Student Registration No 815675

MULTI-MODAL TASK INSTRUCTIONS TO ROBOTS BY NAIVE USERS

MPHIL/PHD TRANSFER REPORT

by

JOERG CHRISTIAN WOLF
MAY 2006

Submitted to the

UNIVERSITY OF PLYMOUTH, U.K.

SUPERVISORY TEAM
Dr. Guido Bugmann
Dr. Paul Robinson

Abstract

The Robotic Intelligence Lab is developing a theoretical framework for user-programmable robots. The objective of the project is to investigate multi-modal unconstrained natural instructions to robots in order to design a learning robot. This report gives an overview of the ongoing work on this project and elsewhere. It describes how a corpus centred approach can be used to design an agent that can reason, learn and interact with a human in a natural unconstrained way. The focus of this project so far has been on the collection of the corpus and on how to combine speech and gesture based on rules extracted from the corpus. We are trying to find novel ways of how natural instructions to a robot can be translated into a function usable by a robot. A test system of the full agent is also described. The report concludes on how much the project has progressed and what work is left to do.

Content

Abstract ..... - 3 -

Content ..... - 4 -

1 Introduction ..... - 5 -

2 Literature Review ..... - 7 -
2.1 Multimodal Robots ..... - 7 -
2.2 Natural Language Processing and Natural Language Understanding Systems ..... - 10 -
2.3 Semantic Representation Theories ..... - 13 -
2.4 Conclusion of the Literature Review ..... - 16 -

3 Method ..... - 17 -
3.1 Corpus centred approach ..... - 17 -
3.2 Multi-modal interface and corpus collection ..... - 18 -
3.2.1 User interface ..... - 18 -
3.2.2 Corpus Collection ..... - 19 -
3.2.3 Transcription Tool ..... - 20 -
3.3 Corpus Analysis and knowledge representation ..... - 21 -
3.3.1 Grammar design ..... - 22 -
3.3.2 Semantic Analysis ..... - 22 -
3.3.3 Speech and Gesture timing ..... - 22 -
3.3.4 Reasoning ..... - 22 -
3.4 Test System ..... - 23 -

4 Progress ..... - 25 -
4.1 Overview of progress and planned work ..... - 25 -
4.2 Publications and Feedback ..... - 26 -
4.3 Expected Contribution to Knowledge ..... - 26 -

References ..... - 27 -

1 Introduction

IN the future, service robots will be more common in our households. They should be programmable by anybody interacting with them, since there are far too many possible tasks for the robot to be pre-programmed completely. Users want to change the robots behaviour to their individual preference (Bugmann G 2005, Wermter S 2003). Users may not be experts in programming. Therefore "programming" of service robots should be done in the language of humans. "Programming" between humans is giving instructions from person to person. So there is a clear need for researching human to human instructions. Humans instruct (teach) by speaking demonstration of actions and gestures. Therefore a robot must be able to accept these instructions without the need for the instructor to change the way of communication (example see Fig. 1).

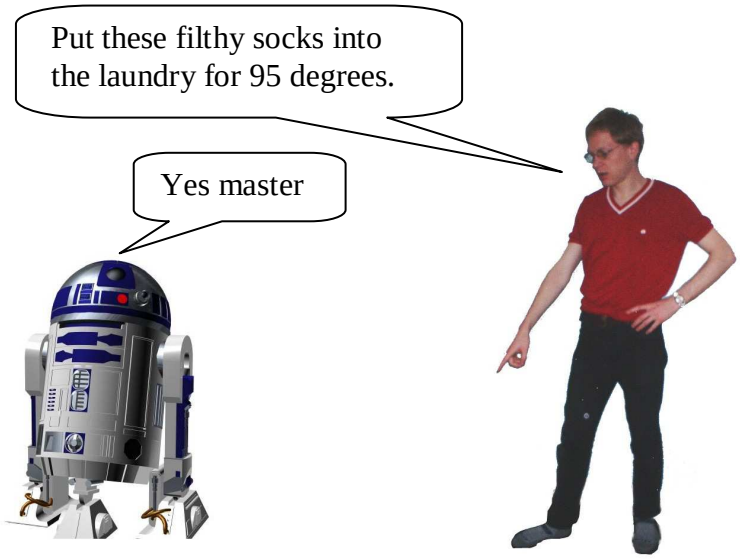

Fig 1: *Example scenario of natural interaction with a service robot.*

The target of this PhD is to contribute to knowledge in the field of human-robot communication. More specifically **how to convert unconstrained multimodal instructions (spoken natural language + gestures) into a knowledge representation usable for robot reasoning and acting**. The emphasis lies in the fact that the communication is unconstrained, so the user can communicate freely (free choice of vocabulary, free natural flow of gestures and speech). We believe that there is currently no service-robotics project focusing on this emphasis.

We try to achieve a real world implementation of a specific situation. The robot must be able to accept a wide range of commands, so that the user can say anything suitable for that situation. This is possible with a corpus based design applied to that situation. Corpus based means that a body of example dialogues from that situation is recorded. Once all the corpus sentences and corpus gestures have been implemented into the agent, the agent will be able to interact and carry out actions. For the implementation the sentences must be translated into grammar and some form of semantics that can be processed by the agent.

Previous research in our group investigated a scenario where route instruction for directions to find a location in a town where given to a robot. It turned out that all the instructions that occurred are sequences. There are domains where sequences are not sufficient. Often general rules are included in task instructions such as "Hoover the ground floor once a week, chairs must be removed before hoovering".

With the experience from the previous project criteria for the selection of a new application/task were determined:

i) The task must contain a wide range of instruction types.
(rules, sequences , repetitions)
ii) The task should be scalable from simple to complex.
iii) The task should preferably have a small vocabulary (less than 1000 words)
iv) The task must be natural to naïve users

Given these constraints, game instruction seems to be a good choice. In particular, card games come in a great variety of type and complexity, yet their vocabulary is restricted. We investigated all two player games listed in "the Oxford A-Z of Card Games" (Parlett, 2004). See (Wolf and Bugmann 2005) for details. The test system (a software agent) that is used in this PhD project will be able to learn a card game from a human-teacher.

Previous research in our group focused purely on verbal instructions which are sufficient in some cases where a demonstration with physical objects is not required. In practice, many tasks are explained using a mixture of verbal instructions, gestures and demonstrations. Thus, a truly natural interface between human and robots must be multi-modal. This is one of the features to include in this PhD works and has been the inspiration of the name of the project: **MIBL** (**M**ulti-**M**odal **I**nstruction **B**ased **L**earning). Multi-modal systems combine gesture and language. Multi-modal robots are investigated by several research groups around the world (Soshi Iba et al 2002, Wermter S et al 2003, R. Dillmann et. Al. 2002). Some relevant projects are described in chapter 2. The challenge of multi-modal robots lies in combining the modalities to form an internal model of the environment. Combining gesture and language is one of the focus points of this PhD work (Wolf and Bugmann 2006). Language is ambiguous and can only be understood by the listener when put into context. Situational knowledge, prior knowledge and gestures are responsible to create the correct interpretation of a sentence. Chapter 3.3 and 3.4 discusses this in more detail. During this PhD an example system is produced that essentially consists of a learning and interacting service robot (agent) plus all the software tools necessary to adapt this agent to a new task domain.

To collect all the information required from a future user we use the corpus based approach to robotics, described in chapter 3.1. During the first year of the PhD, a corpus of dialogues between teacher and student has been recorded and transcribed. For details see (Wolf and Bugmann 2005) and chapter 3.2. The corpus is the basis for the research. It provides information about:

- what type of instructions are used
- what parameters do they have
- how often they occur
- to what level of detail a teacher goes to explain a task
- the dialogue structure
- the dialogue primitives

Based on the corpus, concepts in the field of understanding task instructions for a robot can be established. This concepts aim at answering questions like:

- How to design a grammar from a multi-modal corpus?
- What knowledge representation and reasoning engine is suitable?
- What semantic structures are used in the language of the teacher?
- How to map language into a semantic representation suitable for a service robot?

The above list is not exhaustive. Chapter 4 of this report concludes with how much progress has been made and identifies the work left to do and the expected key contributions to knowledge.

2 Literature Review

The field of service/personal robotics is relatively new. The first International Personal Robot Congress was held in April 1984 [Engelhardt and Edwards,1992]. The field has gained more interest in the last two decades. Technological advances made it possible to construct completely autonomous mobile robots with all the computing power needed onboard and compact electric servos. Some key research areas such as navigation and object recognition are far more popular among the research community than for example “window cleaning” (Bugmann G 2005). However to build a true service robot researchers must face real world problems and problems at hand. The University of Plymouth has recently advertised three PhD scholarships addressing problems at hand on how to make domestic/service robots reality: “Imitation of intentional behaviour on artificial agents” principal investigator Dr Tony Belpaeme , “Vision-based manipulation of fabric in domestic environments” principal investigator Dr Phil Culverhouse and “The natural language robot football trainer” principal investigator Dr Paul Robinson.

From the ongoing literature review, only the most relevant reviewed publications are included in this transfer report. The topics investigated are multimodal Robots (chapter 2.1), natural language processing and understanding (chapter 2.2) and semantic representation theories (chapter 2.3).

2.1 Multimodal Robots

The German Collaborative Research Center has build two humanoid robots (Albert and ARMAR) with the target of interacting with humans in a service robot scenario (R. Dillmann et al 2002). Their emphasis lies in building a complete system that can interact through observation and tracking of objects, gesture recognition and speech recognition. The research group recognizes that interactive programming must be a One-Shot-Learning process or it would be very annoying to the user. Another important point from the paper is that there seem to be no system so far that integrates the control, basic interaction methods and programming techniques for humanoid robots into a single system. The robot build by the research institute can learn to fetch and carry tasks and can be taught fine manipulations of simple objects.

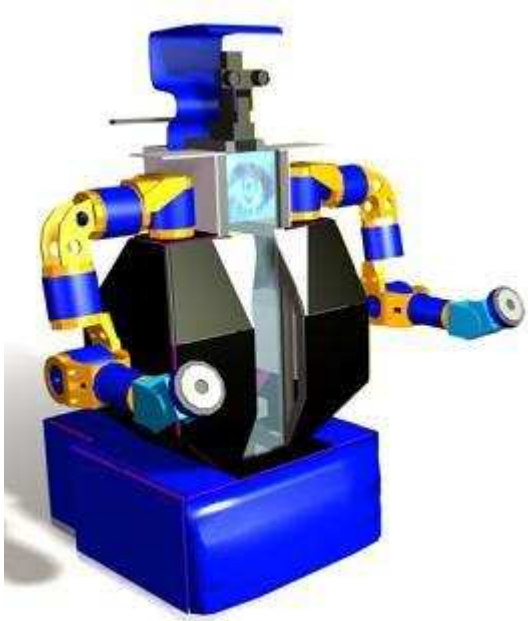

Fig. 2.: *Albert 2*

Data from the recognition of trajectories and grasping is segmented so it can broken down into a sequence in a semantic format. This system confirms with the ideas of this PhD in this respect. However sequence learning alone is not enough. For more advanced tasks rule learning is necessary.

Another project on natural multi-modal interaction and learning of a robot is going on at the Bielefeld University(Haasch A et al 2004). Their robot, called BIRON, has the capability to detect which person out of a group it has to pay attention to. The person can then engage in a simple dialogue with the robot introducing objects to the robot. (see figure on the right). BIRON can focus it's microphone beams on the person thus improve speech recognition performance. The domain of the reasoning and speech recognition engine of BIRON is limited to a simple dialogue. BIRON only understands simple sentences that introduce objects. (see figure on the right.)

Fig 3*: Typical interaction with BIRON. "This is a plant". picture from (Haasch A et al 2004)*

The robot has a vision system with gesture recognition and object recognition, a natural language interface and laser range finders.

The Bielefeld group recognized some important points which are relevant to our research:

- Combining uni-modal processing results into a multi-modal data-association framework makes the system robust against errors.
- Human communication partners can not be expected to wear special equipment such as close-talking microphone or data-gloves.
- a semantic-based grammar is necessary to extract the meaning of the sentence (parsing and subsequent interpretation is not acceptable)
- missing information in an utterance can often be acquired from the scene (other sensors)
- the system uses a horizontal hierarchy (Reactive Layer, Intermediate Layer, Deliberate Layer. (see figure 4)

The research in Bielefeld concentrated on the reactive layer (Person Attention etc.) The dialogue and high level reasoning has not been investigated enough to make this service robot all commands necessary in it's domain. The domain used is a scenario where the service robot has just been bought and is shown around the house. The human-user introduces objects in the house to the robot. The robot understands sentences such as "This is a plant", however it might not understand sentences such as "Please water canteen only once every fortnight and the other plants weekly". That is where my PhD work could complete this robotic system.

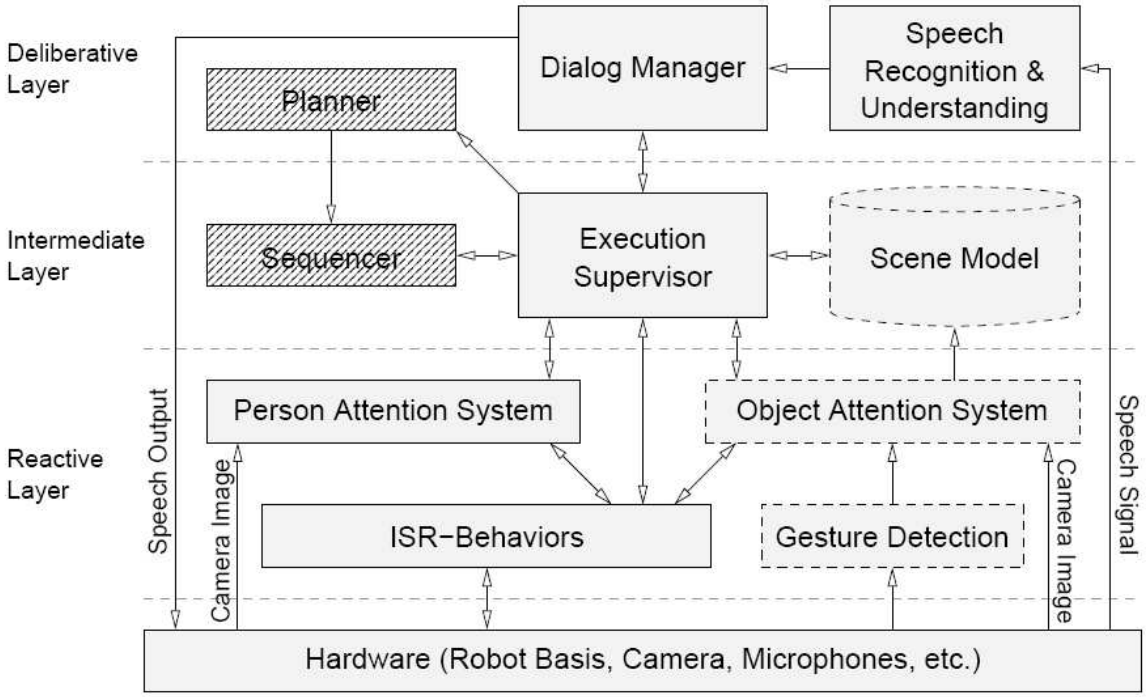

Fig 4: horizontal layers in the hierarchy ensure that low level behaviour (reactive layer) continues to operate while high level plans are executed (Intermediate Layer). Speech recognition is based in the deliberative layer, since recognized sentences contain high level spoken instructions that command the robot.

For further reading, there is another project by Bielefeld University (Steil et. al 2004) about a robot called GRAVIS. The project concentrates on gesture recognition and learning of grasping of objects. The dialogue system is based on an investigation of a corpus of human-human and simulated human-machine dialogs. Language and gesture integration is achieved with a Bayesian network. In contrast, in the MIBL project we try to avoid probabilistic approaches as much as possible, if a semantic approach can give a near to 100% correct integration of speech and gesture (see Wolf and Bugmann 2006).

There are a variety of other projects in the field of multi-modal robots which are not reviewed in this report, but of some importance to this PhD work. see (Perzanowski et al 2001) , (Soshi Iba et al 2002), (Jun et al 2005).

2.2 Natural Language Processing and Natural Language Understanding Systems

Literature in the areas of natural language processing and natural language understanding has been reviewed. Mayor historical works ELIZA Weizenbaum, J.(1966,1976), SHRDLU Winograd, T., (1971), MARGIE Schank, R. and Abelson, R., (1977), and current work such as Mann, G., (1996), Bos J. (2002), Bugmann et al (2004)was considered.

ELIZA is a natural language processing system that enables a user to communicate with it via a console (Weizenbaum, Joseph.(1966)). ELIZA posses as a rogerian psychotherapist. A rogerian psychotherapist is very passive and understanding and let the patient talk about his/her problems. This idea of empathic understanding has psychological healing power according to Rogers.

Here is why Weizenbaum decided to make ELIZA a psychotherapist, when he was confronted with the question : "And what was it that motivated this Rogerian guise?"
Weizenbaum answered:

> "From the purely technical programming point of view then, the psychiatric interview form of an ELIZA script has the advantage that it eliminates the need of storing *explicit* information about the real world."

This statement tells us that Weizenbaum recognized that "real world knowledge" i.e. semantic processing using a knowledge base is a difficult thing to implement. The program ELIZA demonstrates also that even it has no "grounded" language it can pose intelligent by replying to the user with sentences that refer to what the user said. For example if the user would say "I'M DEPRESSED." , ELIZA is programmed to answer "I AM SORRY TO HEAR YOU ARE DEPRESSED" because it was programmed to do so by a simple statement along the lines of:

IF sentence has Subject="I" AND Verb="am" AND object="depressed"
THEN Answer="I AM SORRY TO HEAR YOU ARE DEPRESSED"

Even the program only responds to key-words the users are under the impression to be understood by ELIZA. As the example above shows, however, there is no attempt to connect the rule sets to infer new knowledge or even to ground it to the physical world. ELIZA became a very popular program, since it was one of the first attempts to imitate humanlike communication.

SHRDLU is a program written by Terry Winograd in 1968. It is able to understand natural language text input. He showed by doing this implementation, that if language is confined to a domain ("a micro world"), the computer is able to understand and act upon user requests. The micro world he chosen is a table with blocks, cubes, pyramids and a box. These objects have colours and sizes assigned to them. This project has become quite famous in A.I. under the name "Blocks World" as an idiom for simplifying a problem by restricting the complexity of the environment. It has a vocabulary of around 200 words.

Winograd recognized that syntactics, semantics and logical inference are inseparable in his PhD thesis Winograd, T., (1971). He represents knowledge as procedures, rather than as declarative statements. A procedure can make use of:

- grammar
- semantics
- deductive logic
- other procedures

As the system parses a sentence it will make use of the grammar procedures which can also call semantic interpretation procedures during the parsing process. This is a flexible and powerful method of language parsing.

This increases the flexibility of his representations, since a procedure can call and combine with any other procedures. This is the reason why Winograd has chosen to implement SHRDLU in Lisp. Lisp has the capability to ignore the difference between procedures and data.

The grammar used in SHRDLU is a form of context sensitive grammar called systemic grammar. Systemic grammar helps to organize the correlation between features of natural language constituents and their semantics. This is important for understanding systems, and this was probably the reason why Terry Winograd has chosen systemic grammar. Winograd recognized that context free grammars are over-generative. The grammar rules are written in "PROGRAMMAR", a general parsing system which compiles the grammar to Lisp code. Winograd admits that it was not practical to implement the whole of systemic grammar, and that the resulting grammar is more "practical". It should be noted that the implemented grammar is not a complete valid grammar for English language. And it is definitely not a standard English grammar. However, it enables the extraction of the semantics of most sentences in order to build a NLU system.

In the late seventies and eighties Roger Schank developed several natural language understanding systems. Schank was working with a group of scientists (Cullingford, Rieger, Goldman, Abelson, Riesbeck, Lehnert and others) perusing the same basic ideas : creating a methodology that leads towards the eventual computer understanding of natural language.

MARGIE was one of the first parsers that created conceptual representations directly from the input text without doing an intermediate syntactic description of the sentence.

SAM (*Script Applier Mechanism*) is a natural language understanding program in the domain of stories. It is a successor of MARGIE (Schank and Abelson 1977). SAM was created by Richard Cullingford and Riesbeck in 1975.

The programs for natural language understanding (NLU) developed by them make use of *conceptual dependency theory*. However, the inventor of *conceptual dependency theory* John Sowa argues that the implementations that Schank's research group used, does not explore the full potential of conceptual dependency (Mann 1995). For example, a word is fixed to a single meaning (word-sense) where a word could have multiple meanings.

Schank goes into great detail of what "understanding" means. To clarify the level of understanding, systems build upon his theory have, a summary is given below.
The system is able to:

- create a linked causal chain of conceptualizations that represent what took place in a story (a paragraph of written text).
- make inferences from the created concepts
- turn created concepts back into text in any language. (paraphrasing)

Since the programs use background knowledge the following is possible with the systems:

- Inferences can be made which are now specifically mentioned from the given text.

In order to encode background knowledge of a particular context, Schank invented the idea of using "*scripts*". A *script* is a structure that describes appropriate sequences of events.

Scripts are used if a situation has a stereotyped sequence of action. Stereotype sequences are situations that are well series of events. For instance in the context of a customer going shopping the following *script* could be used:

Shopping Script

Entry conditions:
- customer needs an item
- customer has money

Results:
- customer has item
- customer has less money

Script Header

Script scenes/events:
- entering the shop
- looking for the item
- taking the item or requesting the item
- paying for the item
- leaving the shop

Fig 5: example of a script

Script items are first hypotheses of events that are going to happen in a particular situation.
The events are in an order, one event happens after another. Schank calls this a "*causal chain*". As the natural language text is processed script events are instantiated with values. A kind of slot filling. If an event happens it can enable the occurrence of another event. In the example above: If the customer has taken an item off the self then the event "paying for the item" is enabled. Since a customer in a shop can only pay if there are items he/she wishes to pay for.

Unfortunately scripts only work for stereotypical situations; therefore they are by no means the answer to how to understand natural language text. Like the title of Schank's book says ("Scripts Plans Goals and Understanding" (Schank and Abelson 1977)), there three theoretical entities necessary, namely Scripts, Plans and Goals to understand natural language.

Plans & Goals
It there is no script available, there needs to be a method of understanding a text. The first thing to do then is to identify the main "*goal***"** of the entities in the text. Suppose the text starts with "John is hungry" then the goal of John is to find food. There might be several sub-goals that are identified during the processing of the text, such as going to a location where food can be found.
If a *goal* can be identified then the computer is able to:
- make prediction what might happen
- build up a script on how to achieve the goal by following the text
- put the text and word meanings in the right context (not specifically mentioned in Schank's book)

To deal with situations, that are not available as scripts, mechanisms (conceptualisations) that underlie the normal scripts must be accessed. Any conceptualizations that are instantiated must be placed so that it is possible to trace a path between them. The path is called a "*plan*". Although Schank's scripts, plans and goals idea lacks flexibility, it may be the most practical approach if a service robot is confined to a limited set of skills. Especially if a practical/commercial service robot with natural language interface would be build at present or in the near future it would use a script based learning approach. Therefore his work is considered in my PhD work.

2.3 Semantic Representation Theories

Conceptual Graphs (CG) are related to *semantic networks*. They have been invented by John F. Sowa. *Conceptual Graphs* can represent concepts and their relationships. They are a powerful tool to create a knowledge base. CGs have the following useful properties:

- they are human readable (hence they can be turned into natural language expressions)
- they can be created from natural language expressions
- they can turned into predicate logic statements (with certain constrains)

Conceptual graphs are best explained on an example. Below an example of the sentence: "John is going to Boston by bus" taken from (Sowa J.F. website)

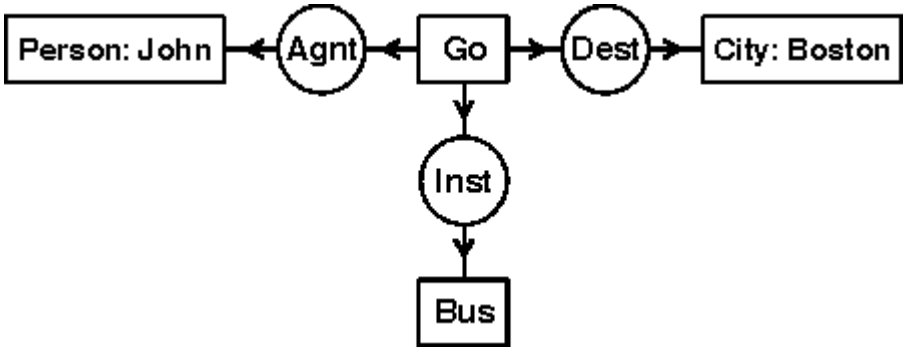

Figure 6: The square boxes indicate concepts and the circles indicate relations. Note that the concept Person has a referent "John" while the instantiation (referent) of the Bus is unknown.

If graphics are not available a "linear representation of the graph can be written as:

```
[Go]-
   (Agnt)->[Person: John]
   (Dest)->[City: Boston]
   (Inst)->[Bus].
```

The following statements show how to read the graph.
A person John *is* an Agent *of* go.
Go *has* a Destination *which is* a City Boston.
Bus *is* a Instrument *of* go.

A concept always has a Type and can have a Referent. A referent is a particular object/concept. The Type must be based on ontology. (Ontology is a tree of types starting with the most general type at the top). A concept can either stand alone or be connected to a relation. It is not allowed to connect two relations directly with each other.

```
[Type: Referent] <-(Relation)-> [Type: Referent]
```

A single concept may be: `[Bus]` Which means “There is a bus”.

```
[Proposition:
  [Woman: *x]->(Attr)->[Beautiful]
]
```

"There exists a woman x who is beautiful.”

Language can be mapped into a conceptual representation using a conceptual parser. The conceptual representation is a representation of the dependency of the parsed text. In this PhD project, conceptual graphs have been a useful representation to clarify the structure of sentences and to extract an ontology design of the domain. This clarification enables the system designer to map sentences into logic.

The research field of knowledge representation has an ongoing argument between non-symbolic (computational) intelligence and symbolic intelligence. Non-symbolic approaches to robotics systems have attracted a large amount of attention in the last two decades. Non-symbolic approaches are very useful in low-level sensor/actuator control. However human language is inherently symbolic and refers to concepts like “the table” and not X=102, Y=314.3. Considering the near future I suggest the most successful system have a mixture of non-symbolic AI on the sense/act level and symbolic AI on the reasoning level [REF Deb Roy: Grounded Situation Models for Robots: Bridging language, Perception, and Action]. The key of such a combined system is a categorizer that maps numerical information into symbolic information. The hybrid system has several advantages.

- decision making is transparent to developers
- language can be integrated on a symbolic level
- symbols can be properly grounded through non-symbolic sensor/actuator information

In an effort to create such a non-symbolic (computational) system that can make the connection to symbols, Deb Roy from MIT created a framework Roy, D. (2005), which is described here. The framework for semiotic schemas is built upon creating a meaning from sensor data and motor acts. It is therefore a so called bottom-up approach to machine learning systems. Every piece of knowledge stored in the robots “brain” can be referred back to the physical world through sensor data and motor acts. It is a grounded system. The knowledge can also be used to make predictions about the future and compare these to actual sensations.

The theory builds on the concepts of

- Signs
- Beliefs
- Projections
- Schemas

Every of these will be explained in more detail.
A *natural sign* is a continuously varying variable, usually an input to the system. A *sign* could be, for example, the width of an object seen by a robot with a vision system. Suppose a sign is monitored over a period of time, a statistical distribution of the input can be formed.

This is called an *analog belief*. So sensors are mapped to *analog beliefs*. See the notation below in figure 7.

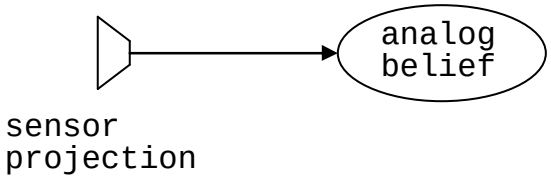

Fig 7: a sensor (natural sign) is montitored to create an "average" belief state

One may wonder how this "analogue" statistical distributions can be put into categories. Deb Roy introduces categorizers as a link between analog beliefs and discrete *categorical beliefs*. In the graphical notation analog beliefs are oval and categorical beliefs are rectangluar.

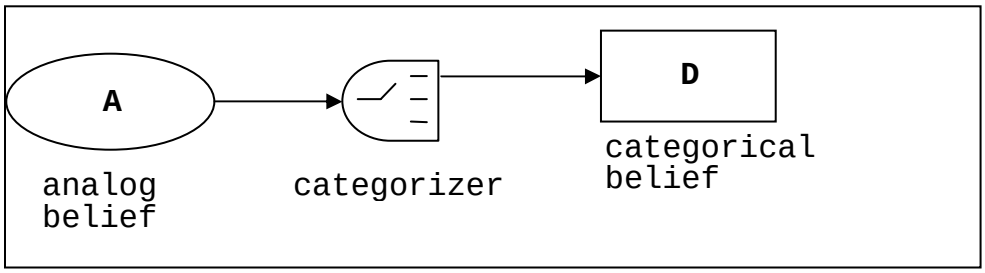

Fig 8: a categorizer makes discrete decisions based on an analog belief

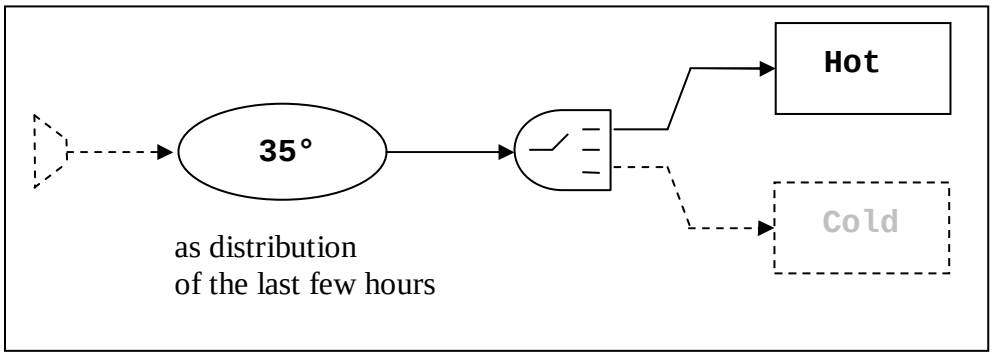

Fig 9: a robot that can feel temperature and belief if it is hot or cold

So far sensor inputs have been measured and categorized, but a robot also has actuators. To drive these *action projections* are introduced. Actions are associated with a lowest-level action primitive function, for example driving a robot joint to a certain angle. An action can either succeed or fail. Here the graphical representation.

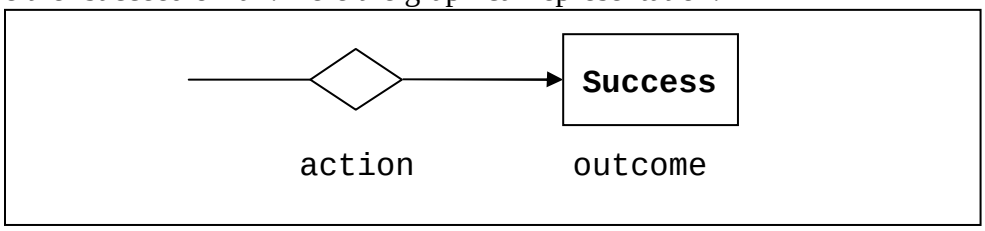

Fig 10: action projection with successful outcome

Actuators can have feedback built in for active sensing. A hand-actuator for instance, can have force sensors to feel how firm the grip is. Sensing while an action is performed is called *active perception* with the following notation, Fig 11:

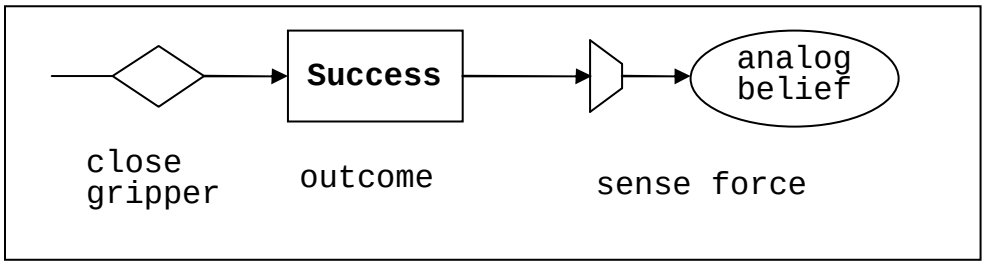

Fig 11: active perception

Deb Roy implemented this framework into a robot called Ripley, which has a 7 degrees-of-freedom arm, vision system and a speech interface. The robot was designed for grounded language experiments (Roy et.al. 2004). The framework is an attempt to connect the symbolic world of language to the non-symbolic world of sensors and actuators. Dr Roy argues that not-grounded systems would need a human in the loop (during design and implementation) to connect sensor data to a representation system in the robot, whereas his approach enables statistical mapping between the sensors/actuators and the introduced symbols.

The frame work of semiotic schemas is used as an inspiration to this PhD work. The most relevant concept for our work is that physically grounded analogue data can be converted to symbolic categories. This allows symbolic processing and the integration of language into the robots reasoning system, even though the robots low level AI is sub-symbolic.

2.4 Conclusion of the Literature Review

Every reviewed project in this report and some more reviewed projects, which are not described in depth here (Spiliotopoulos D et. Al 2001)(Lopes et. al. 2003), have been an inspiration to my work. Almost all reviewed papers mention that future service robots must have natural language capabilities. It is also becomes clear from reviewing that many projects that set out to build a service robot concentrate either on sensor actuators or on reasoning and language. There has not been a service robot project yet that has implemented all of these areas into a single robot to a satisfactory degree. It shows how difficult it is to build a competent service robot. It usually takes several years to develop a basic mobile service robot platform with actuators. To prevent such a long initial development phase it was decided to use the touch screen interface on a table instead.

3 Method

3.1 Corpus centred approach

Previous work carried out by our laboratory on the Instruction Based Learning project (IBL) (Kyriacou, 2004; Bugmann *et. al.* 2004) has shown that it is possible to extract information from a representative sample of the teacher's utterances (the "corpus") in order to:

- Identify primitive procedures that the robot has to be able to carry out (the robot's "prior knowledge")
- Write and tune speech-recognition software to call and combine these primitive procedures.

This approach to the definition of the robot's functionality and natural-language interface (NLI) has been described as "corpus-based robotics" (Bugmann *et. al.* 2004) and is outlined in figure 12.

Figure 12 compares Robot and Corpus-Centred Natural Language Interface (NLI) design. In the Corpus-centred approach, the content of samples of instructions between humans defines at the same time the vocabulary to be dealt with by the speech interface and the required functionality of the robot. In the robot-centred approach, the functionality is defined first, then the access vocabulary, then the NLI. Corpus centred robot design is a completely new idea from Guido Bugmann. So far our research group carried out one project (Instruction Based Learning) IBL based on this idea. There is still much to be learned when using this approach. Also there is much room for improvements for this novel approach. These improvements could turn out to be contributions to knowledge. Here are some research questions which we are trying to answer with this work:

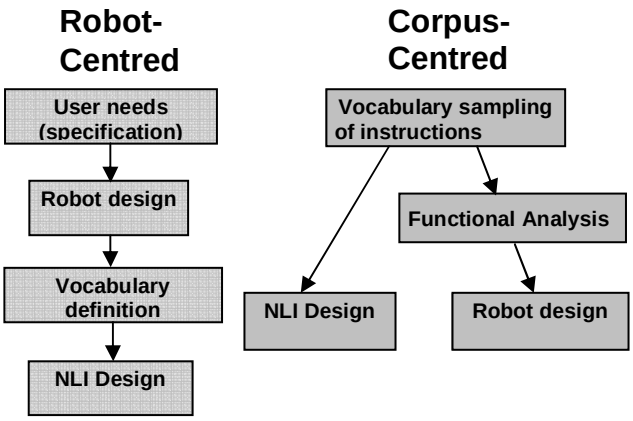

Fig 12: Robot centred vs Corpus centred design

- How can the lengthy process of corpus collection and annotation be made more efficient?
- How can multi-modal corpus collection methods be adopted for the application in the robotics domain?
- How can a corpus be collected before a robot is built?
- Is corpus centred robot design commercially valuable?
- How to extract data (semantics…) from the multi-modal corpus?

The IBL project focused on route instructions given to robots. A dialogue such as the following was possible between the user and the robot:

User: "Go to the University."
Robot: "How do I go there?"
User: "Take the third turning to the left..."
Robot: "Next instruction please."
User: "...take the third exit off the roundabout..."
Robot: "Next instruction please."

User: "The University will be on our right."
Robot: "OK, it's done."

Since the IBL project was using route instructions, the resulting system was developed to deal with sequential instructions, and could not handle other forms of instructions, such as general rules, which apply at any time during the task, such as "Stop at the petrol station if you run low on petrol". The system could not deal with conditionals, such as the one above, that were not found explicitly in the corpus (Lauria *et al*., 2002). In route instructions, sentences starting with "if" instructions are generally just a colloquial way of expressing a sequential instruction, as in the following example from the IBL corpus: "...okay if you carry on straight along this road and if you take the third left you will go over a bridge..."

Therefore, to develop a more general instruction system, there is a need for looking at a different application, where instructions not only include sequences, but also other instruction structures. In imperative programs these would be decisions and repetitions. However, in the declarative paradigm, programs consist of lists of goals and a set of rules (see e.g. PROLOG). It is unclear which paradigm is a more useful representation of human instructions. This is one of the main questions that need to be addressed by analysing a new corpus of instructions in a different domain. Currently declarative knowledge is used to reason in the test system, see chapter 3.4.

3.2 Multi-modal interface and corpus collection

As argued in the introduction, a truly natural unconstrained human-robot communication interface must be multi-modal. In future robots, multi-modal interfaces will require complex sensory processing, such as gesture and face recognition. As this project focuses on the problem of learning, we decided to devise a simplified interface that would still allow natural communication with human users.

3.2.1 User interface

Our solution to the problem is to use a touch screen that allows at the same time to acquire human gesture information by the robot (without complex sensory processing) and execution of game moves (without complex actuators). The screen represents the world as the robot would see it through it's vision system. The user is able to point at and manipulate objects on the screen as a demonstration of how to do the task. At the same time the user gives verbal instructions. Touch-screens have been used in multimodal human-robot interfaces for different applications, for example by (Perzanowski *et al.*, 2001), or for investigations in human communication (De Ruiter *et al.*, 2003)

A great advantage of using a screen representing the robot's world is that the robot can be simulated, while the interaction and interface to the robot does not change. It also allows focusing research on human-robot interfaces without having to build a robot first.

The software developed to display the playing cards is based on the Qt (Trolltech®[1] 2005) and the OpenGL® API[2] and is platform independent. Qt is a cross-platform C++ GUI development library. OpenGL® is a standard for a 3D/2D cross-platform Graphics API. The playing cards are described as objects with parameters such as size, texture, position,

[1] *Qt is a trademark of Trolltech in Norway and other countries.*
http://www.trolltech.com/

[2] *OpenGL® and the oval logo are trademarks or registered trademarks of Silicon Graphics, Inc. in the United States and/or other countries worldwide*.

orientation, static or movable. Therefore the system can be used not only for card games. The display software could display any real world object that the robot knows about. The user can manipulate these objects intuitively. The computers used for displaying the cards are linked to a server via TCP/IP.

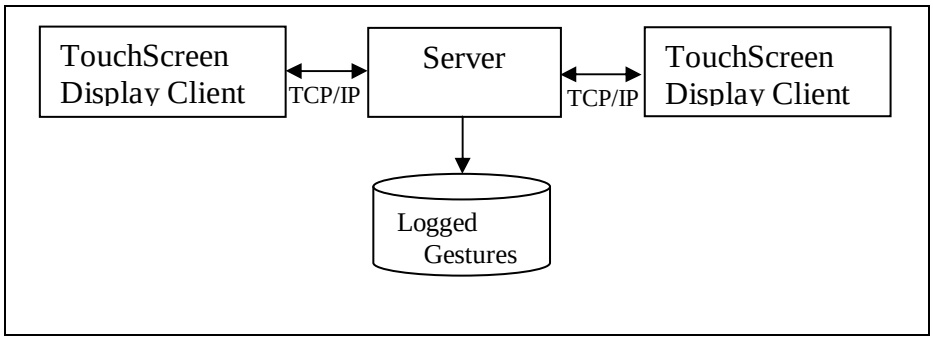

Figure 13: Corpus recording with two touch screens. If an object is manipulated in one screen it is also visible in the other, since the data is forwarded by the server. All data is logged and can be replayed for transcription.

All events of object manipulations are logged at the server and forwarded to all other connected clients. So if objects are moved on one screen, they move on the other screens as well.

3.2.2 Corpus Collection

A teacher and a student sit at a desk (Figure 14). The two are separated by a screen so they can not see each other. The desk has touch screens build into its surface. Playing cards are shown on the screens. The cards can be moved around on the screen by touching and dragging them around. Both players have a common area for cards, and an area that can only be seen by one player (black area on the touch screen in figure 14 symbolizing the cards in the hand. The common area is located near the screen and represents the virtual playing-table.

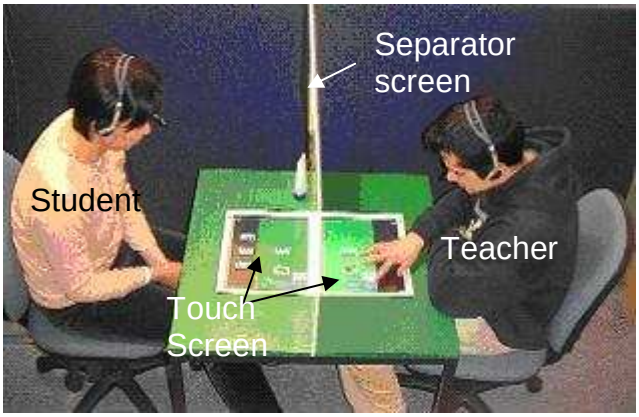

Figure 14: Setup for corpus collection

The teacher explains a card game to the student. The interaction is filmed and the dialogue recorded. To ensure high quality recordings, the subjects wear headset microphones. The coordinates and movements of the cards on the touch screens are recorded simultaneously. The data can be synchronized with a time stamp. This simultaneous recordings of voice and touch screen data from object manipulations constitutes a multi-modal corpus.

We recorded 21 dialogues between teachers and students. Students who learned the game in one session became the teachers in the next (see figure on the right). In the design of the protocol we tried to avoid two forms of bias, the vocabulary bias and the instruction strategy bias.

Pilot studies revealed that a teacher subject tends to use expressions and methods similar to those used when he/she was taught. To avoid this bias, we decided that the initial teacher (Student S0 at the top of the tree) would learn the rules of the game from a set of rules written on separate sheets and presented in random order. Subjects usually proceed by re-ordering the sheets to help learning the game. Two sets were used with different words for the same rules.

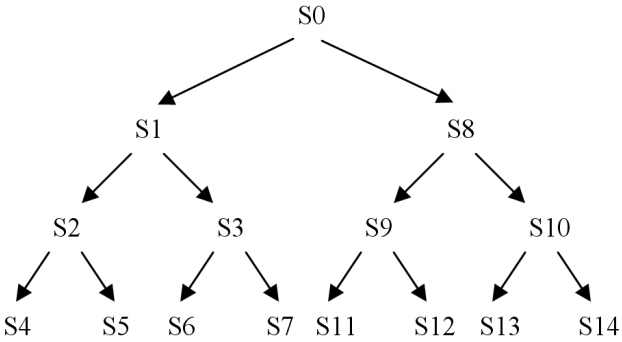

Figure 15: Tree of teaching dialogues. Two trees of this type were used to record dialogues. There are 14 dialogues in each tree, represented by the arrows and organized in three layers. Si is the subject number i.

In order to maintain the chain if a subject decided to drop out, the experiment was designed in a tree structure where one teacher teaches two students, and then these students become teachers themselves. Figure 3 shows one of two trees used in this experiment. Example: Teacher S0 teaches student S1 and S8. After that S1 becomes a teacher and teaches S2 and S3.

We left at least one day between learning the game and having to teach it. This generally led to a fading of the memory of the precise words and order of instructions. Thus the chain design is expected to reduce the bias in vocabulary and lead to an increased variety of instruction styles in the corpus.

3.2.3 Transcription Tool

For the transcription the server module is replaced with the transcription tool MuTra (**Mu**lti-modal **Tra**nscription), which was written for this purpose. It can replay mono and stereo WAV and NIST/Sphere sound files in 8 bit and 16 bit format as well as the object manipulations on the screen.

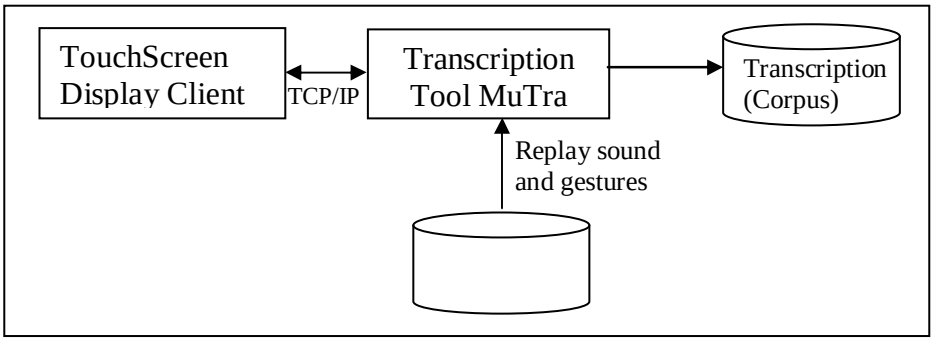

Fig 16: MuTra Transcription tool allows sound and gestures to be replayed in order to transcribe the multi-modal data as XML.

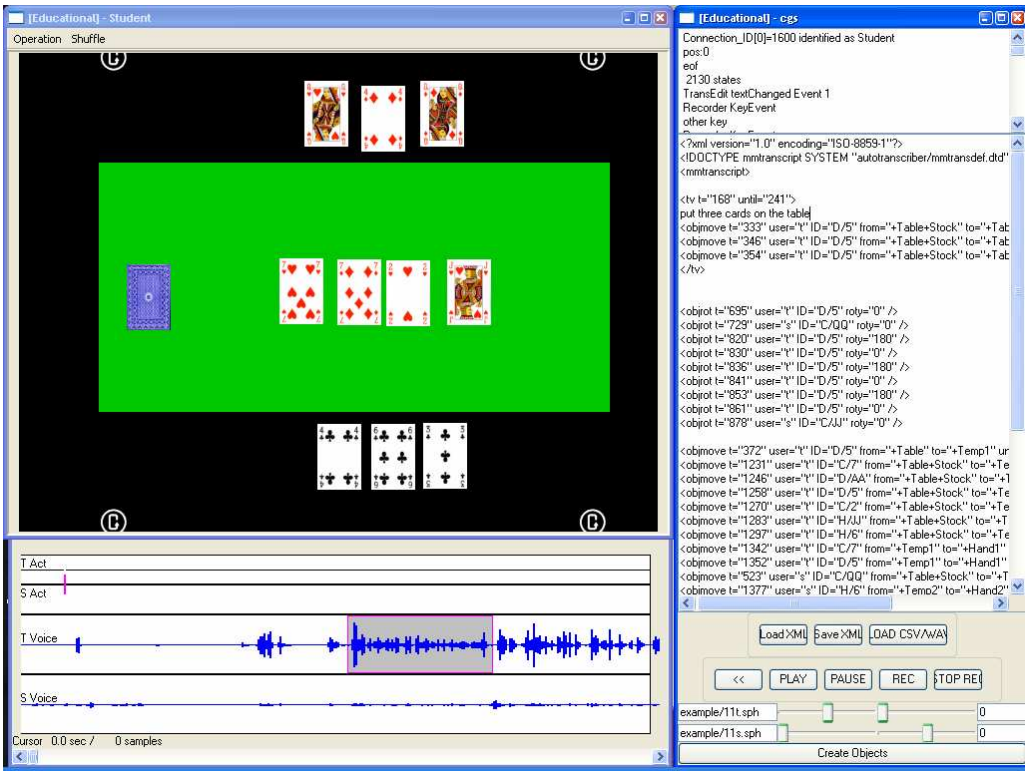

Fig 17: Screen shot of the Multi-modal transcription tool MuTra

The screen shot shows the touch screen client (top left) and the MuTra software (other two windows). The user can mark an utterance in the wave-display (bottom left). MuTra automatically creates an XML-tag with the corresponding timestamp on the right. Now the user has the opportunity to transcribe (write down) the utterance as text. The created XML file can now be used with parsers to do corpus analysis. The XML structure allows extracting information such as timing, dealing or game phase, relevant grouping of gestures that go with the utterance and so on. The above described OpenGL/Qt client software display cards in this scenario. However it can be used to display any 3D object on the screen that is relevant to the task explanation. The transcription tool has been made available as open source under the web address http://www.swrtec.de/swrtec/mibl/mutra/. The tool has been used for teaching at the university as well.

3.3 Corpus Analysis and knowledge representation

Corpus Analysis is mainly concerned with the investigation of the structure of data and their correlation to semantics. Corpus analysis can be carried out for each modality. In the case of gesture data, one can categorize gestures into groups such as movements of objects, grasping of objects, pointing at objects and so on. Corpus analysis also provides information the timing of these gestures. Timing plays a crucial role when combining the modalities. Combining (unification) of modalities is required since the information of the modalities complete each other. For example sentences like “turn over this card” or “This is a plant” are examples that refer to specific instances of physical objects which can only be identified with the accompanied gesture.

3.3.1 Grammar design

Corpus analysis also provides all information required to write a grammar and tune speech-recognition software. Currently a statistical language model has been trained with the corpus using NUANCE 8.5, a user independent speech recognition system.

Analysing the utterances of the transcriptions reveals primitive procedures that the robot has to be able to carry out (the robot's "prior knowledge"). Such "language primitives" are specific to the level at which humans communicate with each other. They can constitute complex robot procedures which may require the use of a micro planer.

3.3.2 Semantic Analysis

The following language primitive have been identified in the dealing phase:

```
start_of_sequence(name)
end_of_sequence()
deal(objects,amount,target)
move(objects,amount,source,target)
turn(objects)
owner(objects,player)
visible(objects,player)
count(objects,amount)
```

Many of these primitives can only be completely specified and resolved using a combination of speech and gesture information.

3.3.3 Speech and Gesture timing

A detailed analysis was carried out measuring the timing between gesture and speech of the teacher. The MIBL corpus showed that verbal instructions are always in the same order as the corresponding gestures. By generating a histogram that shows time-spans between start of an utterance and the corresponding gestures, filter rules can be established. More filter rules can be established by looking at the end-of-speech timing and end-of-gesture timing. Applying all the filter rules that are only based on timing leaves still a large percentage of ambiguous cases, where it is not clear to which utterance and gesture belongs together. This is due to the nature of unconstrained flow of speech and gesture, where utterances the pause of two utterances can be as little as few tenth of a second. This is evidence that gesture and language must be group using semantics as well. In the case of card games the primitive and card origin and destination usually give enough clues to disambiguate the remaining cases, where timing is not sufficient. After the gesture and utterance have been grouped they can be combined (unified).

3.3.4 Reasoning

Work is currently underway to develop first-order predicate logic statements that carry out the unification, although temporal logic could be considered as well. A Prolog rule that compares the parameters of the language primitive to the parameters of the gestures is at the core of the mechanism.

The following 4 cases can occur as a result of pairing:

Completion	A gesture and an utterance are individually incomplete, but complete each other. $n_s = 1$, all variables are resolved
Confirmation	A gesture and an utterance are individually complete. When combining they match. $n_s = 1$, no variables exist
Contradiction	The gesture supplies contradicting semantics when compared to the utterance. $n_s = 0$
Under-specification	The gesture and language combined are still semantically underspecified. Therefore several possible candidates are returned. $n_s > 1$

Another reasoning engine that is applied is the micro-planner, which is described in Chapter 3.4. The high level reasoning and planning part has only been investigated very little. It will be based on a problem solver using semantics like in chapter 3.3.2. The goal of the problem solver is to work out the next possible move of cards, which is valid within the learned rules of the game. The dialogue manager is working closely together with this problem solver. The dialogue manager will probably be a state engine, where a change in state is triggered by a set of states which become true. A typical state variable is the robot mode which can either be learning or playing.

3.4 Test System

The test system implements the ideas and that have been developed so far on mapping natural instructions into semantics. Shown in figure 18 is an overview of the test system. Interesting is that (Perzanowski et al 1998, 2000) and the Bielefeld group have produced a related system proposal independently. Comparing to (Wolf and Bugmann 2006) the diagram shown below is now rearranged to fit Brooks layered architecture.

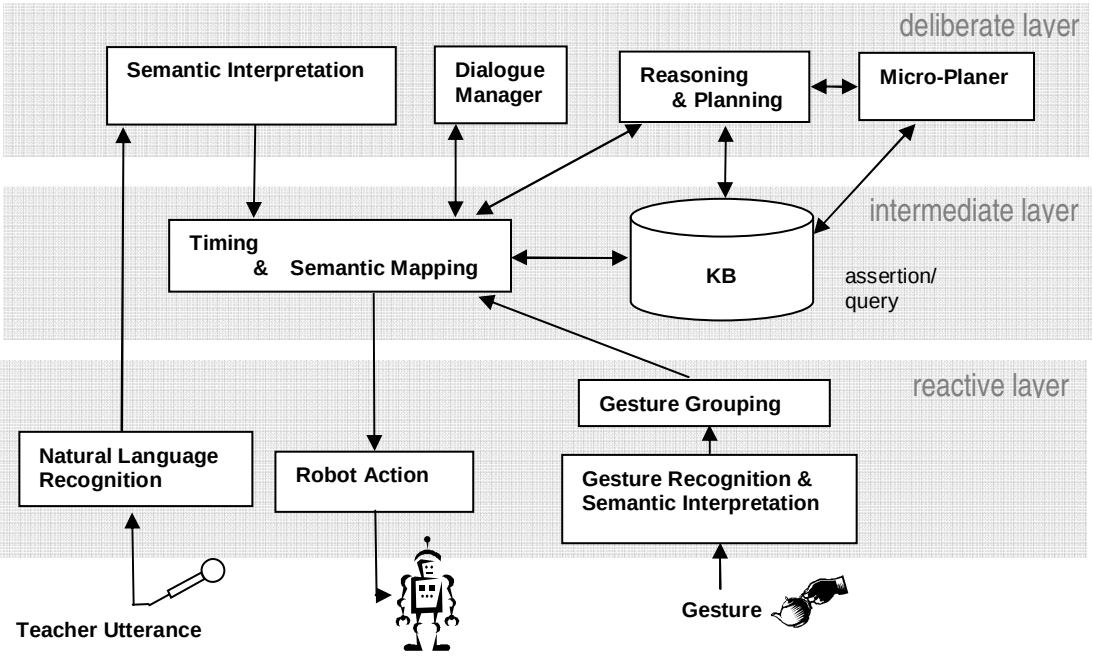

Fig. 18. Overview of the MIBL Robot.

One of the first researchers to recognize the horizontally layered architecture to implement the sense-plan-act theorem was Brooks, R. A. (1986). The major advantage of this architecture is that it is modular. Low level modules (reactive layer) can operate simultaneously with higher levels. Higher levels can overwrite lower levels. The lowest layer is called reactive since it would contain behaviours like simple obstacle avoidance.

We are currently using a statistical language model for the language recognition. A robust interpretation grammar extracts the semantics. A multithreaded application (one for gesture and one for speech recognition) forwards information to the Timing & Semantic Mapping process. The semantics are unified and the micro-planer is consulted.

The micro-planer produces a detailed plan of what the robot should do. Sentences such as “take out all the eights, nines and tens from the deck”, are one primitive to a human, but require a variety of robot-actions to be carried out at the low level (i.e. moves and comparisons). The micro-planer is a problem solver which returns the steps required for the robot to achieve the language-primitive. If a single solution-path is returned, the problem is solved. The path is executed if needed, and stored if the robot is in its learning phase.

The resultant plan can be a robot action or a change in the knowledge base. Robot actions range from moving cards to replying to the user via a text-to-speech processor.

4 Progress

4.1 Overview of progress and planned work

Referring back to the FTR1 form, the aim of the research has not changed. Here is my estimation on how much the project is completed so far.

MPhil Component	
bibliographical review (but continues throughout project)	100%
multi-modal corpus design:	
corpus collection	100 %
corpus transcription	100 %
corpus analysis	40 %
test system	80 %

PhD Component	
results with new subjects	0 %
multi-modal issues	50 %
error correction	0 %

Figure 19: Progress overview with respect to FTR1

For a Gant Chart of the project plan please refer to the Appendix. The collection of a corpus and its transcription is a long winded process and it has taken longer than expected. However during the process of corpus transcription a variety of research tools have been developed that will enable to speed up multi-modal corpus collection in the future.

The micro-planner and the unification of the primitives coming from the modalities gesture and speech have been implemented. The main reasoning engine however is still in the design phase. The micro-planner has been successfully tested.

The corpus analysis concentrated on the dealing phase of the game so far. The next step is to analyse the primitives of the playing phase. This will produce some interesting results for research, since the explanation of the playing phase contains rules and turn taking which are challenges for dialogue design and the reasoning engine.

The project plan (Appendix) shows that the first test system is running now. In the next few month new subjects will be invited to test it. From the experience with the subjects we will be able to measure the performance of the test system as its development progresses. One of the more ambitious ideas it to test the system on another domain (not card games or route instructions), if time permits.

Another way of looking at the progress of the project (except the Gant chart), is to divide the work on the corpus in the phases of a card game (dealing, game and end-of-game). Once all three stages have been implemented into the robot, the robot can learn a card game. The completion of an agent implementation is of course a big mile-stone in the project. The progress so far is marked in grey shaded areas in figure 20. It should be noted that figure below does not show the work in proportion to the time-scale it will take to complete.

	dealing phase	game phase	end-of-game phase
corpus collection and transcription			
grammar development			
identification of functions			
reasoning and action			

Figure 20: Progress with respect to corpus.

4.2 Publications and Feedback

So far I have successfully published 8 publications in the field of Robotics. Four of these publications are on the field of my PhD (See Appendix). This four consist of one journal paper, IEEE conference paper and two more conference papers. The research project has spawn off a significant amount of materials that are worth publishing. The next publication planned is a journal paper describing the test system currently built.

I have attended several seminars in the University of Plymouth that have been given by guests to our research group in robotics. Talking to these visitors and demonstrating my project to them has provided me with valuable feedback. I have also received feedback by e-mail and from conferences. The reviewers of my publications have helped me to see my work from a different perspective.

4.3 Expected Contribution to Knowledge

So far original work has already been carried out in the field of multi-modal corpus collection and multi-modal information fusion. The rule based reasoning engine however is still under investigation and may have significant value.

The following potential contributions have been identified:

- *High level learning and reasoning engine for a service robot:* Most current service robots are only able to learn simple movement or a simple sequence of actions from a human instructor. A system will be proposed that can learn *rules* through speech and gesture demonstration *as well as sequences*.

- *A framework for user programmable robots.* Architecture for the higher levels of a multi-modal robot

- *A detailed method for corpus based robotics*:
  going from an application-scenario, collecting a Multi-modal corpus, creating a grammar and corpus analysis method that can identify primitives and gesture timing data, Implementation pattern of AI routines and finally to a working robot.

- *novel method of corpus collection for Multi-modal corpora* for service robots: The use of a touch screen and not allowing direct visibility between human-to-human gestures provides a novel method of collecting data for free-flowing future human-robot interaction without the need of building the robot first.

- *method of information fusion of gesture and language:* non statistical approach of gesture and language integration based on timing and semantics of a corpus.

Considering the significance of the expected original contributions to knowledge listed above, I apply therefore to transfer this project from MPhil to PhD.

References

Bos Johan (2002), “Compilation of Unification Grammars with Compositional Semantics to Speech Recognition Packages.” COLING 2002, Proceedings of the 19th International Conference on Computational Linguistics, Pages 106-112.

Brooks, R. A. (1986) "A Robust Layered Control System for a Mobile Robot", IEEE Journal of Robotics and Automation, Vol. 2, No. 1, March 1986, pp. 14–23; also MIT AI Memo 864, September 1985. (http://people.csail.mit.edu/brooks/papers/AIM-864.pdf)

Bugmann G., "The What and When of Service Robotics”, Industrial Robot :An International Journal, Emerald , Volume 32 Issue 6 2005, p.437

Bugmann,G., Klein, E., Lauria, S., Bos, J. and Kyriacou T. , (2004) “Corpus-Based Robotics: A Route Instruction Example” in Proceedings of IAS-8, 10-13 March 2004, Amsterdam, pp. 96-103.

De Ruiter, J. ,P. , Rossignol, S., Vuurpijl, L., Cunningham D.W. and Levelt, W.J.M., (2003). SLOT:A research platform for investigating multimodal communication. In Proc. Of Behavior Research Methods, Instruments & Computers 2003, 35(3),408-419

Dillmann, R, Ehrenmann M, Steinhaus P, Rogalla O, Zöllner R, “Human Friendly Programming of Humanoid Robots - The German Collaborative Research Center” Tsukuba Research Center, AIST, Tsukuba, Ibaraki, JAPAN, Dec.11-12, 2002

Engelhardt K.,G., and Edwards R., A., “Human-robot integration for service robotics”, Chapter 16 from Human-Robot Interaction, Taylor & Francis Ltd. , London, 1992.
Libertas Ref# 629.892 HUM

Haasch A., Hohenner S., Hüwel S., Kleinehagenbrock M., Lang S., Toptsis I., Fink G. A. , Fritsch J., Wrede B., and Sagerer G: “BIRON - The Bielefeld Robot Companion” (In E. Prassler, G. Lawitzky, P. Fiorini, and M. Hägele, editors), Proc. Int. Workshop on Advances in Service Robotics, pages 27-32, Stuttgart, Germany, May 2004. Fraunhofer IRB Verlag.

Iba s., Paredis C., and Khosla P., “Interactive Multi-Modal Robot Programming”,
International Journal of Robotics Research, Vol. 24, No. 1, January, 2005, pp. 83-104.
also came out in Proceedings of the 2002 IEEE International Conference on Robotics and Automation, Washington D.C., May 11-15, 2002

Kyriacou T., (2004), “Vision-Based Urban Navigation Procedures for verbally instructed robots.”, PhD Thesis, University of Plymouth, U.K.

Lauria, S., Kyriacou, T., Bugmann, G., Bos, J. and Klein, E. (2002). Converting Natural Language Route Instructions into Robot Executable Procedures. In Proc. of the 2002 IEEE International Workshop on Robot and Human Interactive Communication (Roman'02), Berlin, Germany, pp. 223-228.

Lopes, L.S.; Teixeira, A.J.S.; Rodrigues, M.; Gomes, D.; Girao, J.; Teixiera, C.; Senica, N.; Ferreira, L.; Soares, P.; (2003) “A robot with natural interaction capabilities” Emerging

Technologies and Factory Automation, 2003. Proceedings. ETFA '03. IEEE Conference, Volume 1, 16-19 Sept. 2003 Page(s):605 - 612 vol.1

Mann, G., (1996), “Control of a Navigating Rational Agent by Natural Language”, PhD thesis, Department of Artificial Intelligence, School of Computer Science & Engineering, The University of New South Wales, Sydney, Australia

Mellish, C.S., (1985), “Computer interpretation of natural Language descriptions”, Ellis Horwood Limited, Chichester, ISBN 0-85312-828-6

Miura J , Iwase K., and Shirai. Y. "Interactive Teaching of a Mobile Robot," *Proc. 2005 IEEE Int. Conf. on Robotics and Automation*, pp. 3389-3394, Barcelona, Spain, April 2005

Parlett, D., (2004), “the Oxford A-Z of Card Games”, Oxford University Press, Second Ed.

Perzanowski, D., Shultz A.C. and Adams W., “Integrating Natural Language and Gesture in a Robotics Domain” *in Proceedings of the IEEE International Symposium on Intelligent Control: ISIC/CIRA/ISAS Joint Conference, Gaithersburg, MD: National Institute of Standards and Technology*, 247-252, September 1998.

Perzanowski, D., Adams W., Shultz A.C. and Elaine Marsh “Towards Seamless Integration in a Multi-modal Interface.” *in Proceedings of the Workshop on Interactive Robotics and Entertainment*, Carnegie Mellon University: AAAI Press, 3-9, April 2000.

Perzanowski, D., Schultz, A., C., Adams W., Marsh, E. and Bugajska, M., (2001), "Building a Multimodal Human-Robot Interface", IEEE Intelligent Systems, 16 (1), IEEE Computer Society, 16-21.

Riesbeck, C.K., (1986), “From Conceptual Analyzer to Direct Memory Access Parsing:An Overview”, Advances in Cognitive Science, Chapter 8 (http://www.cogsci.northwestern.edu/courses/cg207/Readings/Riesbeck_From_CA_to_DMAP.pdf)

Roy, D. (2005), “Semiotic Schemas: A framework for Grounding Language in Action and Perception”, Elsevier, Artificial Intelligence Journal, Volume 167, Issues 1-2, Pages 170-205 (http://web.media.mit.edu/~dkroy/papers/pdf/aij_current.pdf)

Roy,D., Hsiao, K., Mavridis, N., (2004), ”Mental imagery for a conversational robot”, IEEE Transactions of Systems,Man and Cybernetics, Part B, 34(3):1374-1383, 2004

Schank, R. and Abelson, R., (1977), “Scripts Plans Goals and Understanding”, Book published by Lawrence Erlbaum Associates Inc, New Jersey, U.S.A., ISBN 0-470-99033-3

Sowa J.F.(2005) , website “http://www.jfsowa.com/cg/cgexamp.htm”, last update March 2005.

Spiliotopoulos D., Androutsopoulos I. and Spyropoulos C.D., "Human-Robot Interaction Based on Spoken Natural Language Dialogue". Presented at the European Workshop on Service and Humanoid Robots (Servicerob 2001), Santorini, Greece, 2001.

Steil J.J., Röthling F., Haschke R. , Ritter H. (2004),"Situated robot learning for multi-modal instruction and imitation of grasping" *Robotics and Autonomous Systems*, Special Issue on "Robot Learning by Demonstration", (47), 129-141, 2004

Weizenbaum, Joseph.(1966) "ELIZA - A Computer Program for the Study of Natural Language Communication between Man and Machine," Communications of the Association for Computing Machinery 9: 36-45.

Weizenbaum, Joseph.(1976) "Computer power and human reason." San Francisco, CA: W.H. Freeman

Wermter S., Elshaw M., Weber C., Panchev C., Erwin H. Towards Integrating Learning by Demonstration and Learning by Instruction in a Multimodal Robotics. *Proceedings of the IROS-2003 Workshop on Robot Learning by Demonstration*, pp. 72-79, October 2003.

Winograd, T., (1971), "Procedures as a Representation for Data in a Computer Program for Understanding Natural Language", published in three forms:
MIT AI Technical Report 235, February 1971 (http://hdl.handle.net/1721.1/7095)
Journal, Cognitive Psychology Vol. 3 No 1, 1972
Understanding Natural Language (Academic Press, 1972).

Wolf J.C. and Bugmann G. Multimodal Corpus Collection for the Design of User-Programmable Robots. Proc. Taros 2005, London, pp. 251-255.
MuTra link: (http://www.swrtec.de/swrtec/mibl/mutra/index.php)

Wolf J.C. and Bugmann G., "Linking Speech and Gesture in Multimodal Instruction Systems" in the Proceedings of The 15th IEEE International Symposium on Robot and Human Interactive Communication (Ro-Man), Hatfield, U.K. , 2006

Copyright Note:
Some pictures of this report only have a copyright for personal use. The report can therefore not be published unless the pictures have been taken out or permission from the corresponding sources has been granted.