0.22%	2.63%
Grammatically Wrong Utterances	0.37%	4.39%

Table 2: Error Analysis

use shorter utterances, sometimes even only a single word, when they notice that the robot did not understand them. For example, you can see in Figure 3 in the second example that a user who has not been understood by the robot, just repeated the object ”blue cup”.

In the third example (cf. Figure 3), you can even see a more condensed user utterance, after an error occurred: After the robot did not understand ”I want to drink coffee.” because of the punctuation mark at the end, the user just said ”kitchen-coffee”. Of course, the system would never understand this utterance.

```
User: switch on the lamp
Robbi: Which lamp do you want me
to switch on?
User: the small one

User: get me the blue cup
Robbi: Sorry. I didn’t understand you,
please try again!
User: blue cup

User: I want to drink coffee.
Robbi: I am sorry.
I can’t understand you.
User: kitchen-coffee
```

Figure 3: Example Dialogues from the User Test

In the event of an error, the users tried different words and just waited whether the robot understood them. The problem is that most of the words were correctly understood by the system, but since they could not be interpreted without context, there was no answer from the dialogue system. In the future, we will therefore implement methods to cope with single words and give the user better feedback so that he knows what had been understood.

In addition, it seems to be a general problem that the dialogue system understands some parts of the user utterance, but does not indicate this to the user so that he thinks that it did not understand. In the example below (cf. Figure 4), the user told the robot to get him the blue cup. But since this verbal construction has not been covered by the prototype grammar, the system only understood the object, namely ”the blue cup” which the user did not know but thought that everything was misunderstood and mentioned again ”blue cup”. Better grounding mechanisms have to be implemented in the future for all these cases.

```
User: can you get me the blue cup please
Semantic Representation: [generic:object,NP,_] { [robbi:obj_cup,N,_]
{ THE BLUE CUP } [robbi:please_NT] { PLEASE } }
Robbi: Sorry. I didn’t understand you, please try again!
User: blue cup please
Semantic Representation: [generic:object,NP,_] { [generic:object,NP,Sg]
{ [robbi:obj_cup,N,Sg] { CUP } [robbi:please_NT] { PLEASE } } }
```

Figure 4: Example Dialogues with semantic Representations from the User Test

5 Conclusion & Outlook

In this paper, we presented the results of an internet user test of the dialogue management component of our household. The results showed that most of the errors in human-robot communication are due to new formulations and missing mechanisms to deal with meta-communication and elliptical utterances.

Furthermore, the user test showed that lots of users tried to get the communication back on track by using shorter and shorter utterances. Unfortunately, even if these utterances had been understood correctly, the dialogue manager did not give any feedback to the user, but waited for more input. Therefore, we want to integrate a component which can handle these short utterances and adds them to the common ground. In this way, a clarification dialogue can be initiated to find out what the user wants to do. In addition, this component can also help avoiding errors resulting form elliptical utterances.

Acknowledgments

This work was supported in part by the German Research Foundation (DFG) as part of the SFB 588 and by the European Commission under project CHIL (contract #506909).

References

B. Carpenter. 1992. *The Logic of Typed Feature Structures*. Cambridge University Press.

S. Choularton and R. Dale. 2004. User responses to speech recognition errors: Consistency of behaviour across domains. *Proceedings of the 10th Australian International Conference on Speech Science & Technology*.

M. Denecke. 2002. Rapid prototyping for spoken dialogue systems. *Proceedings of the 19th International Conference on Computational Linguistics*.

E. Fosler-Lusier and H.K. J. Kuo. 2001. Using semantic information for rapid development of language models within asr dialogue systems. *Proceedings of ICASSP*.

P. Gieselmann, C. Fuegen, H. Holzapfel, T. Schaaf, and A. Waibel. 2003. Towards multimodal communication with a household robot. *Proceedings of the Third IEEE International Conference on Humanoid Robots (Humanoids)*.

G. Gorrell. 2003. Recognition error handling in spoken dialogue systems. *Proceedings of the 2nd International Conference on Mobile and Ubiquitous Multimedia*.

J. Hirschberg, D. Litman, and M. Swerts. 2004. Prosodic and other cues to speech recognition failures. *Speech Communication*, 43.

V. I. Levenstein. 1996. Binary codes capable of correcting insertion and reversals. *Cybernetics and Control Theory 10*.

M. Poesio and D. Traum. 1998. Towards an axiomatization of dialogue acts. *Proceedings of the Twente Workshop on the Formal Semantics and Pragmatics of Dialogues (13th Twente Workshop on Language Technology)*.

U.-D. Reips. 2002. Standards for internet-based experimenting. *Experimental Psychology*, 49(4).

W. C. Schmidt. 1997. World-wide web survey research: Benefits, potential problems, and solutions. *Behavior Research Methods, Instruments & Computers*, 29(2).

R. A. Solsona, E. Fosler-Lussier, H.-K. J. Kuo, A. Potamianos, and I. Zitouni. 2002. Adaptive language models for spoken dialogue systems. *Proceedings of the ICASSP*.

R. Stiefelhagen, C. Fuegen, P. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel. 2004. Natural human-robot interaction using speech, gaze and gestures. *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

L. J. Stifelman. 1993. User repairs of speech recognition errors: An intonational analysis. *Technical Report, Speech Research Group, MIT Media Lab*.

D. R. Traum and P. Dillenbourg. 1998. Towards a normative model of grounding in collaboration. *Working notes of the ESSLLI-98 workshop on Mutual Knowledge, Common Ground and Public Information*.

D. R. Traum. 1999. Computational models of grounding in collaborative systems. *Psychological Models of Communication in Collaborative Systems - Papers from the AAAI Fall Symposium*.

W. Xu and A. Rudnicky. 2000. Language modeling for dialog system. *Proceedings of ICSLP*.