Robotics and Autonomous Systems 56 (2008) 1004–1013

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

A dialogue approach to learning object descriptions and semantic categories

Hartwig Holzapfel *, Daniel Neubig, Alex Waibel

interAct Research, Interactive Systems Lab, Universität Karlsruhe, Germany

ARTICLE INFO

Article history:
Available online 2 September 2008

Keywords:
Dialogue management
Learning
Knowledge acquisition
Humanoid robots

ABSTRACT

Acquiring new knowledge through interactive learning mechanisms is a key ability for humanoid robots in a natural environment. Such learning mechanisms need to be performed autonomously, and through interaction with the environment or with other agents/humans. In this paper, we describe a dialogue approach and a dynamic object model for learning semantic categories, object descriptions, and new words acquisition for object learning and integration with visual perception for grounding objects in the real world. The presented system has been implemented and evaluated on the humanoid robot Armar III.

Published by Elsevier B.V.

1. Introduction

An important aspect of humanoid robots in a natural environment, is the ability to acquire new knowledge through learning mechanisms, which enhances an artificial system with the ability to adapt to a changing or new environment. In contrast to most learning algorithms applied in machine learning today, which mainly work with offline learning on training samples, such learning mechanisms need to be performed autonomously, and through interaction with the environment or with other agents/humans. Here, dialogue offers an appropriate means.

In this paper, we address learning of unknown objects in dialogue, which enables a robot to acquire information about unknown objects, and store this information in a knowledge base. A typical problem in this domain is that non-trivial information must be communicated, that spoken interaction results in recognition errors, new words occur in speech that cannot be understood by the system. Thus, the dialogue system needs to conduct dialogue strategies for learning in a way that information about the object can successfully be communicated. And it has to cope with new word learning on speech recognition, grammatical and semantic levels to achieve the learning goal.

The scenario for the system is a household environment for the humanoid robot Armar III, in which the robot is confronted with different everyday-life objects. These objects are parts of tasks that the robot performs, e.g. requests from a human to bring a specific object to someone. Some of these objects that the robot encounters are unknown to the robot. In this case, it is important that the robot can acquire information about the object. The robot needs to acquire verbal information, to understand when the object is referenced by the user, or to talk about the object. It needs to create a model of the object's semantics, which describe the type of object, properties of the object and what the object can be used for. In addition, visual information is stored by the object recognizer and linked to the object's ID.

The presented approach addresses these challenges with a dialogue model for acquiring semantic knowledge, and learning new words from speech recognition. Dialogue strategies are suggested and analyzed, to obtain a semantic category of an object including one shot learning and browsing through the ontology, and to obtain property descriptions of objects, potentially with the usage of unknown words. A dynamic object model is presented, with interlinking an objects database, object ontology and recognition resources that can be updated interactively during runtime. The approach is integrated with visual object recognition and learning for grounding objects in the real world.

The remainder of the paper is organized as follows: Section 2 gives an overview over state-of-the art technology and related work. Section 3 describes the system architecture of our approach to interactive learning of objects. Section 4 describes the system's knowledge sources and its ontology. Section 5 describes detection of unknown information and new words acquisition in dialogue. Section 6 presents an algorithm to symbol grounding for assigning a semantic category to an unknown object in dialogue. Section 7 describes experiments and an evaluation conducted with the system. Section 8 gives a conclusion, and an outlook to future work.

2. Related work

Interactive learning for artificial systems has been studied in several systems. However, the number of approaches that allow interactive knowledge acquisition for humanoid robots is still comparably small. The task to interactively acquire knowledge about objects includes different aspects, which are addressed here, with a discussion of related work. One important aspect of

* Corresponding author. Tel.: +49 721 608 4057; fax: +49 721 607 721.
E-mail addresses: hartwig@ira.uka.de (H. Holzapfel), waibel@ira.uka.de (A. Waibel).

0921-8890/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.robot.2008.08.012

interactive learning, is to detect and understand unknown words in speech recognition. A typical approach for learning new words in speech recognition first addresses detection of an unknown word, which is also called an out-of-vocabulary word (OOV). In a second step, the system then acquires pronunciation (phoneme) and spelling (grapheme) representations of the new word, to update the speech recognizer's dictionary. A further step is necessary to update the speech recognizer's language model, which essentially tells the recognizer where the word can occur in speech, and which probabilities are associated with the word. [17] suggests an approach to learn new words in a multimodal scenario, with the integration of written words on a projection screen. Other work purely relies on speech recognition, such as [22,6,26,24] with different approaches. For example [26] uses multiple recognition, and passes on a single speech utterance with a phoneme-based OOV-model in the first step, and successively narrows down the search vocabulary in the second step. Our work uses the approach described in [24,25], which uses so-called Head-Tail models for acoustic modeling of unknown words. A second recognition run is then performed only on utterances where an OOV has been detected, with a broader vocabulary. It has the advantage that it can be integrated with our speech recognition grammar, which also gives information about a possible semantic meaning of the OOV, based on grammatical construction of the utterance.

During the last few years several approaches have been presented for learning of unknown objects. These are for example [20,32,19], whose main focus however is on the visual side, whereby only known words can be used in speech recognition. Learning of semantic meaning is not addressed there. In the same context the work from [10] addresses cross-modal learning of visual categories. Here, spacial reasoning is applied to associate visual categories to different objects in one image for which a description is given by a human tutor. [27] analyze object models and speech segments that correspond to objects in video sequences from TV shows. Further work addresses learning of speech, for example [13,23] with visual grounding, presuming minimal prior knowledge. The main focus is on grounding and semantics in early stage language acquisition of children. Also, the approach of [30] analyzes very early stage language acquisition, with social learning for the robot pet Aibo. Grounding of event descriptions with visual perception is presented by [29]. Early work for learning in dialogue can be found in the work by [4], with the systems FOUL-UP and POLITICS. Recent work for learning semantics also in dialogue has been presented by [8,9] with the spoken dialog system ''ABILITY'', which is capable of learning new words and phrases during interaction with users. After learning, users could use these new words during their future interactions with the system.

[3,18] describe a system, which is able to develop an ontology with interactive means. The user can insert new objects into the ontology by applying different input modalities. The ontology describes an object hierarchy and properties, attributes and actions can be associated with an object. Sensor feedback is used to detect selected features automatically, which can be confirmed by the user.

In contrast to most of the work referenced here, our approach is intended for human–robot interaction in a household environment, and takes an approach of learning in dialogue. Application of the approach to a humanoid robot defines the type and style of the interaction, the scenario and available sensors. In contrast to work from [3,18], who cover many details of the knowledge representation in a training center, our approach is intended to easily acquire information with comparably short dialogues, in a manner which is acceptable for the communication partner. While the presented approach does not intend to build ontology and object model from scratch, it tries to describe new types of objects with known object categories and concepts. Our approach does not attempt to solve problems of visual processing, such as interactive learning of 3D shapes, or suggest new algorithms for unknown objects segmentation. However, the presented dialogue approach integrates such a component, and could also be integrated with other, similar components. Grounding is an important issue in the field of learning new objects. While approaches presented above do grounding connected with language learning, our approach implements a grounding as a submodule in dialogue, however, the main focus is on the construction of dialogues and dialogue strategies, to categorize a new object and learn properties, potentially with new words in speech recognition. In addition, while some of the approaches presented above address only some aspects of object learning, our approach combines the aspects presented above, namely learning new words, learning semantic concepts and properties, visual object recognition for grounding of the objects in the real world, in a dialogue approach.

3. System overview

Our approach for the interactive learning of objects integrates several knowledge sources with the following aspects:

- *semantic information* about the object is acquired in dialogue. Semantic information covers the type of the object and several properties.
- *different descriptions for spoken reference* can be acquired for a new object, which includes introduction of new words
- *visual information* is used for visual recognition and associates internal object representation to perceptions in the real world.

Fig. 1 shows the integration of the different components within the dialogue system. Dialogue management is handled by the Tapas dialogue tools [14]. The central component in this work is the dialogue manager, which handles user input from speech recognition and object recognition results from the robot. Dialogue strategies conducted with the system make use of several knowledge source, which will be introduced in more detail in the following. The dialogue manager uses (semantic) typed feature structures (TFS) for language-independent knowledge representation for input, discourse and state description. On top of the discourse, an abstract state model defines a context for strategy execution, which selects moves from the dialogue manager's action model to interact with the environment. The slot model defines pieces of information that are collected during a dialogue, in terms of a goal-based strategy.

The learning target in the presented work is kitchen and household objects. In the presented system, learning is conducted as an interactive dialogue approach. The dialogue manager processes requests from the user, integrates object recognition hypotheses, and conducts dialogue strategies to learn information, when a learning dialogue has been initiated during human–robot interaction dialogues. Such a 'learning dialogue' is designed to acquire new information for known or unknown objects or to clarify information, each to update the system's knowledge model. Learning covers new semantic categories, new descriptions for existing objects including new words, learning of object properties, and association with visual object IDs. Before a learning dialogue can be initiated, certain triggers are used to determine when an unknown object has been found, e.g. by the object recognition component, or when unknown words occur.

Speech recognition, unknown word detection and new words learning is performed with the Janus speech recognizer and Ibis decoder, presented by [28]. It is integrated with the dialogue manager Tapas in an interactive system for human–robot interaction. An overview of the multimodal perceptual components for the robotic system with further details about additional components can be found in [31].

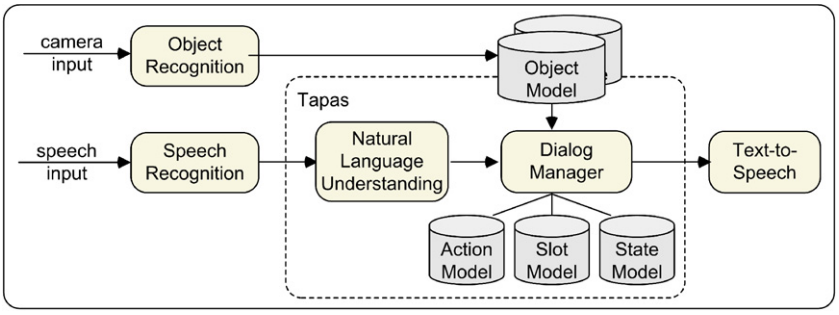

Fig. 1. System overview of the different components in the dialogue system, and data flow of perceptual input with response generation by the dialogue manager.

Experiments reported later in the paper have been conducted on the humanoid robot Armar III, which is described by [1]. Conducting the experiments with the humanoid robot Armar III, leads to a typical human–robot interaction scenario, which defines the type of interaction, and defines the perceptual system for our approach. While from a technical point of view, the humanoid robot is only used as a perceptual system which can go to and look at different places, users reported that interactions with the humanoid robot is fun, and the robot represents a communication partner they can talk to. Using the humanoid robot also serves as a proof of concept that the approach works on the target platform.

Visual processing uses stereo vision from the robotic head's cameras. For visual processing, detection and recognition of objects, we have integrated an object recognizer provided by Azad et al. [2] and the software toolkit IVT.[1] Though visual object recognition is not the main focus of the paper, we want to give a brief description of the recognizer's functionality to the extent that is necessary to follow the experiments. It can recognize textured objects using SIFT features [21], and untextured objects using 3D shape models and color. Because learning of 3D shape models requires complex modeling, and scanning of the object from different angles to observe its structure, this approach is currently not realistically applicable for interactive learning in real-time. Rather, the use of SIFT features allows to learn an object from features extracted from a single image taken from the scene with stereo vision, during the learning dialogue and in real-time. Another advantage of this approach, is that the object's features are mostly independent of scaling, angle of view, rotation, light conditions and their position in the input image.

The object recognizer is able to recognize objects and detect unknown objects in real-time, which is triggered by the dialogue system. For learning of new objects, the object recognizer can store acquired visual features, together with a given label during runtime, such that the object can be recognized immediately after learning. The label is generated by the dialog system and represents an internal 'ID' that is used to identify an object instance. The visual features are automatically segmented from a scene, using stereo vision, depth information and occurrence of visual features. The features for unknown object detection are kept in the memory, until a decision is provided by the dialogue manager to store the unknown object or to discard the features. More details regarding the visual object recognizer can be found in the referenced publication [2].

In the following, first the dialogue manager's knowledge model is described, and afterwards, the dialogue strategies for learning are introduced.

[1] Integrating Vision Toolkit - IVT: http://ivt.sourceforge.net.

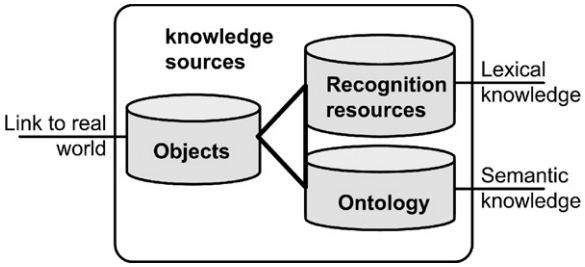

Fig. 2. Knowledge bases which represent lexical knowledge, semantic knowledge and a database of known objects.

4. Object representation and knowledge model

The knowledge bases in our system comprise representational knowledge and interaction knowledge. The representational knowledge bases define the aspects of the knowledge which the system can talk about, and which it can extend by acquiring new information. Interaction knowledge tells the system how to obtain the knowledge via a communication process i.e. the dialogue strategy for acquiring new information.

4.1. Object model

The (representational) knowledge sources and their interrelation are shown in Fig. 2. The figure shows three knowledge models: an objects database, recognition resources and a semantic ontology. Each object is stored as an entry in the objects database and is encoded in part by each of the three knowledge sources. The objects database thus contains all "real world" objects, i.e. instances of all known objects and objects in the environment model. The database is realized as a relational database, and an object is represented by a database entry. An object entry in the database includes an ID (unique label), the semantic category (type), values for object properties, and an association to a list of observed textual descriptions. The ID is the same label that is used by the object recognizer to identify an object, for example 'granini_juice_0001'. The type values associates an object instances with ontological concepts, for example the object 'granini juice' is associated with the concept 'obj_juice'. Values of object properties store additional information about the object instance, such as brand 'granini' or color 'yellow'. Observed textual descriptions could be 'granini juice'.

Type information and semantic categories of objects are modeled in an ontology. The object ontology provides inheritance information (*isA* hierarchy of concepts with multiple inheritance) and defines properties that can be associated to objects. To be able to talk about object types, e.g. refer to the concept 'obj_juice' by using the word 'juice', an additional mapping file is defined which is used for grammar creation in speech recognition and understanding, and for spoken output.

Considering spoken interaction, the recognition resources represent lexical and grammatical information of the objects. The grammar describes how objects are embedded in grammatical constructs, i.e. their lexical representation and how the objects are referenced in speech. In the following example: 'please open the granini juice for me', the term 'granini juice' is a description of an object which is stored in the database. The lexical tokens - here 'granini juice' - are read from the database and dynamically update the grammar at a predefined position defined by semantic categories.

4.2. Language understanding and grammars

The definition of grammars in our system follows the approach and formalism of semantic context free grammars, see [7]. This formalism defines a grammar based on semantic categories, in addition to syntactic information with the formalism of vectorized grammar nodes. A grammar node defines three values in the following way <*sem_concept, syntactic_category, subcategory*>. With this construction, the grammar inherently carries semantic information in its grammatical structure. The grammar's syntax is defined in the Java Speech Grammar Format (JSGF).[2]

The grammar is shared by the speech recognizer, which uses the grammar as a language model, and by the dialogue manager which uses these grammars for natural language understanding and contextual weight adaptation. In the presented approach, parts of the grammar are generated automatically from database and ontological information. Rule generation from database information makes use of semantic categories and rule inheritance, which is defined in the following way. A non-terminal symbol that is defined on the right hand side of a rule, e.g. <*obj_openable,NP,_*>, is automatically extended to its descendants, e.g. <*obj_juice,NP,_*>, if <*obj_openable,NP,_*> is not defined in the rule set. Such inheritance approaches are applied to functional object categories, e.g. openable, portable, eatable, etc. These functional categories are used throughout the grammar to integrate actions / speech acts with objects that are applicable to these actions. For example 'please open the granini juice for me' is covered by a grammar rule that interrelates the speech act *act_open* with an object of type openable. The simplified rule looks as follows:

```
public <act_open,VP,_> =
        <please> <open,V,_> <obj_openable,NP,_>
        <recv_me>;
```

The syntactic categories used in the example are *VP* for verb-phrase, *V* for verb, and *NP* for noun-phrase. Subcategories are not used here, but are used in the grammar, e.g. for singular and plural rules or contextual utterances.

The actual generation of grammar rules from database information is realized with the following approach. So called 'import' statements which are specified as the right-hand side of a grammar rule, define grammar rule generation with database content. The presented grammar generation approach from database information extends previous work on a multimedia access dialogue system [11], by the definition of more complex import statements to match object descriptions, and supporting interactive extension of the models. The left hand side of the rule is a standard non-terminal symbol, e.g. <*obj_juice_db* >, the right hand side of the rule is started with a VOID element, which conforms to the JSGF syntax specification. The import definition includes DB connection, imported fields and semantic conversion rules with the syntax *import DB − ref entry*$_1$ *entry*$_2$ . . . *entry*$_n$. Each entry consists of a table-field pair with an optional list of semantic values in the form of *table field* {*sem_type*$_1$ *sem_value*$_1$ . . . *sem_type*$_k$ *sem_value*$_k$}. For example, the import defined by the entries

objects_juice brand {BRAND objects_juice:brand } objects_juice type {TYPE import }

generates the right-hand the side productions

```
  granini { BRAND granini } juice { TYPE juice}
| valensina { BRAND valensina } juice { TYPE juice}
```

from the given database entries

Type	Flavor	Brand	Onto type
Juice	Apple	Granini	*obj_juice*
Juice	Orange	Granini	*obj_juice*
Juice	Orange	Valensina	*obj_juice*

With these tools, more complex grammar constructs are possible, such as combining two rules <*prp_object_db*><*obj_object_type_db*>. This allows understanding of any known property in combination with any known object type, for example 'red cup', 'blue DVD', but also combinations can be parsed that have not been observed before, such as 'green juice'. The latter example is necessary to understand assignment of yet unobserved properties. If one wants to restrict grammar coverage to only known property-object combinations, the import statement is specified accordingly with more than one imported field, as done in the example above.

As mentioned above, the grammars are shared by the dialogue manager and the speech recognizer. For the purpose of using the grammars as language models, a self-contained standard context free grammar is generated. This is done by the Tapas dialogue tools in a compilation step at system startup. During system runtime, the speech recognizer's grammar and the dialogue system's grammar share the same structure, but are different instances. Automatic updates to the grammar, which result from the learning method, are always modifications to the database rather than to the grammar structure. The learning step updates the database and modifies the runtime objects of speech recognizer and dialogue manager accordingly during runtime, by adding new entries to the corresponding grammar rules. With this approach, the grammar instances of dialogue manager and speech recognizer are always kept synchronized, a prerequisite for tight coupling of both components. The advantage of tight coupling is that speech recognition output already represents a parse-tree, and no additional parsing is necessary, to initiate language understanding, which maps grammar nodes to TFS nodes. Another advantage of tight coupling, is that the dialogue manager can maintain a generic expectation model. For example when the system asks the user to name the color of an object, the expectation model contains ontological concepts that can describe a color, and subsequently the speech recognizer's grammar rules are adapted to better fit the expected input. Since the expectation model contains (among others) speech acts such as *inform_color* and property descriptions such *prp_color* which are mapped to grammar rules <*inform_color, VP,_*>, <*prp_color, A,_*>, <*prp_color, AP,_*>, the presented learning approach does not interfere with this model and works in combination with this approach as well. As it has been shown previously, contextual weighting improves speech recognition accuracy significantly [16], especially for short responses, such as 'yes', 'red', or 'yellow'. It offers a benefit especially for large grammars, e.g. to prevent incorrect recognition of an object type, when a color has been said.

[2] http://java.sun.com/products/java-media/speech/forDevelopers/JSGF/.

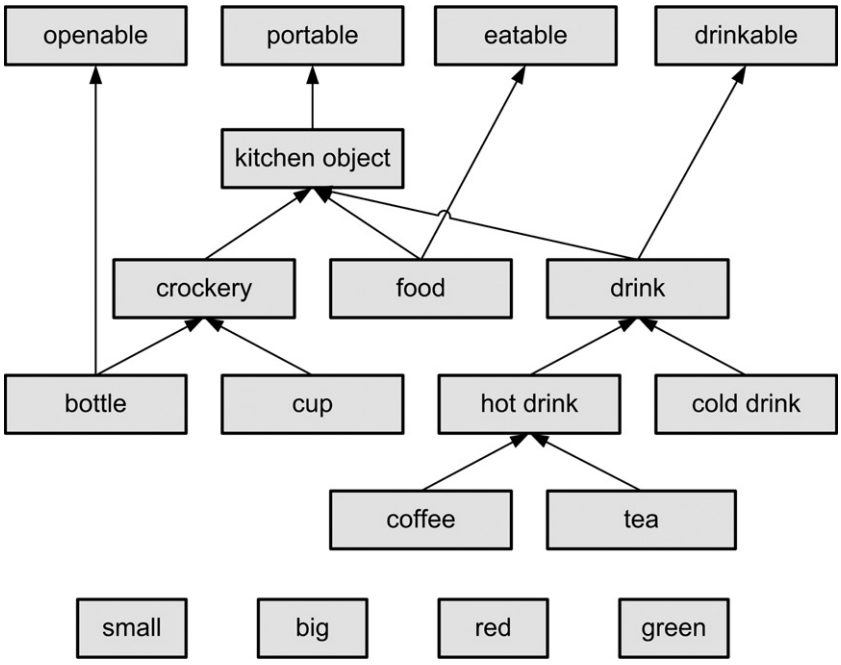

Fig. 3. Ontology organization with functional concepts, type hierarchy and properties.

4.3. Ontology

The robot's knowledge about objects is represented in a specific object model. The model specifies object classes, properties and views of the object. Classes and properties are modeled in an ontology, where a real object can be associated with multiple properties and classes. This allows different attributes to be associated with one object. Examples for properties are color, name, and title of an object.

Fig. 3 shows an excerpt from the system's ontology. Properties are listed at the lower part of the figure. The middle section shows object classes (also referred to as types). The upper section shows functional concepts that model how an object can be used.

The ontology defines object classes hierarchically. General objects are displayed at the top; more specific (inheriting) objects are displayed further down in the ontology. Each object can inherit from one or more functional concepts. Each child of an object inherits the parent's functional classes. This inheritance relation is used in the definition of semantics. As mentioned in the system overview, typed feature structures (TFS) [5] are used to represent semantics in the dialogue system. The definition of TFS allows types from a hierarchy, including multiple inheritance.

For example, an object instance of a kitchen object has the semantic concept 'kitchen object', and all inheriting concepts, such as 'crockery', 'drink' and 'food', are kitchen objects as well. In further inheritance, the concept 'drink' is split into the concepts 'hot drink' and 'cold drink'.

The ontology's functional concepts describe what can be done with an object. For example, all objects which are described in the presented ontology are 'portable'. However, only a bottle is 'openable' whereas coffee is 'drinkable'. These functional classes are used to refer to objects in the semantic grammar for speech recognition and understanding. For example, if the user tells the robot to open something, the concept which is used in the grammar is of type 'openable'. All objects that inherit from 'openable' are automatically inserted into the grammar and can be referenced by the user. The complete list of functional classes used in the experiments covers nine categories: cook, drink, eat, fill, open, play, carry, switch on, and watch.

5. Learning in dialogue and new words acquisition

5.1. Detecting deficient information

A dialogue for learning is initiated by the system during normal interaction, when the system detects deficient information. In the scenario addressed by our system, the goal of most dialogues is to instruct the robot to do a specific task. A typical task-oriented dialogue is conducted when the user instructs the system to bring a specific object, serve something to drink, or put something into the dishwasher. Within such dialogues we have extracted two categories of deficient information.

- the user input cannot be understood correctly by the system given verbal information
- the specified object cannot be found, or an unknown object is detected.

The first case addresses speech recognition and understanding, the second case addresses visual processing of objects in the environment. Both cases can serve as so-called "deficiency detectors".

Deficient information in vision occurs when the object specified by the user cannot be found, or when an unknown object is detected by the system. In either case, the system first needs to detect an unknown object, i.e. obtain visual features for an object which is referred to by the user. If the system does not detect an unknown object, it cannot store any features, and therefore cannot learn information about the object. Thus the detection of features and together with that, segmentation of the object's shape are prerequisites for the learning process. For detection of unknown objects, segmentation and learning of new features, we use the object recognizer described in [2]. In addition to feature detection, the object recognizer uses 3D information for object segmentation. Thus the robot can learn the object when it is held in front of the robot's camera, as shown in Fig. 4 in the leftmost image. The object can also be learned from visual features only, when no 3D segmentation is possible and the background does not have rich texture, as is shown in Fig. 4 in the rightmost image. For the experiments described in this paper, objects were put at a specific location, next to the sink. This way the test subjects did not have to

Fig. 4. Snapshots taken from the robot camera. From left to right: object held in front of the robot's camera, multiple objects recognition, unknown object recognition during the experiment with feature extraction and shape segmentation.

pay attention to where to put the object, so that the robot can find it again and comparable dialogues could be produced. The objects where put on a black surface, with a standard kitchen background, e.g. parts of a cupboard and the sink can be seen in the pictures taken by the robot. In the experiments, the objects' shapes could be segmented reliably from feature clusters only.

Deficient information in speech recognition occurs when the user produces input that cannot correctly be recognized by the system. [12] describes different error situations that occur in human–robot interaction, for which data from text-based interactions and interactions with the real robot have been analyzed. The largest number of miscommunication errors occurs due to new syntactic and semantic concepts, i.e. new formulations, new objects, new goals, and meta-communication. In cases of unknown objects, user input typically leads to sentences that are not covered by the grammar. As described earlier in this paper, the grammar is created automatically from database entries, so that only attributes describing known objects are covered by the grammar. This has the advantage that speech recognition performs well for known utterances, but the disadvantage that new formulations are not covered by the grammar. To prevent this problem, the standard approach in speech recognition would be to extend the vocabulary until all words which have to be covered are contained in the vocabulary. However, in case of object names it is not clear which words need to be covered by the vocabulary in advance, since unpredictable words can occur. In speech recognition evaluation this effect is typically very small, since the standard word-error-rate (WER) is hardly affected, if once in a while, a word cannot be recognized. For the robot in turn, exactly these words can be very important. To show the effect of WER let us consider the example 'please open the granini juice for me' which has been used previously. If the word 'granini' (let this be an unknown word) is misrecognized, the WER is affected in the same way, as if the word 'please' was not understood. However, in the first case, the main information for disambiguating the object in the environment is lost. Extending the vocabulary with a very large number of possible words is not a good option, since speech recognition rates for known objects would drop drastically. However, approaches are known to detect unknown words in speech. We use out-of-vocabulary words (OOVs) which are recognized by the system when an unknown word has been spoken. Our approach uses an implementation of so called Head-Tail models [24] for detection of unknown words. Given an example sentence, which contains the command to switch on an unknown object, the grammar might recognize: 'please open the OOV juice for me'. Here, speech recognition detects an unknown word, which is encoded as OOV. For the detection, both language model scores (defined by the grammar) and acoustic scores (acoustic speech recognition models) are considered. The example sentence also gives us a first hint about the semantic category of the unknown word by observing verb-object subcategorization information, by the semantic frame given through the grammatical construct. Using OOV models has originally been studied for *n*-gram models. In [15] this approach is also described for usage with context free grammar for the recognition of unknown names. The same approach has been adopted for the present system. Following this approach, unknown words can only occur at specific positions in the grammar. The used grammar formalism defines 'oov' symbols in the grammar in the following way. For example a noun phrase describing an object could be

```
public <obj_object,NP,_> =
          oov |
          <obj_juice_db>|
          <prp_juice,A,_> <obj_juice_db>;
```

Here, the oov replaces a full noun phrase. In analogy, the oov can also replace a property, syntactically represented as an adjective or a noun.

5.2. New words learning

Once unknown words have been detected in the utterance, these words can be learned by the system in dialogue. During the experiments, these words are either properties of objects, object types, or part of the object names. In addition to the dialogues to obtain semantic information of the object, which is described in the next section, the system needs to acquire spelling and phonetic information of the word and update the speech recognizer's vocabulary, dictionary and language model. A pronunciation for a new word is generated with a grapheme-to-phoneme converter, which is available with text-to-speech tools, such as Festival or Cepstral. Both a grapheme representation, which is obtained e.g. from spelling, and the phoneme representation are needed to update the speech recognizer's dictionary. In addition to the dictionary, the speech recognizer's language model and the dialogue manager's understanding grammars are updated. The speech recognition grammar is shared by the dialogue manager and the speech recognizer and thus have the same structure. Both can be extended on the fly, and are updated during dialogue, once the new word has been confirmed by the user.

6. Interactive learning of semantic categories for objects

6.1. Learning object properties

The algorithm for learning properties and semantics of an unknown object includes obtaining a description from speech and clarifying properties with their values and semantic types, which is done in a dialogue with the human. The dialogue for learning properties allows the user to formulate any property of the object which he thinks is useful. The system already understands different property values, such as color and size. Other properties, such as title or name (e.g. a DVD has usually been referenced by its title) is restricted to names stored in the database. When the user formulates a description which is not covered by existing property values, the speech recognizer can detect this as unknown words, and reports an OOV detection to the dialogue manager. In the case of OOV detection, the user is asked again to say only

the property of the object, since additional repeats increase the chance of understanding the word correctly. If the word cannot be understood correctly, which is determined by obtaining feedback from the user, the unknown word can also be spelled by the user. The user is only asked for spelling, if the OOV-part of utterance is relatively short (which is determined by phoneme recognition on the utterance). For spoken output, standard grapheme-to-phoneme rules of the text-to-speech synthesis component are used. If the user confirms the word, it is then learned by the system, by adding the word to the speech recognizer's dictionary, and to the speech recognition and understanding grammars. The new word can then immediately be used within the same dialogue. For better understanding of several words which form the title of an object e.g. 'a book on advances in robot control' an additional speech recognition module with *n*-gram language model and a large vocabulary can be used.

A new word learning dialogue is also initiated when the user refers to an object, e.g. 'bring me the red cup' and an OOV is detected for the utterance. In this case the system first needs to find out whether the unknown word is part of the object's type description or if it represents a property. The learning dialogue is the conducted as described above.

6.2. Learning an object's category

Learning of object types is conducted with an approach that combines open input by the user, who can name a category, and a prompted mode which implements browsing through the ontology. In the open input mode, the user can name a category which he would use to classify the object. The open input mode is also referred to as one shot learning, since one input by the user is enough to describe the category. A simple one shot learning dialog follows the example:

System: What type of object is this?	Open question
User: It is a juice	Type is set to juice
System: Did you say that the object is a juice?	Confirmation
User: Yes	Type confirmed

One shot learning has the advantage of quickly obtaining a hypothesis for a category. Drawbacks are that it is not necessarily obvious to the user, how the robot's internal object hierarchy is structured and the user does not know what the system can understand. For example, it was observed that functional categories pose even stronger problems to the one shot learning approach than object types. As a reply to the question 'what can you do with this object' some persons replied with very complex statements, and some had to think for some time before they could come up with an answer. Thus, in the present experiments, open questions are only asked regarding the type of the object, and functional classes can be queried by system initiative only. For example the system can ask 'can you eat this?' or 'is this edible?' when asking for the functional concept 'eatable'. Thus, the dialogue is improved, when the system can choose the wording. The browsing mode addresses exactly this problem, and can choose from questions for object types and functional classes for disambiguation. It starts at a base category and iteratively tries to classify the object as one of the subclasses of the current category. This way, the structure of the ontology can be communicated and input by the user is restricted to a smaller set of possible meanings than in the open input case. Drawbacks of the browse mode are that this mode can be tiring for users, and that for large ontologies, descending the hierarchy can even take too many turns to be practically applicable. An example of the browsing mode is as follows:

System: is the item a kitchen object?	Ask type
User: yes	Type: kitchen_object
System: can you eat this object?	Ask function
User: no	Type: kitchen_object
System: can you drink this object?	Ask function
User: yes	type: Drink
System: is this a hot drink?	Ask type
User: no	type: Drink
System: is this a juice?	Ask type
User: yes	Type: juice

The combined approach begins with a single one shot approach, and then give the opportunity to refine the category be browsing the neighborhood. The dialogue to conduct this strategy begins with a question to specify the class of the object (open input). The input is confirmed. If no children of the class are found in the hierarchy, the dialog ends here. Otherwise the robot switches to the browsing mode until a leaf node has been found in the hierarchy, or no further refinement is given by the user. The questions in browsing mode address children of the selected type or functional concepts to disambiguate subclasses and are formulated as yes/no questions. Fig. 5 depicts this algorithm in a flow diagram. The start-node named "find initial class" represents the one-shot learning node. After the one-shot learning, the learned class can be refined by browsing the ontology's type hierarchy or functional concepts. After posting one question to the user and a confirmation response (bottom node in the flow diagram), the loop is entered again. The combined approach makes sense because of several aspects. (i) Due to speech recognition and understanding problems the desired category cannot be understood. (ii) The user does not know the category description used by the system. (iii) The user communicates a category that is too general, e.g. 'drink'. This general category can then be refined to obtain a better model.

7. Experiments and evaluation

7.1. Experimental setup

For evaluation of the approach, experiments were conducted with the robot in the kitchen environment. The users could, for example, command the robot to bring a specific object from a location, or report which objects he can see at a specific location. The robot knows about several locations from its environment model, such as the sink, sideboard, stove, cupboard, fridge, etc. The robot can also understand directions such as "next to the sink", "left side of the sideboard", "in the middle of the sideboard", "in the fridge", etc. For identifying a requested object (grounding), the robot can ask for the location, which can be given by speech or using pointing gestures. If multiple objects are found at a location, the robot conducts a simple dialog listing all known objects to clarify which object is unknown. If there is more than one unknown object, the user would have to move the object and present the object to the robot e.g. by holding the object in his hand as shown in Fig. 4 in the leftmost image. For the sake of obtaining comparable dialogues during the presented experiments, the setting was restricted to the sink location, with at most one unknown object and grounding restricted to speech. In case there is an unknown object at the given location, the robot ideally would ask the user to help him to learn the object and identify the object's properties. If an unknown object or unknown words occur during the interaction, learning dialogues are initiated by the system as described in the previous section.

The experiments comprise 52 dialogues which were conducted with six naive users—who haven't interacted with a robot before. The goal of these dialogues was to have the robot serve a

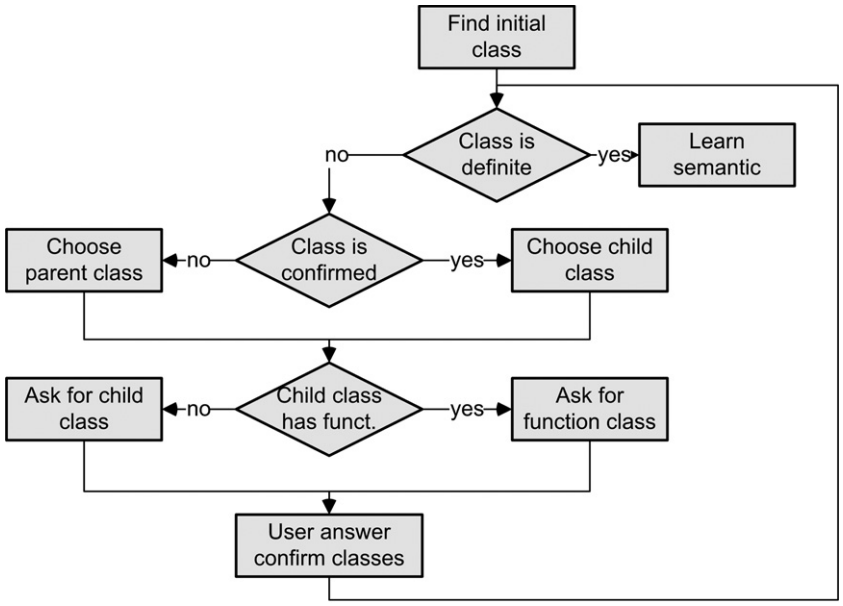

Fig. 5. Learning scheme to acquire semantic categories for an object and dialogue flow.

Table 1
Overview of the experiment and recognition rates of visual object recognition

	#	Cat 1	#	Cat 2	#	Comment
Dialog condition	52	Unknown	40	Known	12	Dialogues with known and unknown objects
Unknown detection	40	Correct	39	Failed	1	Interaction by the user in 5 cases
Known detection	12	Correct	10	Failed	2	Interaction by the user in 2 cases
Detection summary	52	Correct	49	Failed	3	

specific object or get information from the robot which objects he can see at a predefined position in the kitchen. Each of these dialogues includes detection of objects at the sink location. When an unknown object or an unknown word is detected, the learning dialogues were initiated. This way, a dialogue could be very short (if only known object), in this case these dialogues are used to evaluate detection rates. Or, the dialogues could as long as required to reach the learning goal. For example, learning an object's property does not always include learning a new word. In this case, these dialogues are used to evaluate the different learning tasks of properties and concepts.

The users did not know in advance which objects were known to the robot, and which objects were unknown. The interaction started after a brief introduction about the scenario and the robot's task. No details were given about how the robot performs its learning strategies to prevent biasing of the users. The dialog started with a greeting or directly with a request from the user to either serve a specific object, or to report which objects the robot could see. The following evaluation section describes results, success rates and recognition rates from these dialogues.

7.2. Evaluation

Meaningful numbers for the experimented scenario of interactions and learning dialogues are success rates (number of successful dialogues) and dialogue length (measured in number of turns). The first metric is important to measure the effectiveness of the approach. The second metric is important to measure the efficiency and burden for the user. Numbers are reported here for learning object categories and object properties for unknown objects. Also, a comparison of different learning strategies for object categories is made.

An overview of the experiment conditions and conducted dialogs is shown in Table 1. The table shows a total number of 5 conducted dialogues, the separation into known and unknown objects conditions and detection rates of known and unknown objects. A closer look at the different categories, shows that out of 39 objects that could correctly be detected as unknown objects, five objects required interaction by the user. The same situation happened in the known condition, where two objects required interaction by the user. Interaction by the user means that the object could not be detected upon the first try, e.g. because the object was completely or partly out of the robot's field of view. The users then turned the objects into the robot's field of view after which in all these cases, the object was classified correctly. To further analyze the errors that were made by the system, one can look at the failed attempts, which sum up to 3 out of 52. The reasons for failure were that, once visual features were not sufficient for detection, and twice known and unknown categories were confused.

These requests provided the basis for the evaluation of the learning algorithm in dialogue. Learning of an object according to the algorithm described above includes learning of the object description for reference in speech, properties of the object, and the type of the object. The description of the object however, is a combination of object properties and the type of the object. For example, the "red cup" is an example of combining the type of the object (the cup) with a property of the object (red) to create a description that can be used in speech (see Section 4.2 for details). The first part was to understand properties of the object. In the second step the type of the object was narrowed down in more detail. The two parts are addressed by the different learning dialogues described earlier, and are evaluated separately. Table 2 shows the number of dialogues, success rates and average number of turns of dialogues conducted for learning of object properties. Learning of a property value was possible in two ways. Either the word was known (25 dialogues) or the word was recognized as unknown, in which case the word could be spelled (15 dialogues).

Table 2
The three learning tasks and successful completion rates in the experiment

Task	#dialogues	Success	Avg turns
Learn object property	40	83% (33)	1.8
- with known words	25	87% (22)	1.4
- with spelling	15	74% (11)	2.6

Table 3
Application of one shot learning, browsing and the combined approach during the experiments for acquisition of the semantic category

Task	#dialogues	Success	Avg turns
One shot learning (47%)	16	81% (13)	2
Browse (6%)	2	100% (2)	10.5
Combined (47%)	16	81% (13)	4.2
All one shot (100%)	34	68% (23)	2
All combined (100%)	34	82% (28)	3.6

The more complex learning task was to learn the semantic category of an object (described in Section 6.2). Thirty four dialogues were conducted for this task. In 82% of these cases, the dialogue could be completed successfully with the learning algorithm that applies the combined approach. The combined approach was applied in all 34 dialogues. From the conducted dialogues, comparison can be drawn with the one shot learning approach and the browsing strategy. The combination of different possibilities, how a class can be learned by the system, resulted in different combinations of one shot learning and browsing. In 47% of the dialogues, the class was specified directly by the user, and could be learned directly as a pure one shot learning. After the one shot attempt, the dialogue was stopped by the user. The same number of dialogues (additional 16 dialogues) was conducted, where the class was refined after the one shot learning step. The remaining 6% of the dialogues was conducted as pure browsing of the ontology, after the one shot learning approach did not result in a recognized type that could be used for browsing. The browsing dialogue then started with the most general class in the hierarchy. This way, the user could complete the dialogue quickly with one shot learning within only two turns, if it was clear to him how to categorize the object.

Table 3 shows the figures and results from the learning dialogues for acquisition of the semantic category. The top three rows give the numbers for the three approaches as conducted in the experiment. The table shows the number of dialogues conducted for each strategy, the rates and numbers of successful dialogues and the average number of turns per successful dialogue. Since the combined approach starts with a one shot learning hypothesis, and then refines the class in further step with a browsing strategy, comparison can be drawn between one shot learning and the combined approach on all 34 samples. The number of all successful dialogues with the combined strategy is the sum of all successful dialogues. In case of the one shot learning approach, the two cases which could be learned only with the browsing strategy are classified as failures for the one shot learning approach, since no category could be identified. In addition, 3 samples of the remaining dialogues would not report an acceptable result after the one shot learning step.

8. Conclusions and future work

8.1. Conclusions

We have presented a dialogue system that is able to detect deficient information in dialogues, and initiate a learning strategy to acquire information and learn unknown objects. The system is able to learn new words, properties and types of objects. Both properties and types of objects are important to learn, since both contribute to the description of an object, which is used by users to reference an object. During reference to objects, different properties are specified by the users. The speech recognition and understanding grammar thus support a variable combination of different properties and types for each object. Since objects are categorized with different levels of abstraction, it is necessary to model functionalities as separate concepts in the ontology. The robot can then distinguish different functionalities of an object, which can be given from context in speech or from the description of the user.

The combined approach for learning of object classes has shown better success rates than pure one shot learning. The presented algorithm requires only little more interaction with the user (in terms of number of turns), but it produces significantly better results in categorizing the object according to error rates and accuracy. These first results show that the algorithm provides an accurate means to categorize unknown objects in terms of semantic categories within an ontology.

In contrast to pure recognition output of the object recognizer, employing dialogue capabilities significantly improves the final recognition results after confirmation. The dialogue uses implicit and explicit confirmation strategies, which both give the user the opportunity to interrupt the robot and correct the recognition hypothesis in the case of errors.

8.2. Future work

The presented system is able to categorize and learn new objects in dialogue with the user. The resulting knowledge base allows the system to recognize the detected object, talk about the object and understand when the user refers to the new object in speech. Further work could be directed at combining understanding approaches, such as the one presented here, with knowledge acquisition how the robot can manipulate the object. To do so, first, additional perceptual information needs to be collected, e.g. to better segment the object's shape with 3D information acquisition. Integration with vision currently requires that segmentation of an object is possible, e.g. by 3D or feature-based segmentation, and that grounding has already been done, when the learning dialogue is initiated.

Limitations of the presented approach, are that currently all object types and properties are modeled statically. To some extent, dynamic changes in the environment are reflected as properties that change over time, which is already covered by the ability to associate one object with different categories. Also different verbal representations can be associated to objects. However, the system does not cover dynamics in a way that a cup of tea only is associated with tea if it is filled with tea, and that it would be associated with coffee, if it were filled with coffee. Modeling such information requires extending the approach with a state model that keeps track of object properties, such as 'dirty', 'full', etc. Some other properties make only sense if they are interpreted as user-specific properties. For example a person's most favorite cup cannot be generalized as being the most favorite cup of everybody. But this generalization is indeed appropriate for some properties. For example, a red cup remains to be a red cup, or a book continues to have the same title, which does not change over time. For user specific properties, user ID information could be integrated as an additional variable to relate user specific properties to specific users. Another approach can be to correct wrongly stored information or 'forget' information that leads to contradictions in the knowledge base but is not necessary for interaction with the user. To assess how the system evolves over time additional experiments are required, e.g. to quantify effects of storing objects at a wrong position in the ontology.

Acknowledgments

This work was supported in part by the German Research Foundation (DFG) as part of the Collaborative Research Center 588 "Humanoid Robots - Learning and Cooperating Multimodal Robots". A special thanks goes to Pedram Azad who has provided support for the integration of the object recognition component and for experiments with the robot Armar III and to the reviewers of this journal.

References

[1] T. Asfour, K. Regenstein, P. Azad, J. Schröder, A. Bierbaum, N. Vahrenkamp, R. Dillmann, Armar-III: An integrated humanoid platform for sensory-motor control, in: Proceedings of IEEE-RAS International Conference on Humanoid Robots, Genova, Italy, 2006.
[2] P. Azad, T. Asfour, R. Dillmann, Stereo-based 6d object localization for grasping with humanoid robot systems, in: Proceedings International Conference on Intelligent Robots and Systems, IROS, San Diego, USA, 2007.
[3] R. Becher, P. Steinhaus, R. Zöllner, R. Dillmann, Design and implementation of an interactive object modelling system, in: Proceedings of ISR 2006 and Robotik 2006, Düsseldorf, 2006.
[4] J. Carbonell, Towards a self-extending parser, in: Annual Meeting of the Association for Computational Linguistics, 1979.
[5] B. Carpenter, The Logic of Typed Feature Structures, Cambridge University Press, 1992.
[6] G. Choueiter, S. Seneff, J. Glass, New word acquisition using subword modeling, in: Proceedings of Interspeech, 2007.
[7] M. Denecke, Object-oriented techniques in grammar and ontology specification, in: Workshop on Multilingual Speech Communication, Kyoto, Japan, 2000.
[8] S. Dusan, J. Flanagan, Adaptive dialog based upon multimodal language acquisition, in: Proceedings of the Fourth Int. Conf. on Multimodal Interfaces, Pittsburgh, PA, USA, 2002.
[9] S. Dusan, J. Flanagan, A system for multimodal dialogue and language acquisition, in: Invited, The 2nd Romanian Academy Conference on Speech Technology and Human-Computer Dialogue, Romanian Academy, Bucharest, Romania, 2003.
[10] M. Fritz, G.-J.M. Kruijff, B. Schiele, Cross-modal learning of visual categories using different levels of supervision, in: International Conference on Computer Vision Systems, ICVS, Bielefeld, Germany, 2007.
[11] P. Gieselmann, H. Holzapfel, Multimodal context management within intelligent rooms, in: Proceedings of the 10th International Conference on Speech and Computer, SPECOM, Patras, Greece, 2005.
[12] P. Gieselmann, P. Stenneken, How to talk to robots: Evidence from user studies on human–robot communication, in: Proceedings of the Workshop on How People Talk to Computers, Robots, and other Artificial Communication Partners, Bremen, Germany, 2006.
[13] P. Gorniak, D. Roy, Probabilistic grounding of situated speech using plan recognition and reference resolution, in: Proceedings of Seventh International Conference on Multimodal Interfaces, ICMI, 2005.
[14] H. Holzapfel, A dialogue manager for multimodal human–robot interaction and learning of a humanoid robot, Industrial Robots 35 (6) (2008).
[15] H. Holzapfel, T. Schaaf, H.K. Ekenel, C. Schaa, A. Waibel, A robot learns to know people - first contacts of a robot, in: C. Freksa, M. Kohlhase, K. Schill (Eds.), in: Lecture Notes in Computer Science - KI 2006: Advances in Artificial Intelligence, vol. 4314, Springer, 2007, pp. 302–316.
[16] H. Holzapfel, A. Waibel, A multilingual expectations model for contextual utterances in mixed-initiative spoken dialogue, in: Interspeech 2006, ICSLP, Pittsburgh PA, USA, 2006.
[17] E.C. Kaiser, Using redundant speech and handwriting for learning new vocabulary and understanding abbreviations, in: Proceedings of the 8th International Conference on Multimodal Interfaces, ICMI, ACM Press, New York, NY, USA, 2006.
[18] A. Kasper, R. Becher, P. Steinhaus, R. Dillmann, Developing and analyzing intuitive modes for interactive object modeling, in: Proceedings of the Ninth International Conference on Multimodal Interfaces, ICMI, Association for Computer Machinery, Nagoya, Japan, 2007.
[19] S. Kirstein, H. Wersing, E. Körner, Online learning for object recognition with a hierarchical visual cortex model, in: Lecture Notes in Computer Science on Artificial Neural Networks: Biological Inspirations ICANN 2005, vol. 3696/2005, 2005, pp. 487–492.
[20] F. Lömker, G. Sagerer, A multimodal system for object learning, in: Lecture Notes in Computer Science, vol. 2449, 2002, pp. 490–497.
[21] D.G. Lowe, Object recognition from local scale-invariant features, in: International Conference on Computer Vision, ICCV, Corfu, Greece, 1999.
[22] A. Park, J.R. Glass, Unsupervised word acquisition from speech using pattern discovery, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP, Toulouse, France, 2006.
[23] D.K. Roy, Grounded spoken language acquisition: Experiments in word learning, IEEE Transactions on Multimedia 5 (2) (2003) 197–209.
[24] T. Schaaf, Detection of OOV words using generalized word models and a semantic class language model, in: Proceedings of Eurospeech, 2001.
[25] T. Schaaf, Erkennen und Lernen neuer Wörter, Ph.D. thesis, Universität Karlsruhe (TH). 2004.
[26] O. Scharenborg, S. Seneff, Two-pass strategy for handling oovs in a large vocabulary recognition task, in: Proceedings of Interspeech, 2005.
[27] T. Shibata, N. Kato, S. Kurohashi, Automatic object model acquisition and object recognition by integrating linguistic and visual information, in: Proceedings of the 15th ACM International Conference on Multimedia, ACM Multimedia 2007, Augsburg, Germany, 2007.
[28] H. Soltau, F. Metze, C. Fuegen, A. Waibel, A one pass-decoder based on polymorphic linguistic context assignment, in: Proceedings of ASRU'01, Madonna di Campiglio, Trento, Italy, 2001.
[29] L. Steels, J.-C. Baillie, Shared grounding of event descriptions by autonomous robots, Robotics and Autonomous Systems 43 (2–3) (2003) 163–173.
[30] L. Steels, F. Kaplan, Aibos first words. the social learning of language and meaning, Evolution of Communication 4 (1) (2002) 3–32.
[31] R. Stiefelhagen, H.K. Ekenel, C. Fügen, P. Gieselmann, H. Holzapfel, F. Kraft, K. Nickel, M. Voit, A. Waibel, Enabling multimodal human–robot interaction for the Karlsruhe humanoid robot, IEEE Transactions on Robotics 23 (5) (2007) 840–851.
[32] B. Wrede, M. Kleinehagenbrock, J. Fritsch, Towards an integrated robotic system for interactive learning in a social context, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2006.

Hartwig Holzapfel received the Diploma degree in computer science from the Universität Karlsruhe (TH), Karlsruhe, Germany, in 2003. He is a Research Assistant and Ph.D. student at InterACT, the International Center for Advanced Communication Technologies at the Universität Karlsruhe (TH), Germany. His research interests include multimodal dialogue management and interactive learning.

Daniel Neubig received the Diploma degree in computer science from the Universität Karlsruhe (TH), Karlsruhe, Germany, in 2008. He is now a software engineer at AdNovum Informatik AG, Zürich, Switzerland.

Alex Waibel received the B.S. in electrical engineering from Massachusetts Institute of Technology, Cambridge, in 1979, and the M.S. and Ph.D. degrees in computer science from Carnegie Mellon University, Pittsburgh, PA, in 1980 and 1986, respectively. He is currently a Professor of Computer Science at Carnegie Mellon University, and also Professor at the Universität Karlsruhe (TH), Karlsruhe, Germany. He is associated with InterACT, the International Center for Advanced Communication Technologies at both universities. At CarnegieMellon, he also serves as Associate Director of the Language Technologies Institute and holds joint appointments in the Human-Computer Interaction Institute and the Department of Computer Science. He is also associated with the Human Interaction Loop (CHIL) program and the NSF-ITR project STR-DUST. He has founded and cofounded several successful commercial ventures. He was one of the founders of C-STAR, the International Consortium for Speech Translation Research and served as its Chairman from 1998 to 2000. His team has developed the JANUS speech translation system, and, more recently, the first real-time simultaneous speech translation system for lectures. His laboratory has also developed a number of multimodal systems including perceptual meeting rooms, meeting recognizers, meeting browser, and multimodal dialog systems for humanoid robots. He is the holder of several patents. His current research interests include speech recognition, language processing, speech translation, and multimodal and perceptual user interfaces. Prof. Waibel was the recipient of numerous awards for his work and publications.