Fachrichtung 4.7 Allgemeine Linguistik
Universität des Saarlandes

Grounding Robot Gaze Production in a Cross-Modal Category System

Master Thesis

Betreuer:
Dr. Geert-Jan M. Kruijff und
Prof. Dr. Hans Uszkoreit

Maria Staudte

Staudte, Maria
Grounding Robot Gaze Production in a Cross-Modal Category System
Master Thesis,
Saarland University, Saarbrücken, Germany
September 2006, 74 pages

Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, dass ich diese Arbeit selbständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Saarbrücken, September 2006

Maria Staudte

Acknowledgements

I wish to thank . . .

. . . my supervisors Dr. Geert-Jan Kruijff and Prof. Dr. Hans Uszkoreit for giving me the most valuable feedback even in the most odd circumstances and sometimes at very short notice.

. . . my office colleagues Sabrina Wilske, Katja Ignatova and Muralikrishnan Ramasamy and particularly Hendrik Zender for giving constant professional and moral support. Thanks to Sabrina, Hendrik and Henrik Jacobsson also for great LaTeX-support.

. . . colleagues from the psycholinguistic department for interesting and inspiring discussions.

. . . Jens Apel, among other friends, for proof-reading or otherwise contributing to the process of writing this thesis.

Zusammenfassung

Die Augenbewegung – der *Blick* (engl. *gaze*) – einer Person erfüllt mehrere wichtige Funktionen während eines Dialogs zwischen zwei Menschen. Er steuert und beschränkt unter anderem ihre visuelle Wahrnehmung. Zugleich dient er Gesprächsteilnehmern als Anhaltspunkt für den aktuellen Gegenstand der Aufmerksamkeit ihres Gegenübers. D.h. sprechen zwei Menschen über Objekte in ihrer unmittelbaren Umgebung, gibt die Augenbewegung Aufschluss über den Bezug der Äußerung. Was hier für die Kommunikation zwischen Menschen gilt, kann prinzipiell aber auch auf Mensch-Roboter-Interaktion übertragen werden. Um daher die Interaktion zwischen einem Menschen und einem Roboter so einfach wie möglich zu gestalten, sollte der Roboter seinen *Blick* in ähnlicher Weise wie der Mensch einsetzen. Ein natürlicher und zweckmäßiger *Blick*, der unter anderem die Menge der zu verarbeitenden visuellen Stimuli einschränken soll, kann nur aus einem tieferen Verständnis der Situation hervorgehen. Um zu verstehen, wie der *Blick* mit diesem Verständnis zusammenhängt, muss untersucht werden, warum und wohin Menschen in einer bestimmten Situation schauen. Eine Situation in diesem Sinne beinhaltet sowohl visuelle Aspekte der Umgebung als auch sprachliche Aspekte des Dialogkontexts.

Das Phänomen *Blick* wurde am Menschen bereits auf verschiedene Weisen und unter unterschiedlichen Aspekten erforscht. Die Ergebnisse solcher Studien lassen auf eine enge, zeitlich koordinierte Interaktion zwischen inkrementell verarbeiteter Sprache, visueller Verarbeitung und Weltwissen schließen. Weiterhin werden darin Kategoriesysteme, die zwischen den Inhalten der verschiedenen Modalitäten vermitteln, für diese Interaktion verantwortlich gemacht. Um ein künstliches kognitives System, wie z.B. einen Roboter, mit einem nützlichen und aufschlussreichen *Blick* zu auszustatten, verwenden wir diese Erkenntnisse über verantwortliche Mechanismen beim Menschen. Auf der Grundlage jener Erkenntnisse entwickeln wir ein Modell für die Produktion von *Blick* bei Robotern. Es ermöglicht dem Roboter eine vorausschauende Augenbewegung, die beide Funktionen erfüllt: kommunikative Information über die eigene Aufmerksamkeit und die Beschränkung der Menge aller möglichen visuellen Stimuli auf die relevanten. Wir implementieren dieses Modell mittels verteilter Ontologien zur Simulation von Kategoriesystemen kombiniert mit inkrementeller Sprachverarbeitung und visueller Szenenanalyse.

Der vorgestellte Ansatz beinhaltet ein allgemeines Modell für *Blick*-Produktion und einen spezifischen Vorschlag zu dessen Implementierung. Das Modell basiert auf Ergebnissen von Studien zu menschlicher *Blick*-Performanz und trägt damit zum aktuellen Stand der Forschung auf kognitiver als auch technischer Ebene in der Erzeugung von *Blick*-Verhalten bei.

Abstract

Gaze has multiple functions in situated dialogue. It guides perceptual processing and provides feedback to the interlocutor by indicating what is being attended to. These principles for gaze in human-human interaction are also applicable to human-robot-interaction. Thus, to make interaction between a human and a robot as easy and natural as possible, the robot would ideally produce gaze in a way similar to humans. Gaze production that is natural, flexible and that can be used to reduce the load of perceptual processing, needs to be grounded in situational awareness. The prerequisite for that is to understand what makes humans direct gaze to a particular aspect of the situation in the first place, based on what is being talked about.

Gaze in humans has already been studied by means of various methods. The results of these studies reveal a closely time-locked interaction between incremental utterance comprehension, visual scene processing and world knowledge. Moreover, the mechanisms enabling this close interaction have been identified as categorical mediation between modal contents. Therefore, to equip a robot, as an artificial cognitive system, with useful and communicative gaze behaviour, we employ these insights into the underlying mechanisms of human gaze. The model we propose for robot gaze production is based on these insights and is designed to reproduce some of the observed effects. It allows a robot to use gaze as an anticipatory process providing both functions, communicative feedback and reduction of the amount of perceptual processing. We implement this model by means of distributed ontologies for modelling the categorical aspects and combine them with incremental utterance comprehension as well as basic visual scene analysis.

The presented approach contains a general model for gaze production and a specific implementation proposal. The model draws on results for human performance and therefore contributes to the cognitive as well as the engineering scientific point of view on gaze.

Contents

Chapter 1

Introduction

Gaze and its role

In situated dialogue, people look around, i.e. use gaze, when talking and listening. People tend to look to objects before naming them (Griffin, 2004), and during utterance comprehension people typically move their eyes to aspects of the situation that they expect to be mentioned next in the utterance (Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995; Altmann & Kamide, 1999). There are reasons for why people do this.

During utterance comprehension this "looking around" (*saccadic eye movement* or *gaze*) has at least the following two functions. One, it acts as a non-verbal cue for organising turn-taking (Novick, Hansen & Ward, 1996) and for providing feedback. You look at what you understand, or expect, the speaker to be talking about. This way one establishes a common ground in the dialogue, indicating how one resolves references to relevant aspects of the situation. Furthermore, people do not process scenes immediately in full detail, but in a more gradual fashion (Henderson & Ferreira, 2004). This points to another function of gaze. One uses what is being talked about to guide perceptual processing of the situation, focusing on comprehending only what is relevant.

Gaze in HRI and why it is a problem

Several authors experimentally attested the importance of gaze in human-robot interaction (HRI), e.g. Miyauchi, Sakurai, Makamura and Kuno (2004); Sidner, Kidd, Lee and Lesh (2004); Sidner, Lee, Kidd, Lesh and Rich (2005). However, existing approaches to producing robot gaze in HRI do not refer to situational awareness, other than recognizing human head gesture. The robot makes pre-determined saccades, without comprehending where and what it is looking at. This does not scale well to the natural, dynamic scenes (Henderson & Ferreira, 2004) in which service robots are to be deployed, nor does it enable exploring the potential for gaze to help gradually refining the robot's situational awareness. Breazeal, Hoffman and Lockerd (2004), for instance, have employed a model of robot gaze in order to give feedback in HRI. However, they use scripted gaze in a way such that the robot simply glances at an "area of change" to signal understanding, or it

looks at the human otherwise to signal attention. Sidner et al. only implemented the latter function. Yoshikawa, Shinozawa, Ishiguro, Hagita and Miyamoto (2006), on the other hand, have used gaze recognition for the robot's gaze precisely to mimic its human partner by feeding eye-tracking data from the human subject directly into the camera control module. Gaze in these systems is not based on a deeper understanding of the situation. This results in a rigid and merely reactive behavior that is not flexible enough to adapt to novel situations. Gaze, however, should be used also in an anticipating way to fulfill the functions mentioned above. It has to guide selective attention, i.e. the robot looks where it expects to perceive relevant events. Reversely, if the robot does not understand where and what it looks at, perception cannot contribute to the robot's representation of the current scene. Therefore, there can be no gradual refinement of the context understanding which includes language processing.

The fundamental problem is to find out and model what makes one direct gaze to a particular aspect of the situation in the first place, based on what is being talked about. To equip an artificial cognitive system with efficient and communicative gaze behaviour, one needs to examine this question in human gaze and use the results to create a model for robots. Empirical investigations (Altmann & Kamide, 2004; Knoeferle & Crocker, 2006) show that this is an issue of *mediation* between language processing, perceptual processing, and "world knowledge." More precisely, where one expects to see something relevant arises from a mediation between language, perception, and a situational awareness (Endsley, 2000) which combines both *categorical* understanding (Barsalou, 1999; Glenberg, 1997; Glenberg & Kaschak, 2002; Lakoff, 1987) and *spatio-temporal* understanding (Calvert, 2001; Hickok & Poeppel, 2004).

A general model on human gaze production

The general hypotheses on gaze that psycholinguistic and psychological research provide are the following.

a) Human gaze is based on situated understanding of the scene (which is in our case the current real-world environment). This "situational awareness" is a combination of context knowledge, personal experience of the perceiver including acquired world knowledge and the projection thereof into potential events/situations. Knoeferle and Crocker (2006) observed closely time-locked interaction of linguistic analysis and the grounded meaning of communicated content which resulted in expectations about further information, indicated by the subject's gaze.

b) *Situated* understanding is possible because conceptual structures directly relate to sensori-motoric signals and vice versa. Thus, concepts are grounded in percepts and have a meaning that evolves from interaction with the environment, cf. Barsalou (1999); Glenberg and Kaschak (2002). On the other hand, (grounded) content is being mediated between modalities (Glenberg et al., 2005; Barsalou, 2005).

So humans develop expectations about grounded meaning by relating activated concepts to the understood content so far. We explicitly integrate this with observed mediation mechanisms: drawing on other modalities such as vision humans also produce a data-driven or *bottom-up* preference over possible continuations of the linguistic content. Together, these predictions compose *top-down* attentional preferences for concepts, e.g. of objects or actions, yet to be perceived in the scene (cf. Desimone and Duncan (1995) on attention as resolving perceptual competition). Obviously, these anticipations are then resolved with respect to visuo-spatial understanding of the scene such that the human knows where specifically the anticipated object is or the potential action might take place. This then allows her to look at that place and produce exactly the behaviour as described above.

Concrete goal and proposed model

Our goal is to motivate and implement a computational model of the generation of top-down expectations which ultimately lead to a more natural use of gaze in HRI. In this thesis we focus on how categorical understanding in situational awareness can contribute to directing gaze. We provide a model for basic category systems of objects and actions (Barsalou, 1999; Glenberg & Kaschak, 2002) implemented as distributed ontologies. Combining inferencing over these ontologies and associations between these ontologies, categories are activated using input from modalities like vision or speech. Using an incremental model for utterance analysis (Steedman, 2000), and a basic model of scene understanding (Kelleher, Kruijff & Costello, 2006), we give a working model of the interaction between activating categories, priming linguistic analysis, and guiding gaze to objects in a scene we expect to be talked about. We thus model gaze not only as a reactive behaviour (following where a speaker is looking), but also as an anticipatory behaviour by combining incremental language processing with situational awareness. The implementation of this model is embedded into an already existing artificial system developed by the CoSy-Project [1], an EU-funded research project on cognitive systems for cognitive assistants.

Contributions

The scientific contribution of this thesis consists of

- the integration of interdisciplinary research on the use and effects of gaze in visually situated dialogue.

[1] http://www.cognitivesystems.org

- the relation thereof to situated categorical understanding. That is we investigate the role of cross-modal category systems for gaze production and create a model of the identified mechanisms.
- the development of a computational framework based on that model which helps us verify these ideas on a platform for human-robot interaction.

The importance of gaze for HRI has been observed before, but no existing system has grounded gaze production in situational awareness. Instead, it has been scripted or is purely reactive at a perceptual level, thus lacking comprehension. A yet different approach is pursued by Mayberry et al. , for instance, who follow the approach of simulating human gaze by employing neural networks e.g. in Mayberry, Crocker and Knoeferle (2005). We, on the other hand, are interested in understanding the underlying mechanisms. We propose a model of these and, thus, contribute to both a *cognitive* understanding of situated dialogue as well as to a more principled way of producing gaze in HRI.

Overview

An overview of the thesis is as follows. In Chapter 2 we review relevant empirical observations from psycholinguistics and psychology on human gaze production. These observations yield a statement of the requirements for a model of robot gaze production. The requirements and our model are discussed in Chapter 3, and the implementation thereof in Chapter 4. We sketch runs of the implemented model and propose scenarios for evaluation in Chapter 5. The general conclusions and a sketch of possible extensions of this work follow in Chapter 6 which is rounded off with an outlook onto potential applications.

Chapter 2

Background

Summary

In this chapter, we motivate the integration of findings from psychological studies into a model of robot gaze production. In Section 2.1 we introduce studies and their results on the role of gaze in visually grounded dialogue. Section 2.2 presents insights on how knowledge is organised in category systems so that different perceptual and cognitive processes can interact. In Section 2.3 we briefly introduce existing approaches to distributed category systems. We discuss the presented findings and some shortcomings of the studies in Section 2.4 before concluding the chapter in Section 2.5.

Several approaches to robot gaze production exist, among these Miyauchi et al. (2004); Sidner et al. (2004); Yoshikawa et al. (2006); Breazeal et al. (2004). However, these HRI-systems use gaze in a more or less scripted manner. Either it is bound to entirely reproduce the interaction partner's gaze or it is a reaction to conversational cues and to very simplified scene processing indicating whether some change in the setting has occurred. This kind of gaze can improve the perceived naturalness in HRI but it is not flexible and does not scale well to natural dynamic environments. A merely reactive behavior is the result. This does not enable the desired anticipatory use of gaze such that the robot looks where it expects something relevant to occur. Moreover, the lack of understanding of where and what it looks at prevents a contribution of the perceived to the robot's representation of the current scene. Therefore, neither the scene representation nor the linguistic analysis can be gradually refined by integrating on-line information.

To build a better model of robot gaze production, one which is not purely reactive and is adoptable to new situations, we use results from psycholinguistic and psychological research on gaze in humans. We are interested in how gaze evolves naturally, how it is used and what that tells us about the underlying principles. After all, gaze in HRI is supposed to make communication with humans more natural and therefore easier. That means, we need to look at man as technology's standard and study the role of gaze in a) providing feedback and b) guiding attention to what is relevant. In this thesis, we focus on the latter and examine the underlying mechanisms.

2.1 Gaze and utterance comprehension in visually situated dialogue

In this section, we look at what gaze reveals about the interaction between human sentence processing and visual processing. Empirical studies in psycholinguistics have investigated what information listeners use when comprehending spoken utterances. These studies use eye-trackers to monitor where people look at in a scene, and when. Knoeferle and Crocker (2006) argue that these findings identify two core dimensions of the interaction between language and situated experience. One is the *temporal dimension*: Eye movements during utterance comprehension reveal that visual attention is closely time-locked with utterance comprehension. The second one is the *information dimension*, indicating how for utterance comprehension listeners draw not only upon linguistic information, but also upon scene understanding and "world knowledge." Below we discuss studies investigating the latter two aspects.

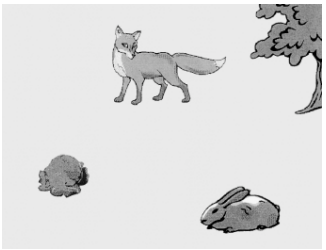


Figure 2.1: "**Der** Hase frisst gleich den Kohl." and "**Den** Hasen frisst gleich der Fuchs."

A number of related studies, e.g. Tanenhaus et al. (1995); Altmann and Kamide (1999); Kamide, Altmann and Haywood (2003), have revealed that listeners focus their attention on objects before these objects are referred to in the utterance. Figure 2.1, for instance, illustrates the setup of Kamide et al. (2003). When someone hears "The hare-nominative eats the cabbage", her gaze already moves to the cabbage in the scene before she has actually heard that word; similarly for "The fox eats the hare-accusative". Knowing that foxes typically eat small animals (not vegetables), and that the argument structure of *eat* reflects this, the listener *expects* that the next object to be mentioned will be the hare, and directs gaze to that object.

The above is an example for how world knowledge influences sentence processing and visual processing, i.e. where one looks at. The scene understanding, on the other hand, can also influence online utterance comprehension. For example, consider the situation in Figure 2.2. Tanenhaus et al. (1995) show that, once the listener has heard "Put the apple on the towel ...", she faces the ambiguity of whether to put the (lone) apple onto the (empty) towel, or to take the apple that is on the towel and put it somewhere else. The ambiguity is revealed as visual search in the scene. Only once she has heard the continuation "... into the box" this ambiguity can be resolved. Interestingly, in (Tanenhaus et al., 1995) the listener cannot directly manipulate the objects. If this is possible (cf. Figure 2.2), Chambers, Tanenhaus and Magnuson (2004) show that also reachability plays a role in comprehending the utterance. Because only one apple is reachable, this is taken as the preferred referent, and as such receives the attention.

Interestingly, the influence among perceptual processes across modalities is not restricted to the comprehension of the current situation. It also affects the projection towards possible future events (Endsley, 2000). Kamide et al. (2003); Altmann and Kamide (2004) show how such projection can affect utterance comprehension. Given a scene with a table, and besides it a glass and a bottle of wine as in Figure 2.3, these studies investigated where listeners look when they hear "The woman will put the glass on the table. Then, she will pick up the wine, and pour it carefully into the glass." It turns out that after hearing the "pouring" phrase, listeners look at the table, not the glass (depicted in Figure 2.3(b)). That indicates that sentence processing and world knowledge can prime gaze and even invoke mental images on anticipated events.

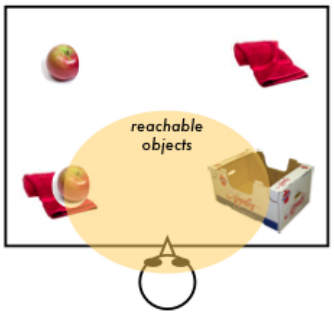
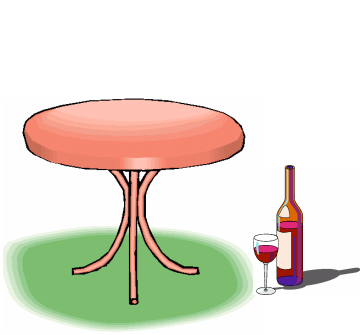
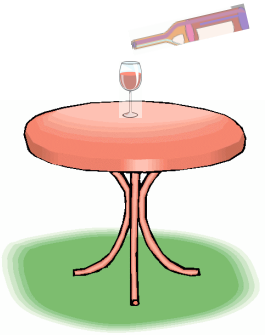


Figure 2.2: "Put the apple on the towel..."



(a) Initial scene, without auditory stimulus.



(b) "The woman will put the glass on the table. Then she will pick up the wine, and pour it carefully into the glass."

Figure 2.3: Visualisation of gaze within the depicted scene: people tend to track the described yet undepicted action in the image.

These studies show that information perceived via different modalities is integrated online and in a time-locked manner. How instantly this integration takes place is nicely illustrated in (Allopenna, Magnuson & Tanenhaus, 1998). The authors have found out that, what can be observed during utterance comprehension with respect to eye-movement and expectations for future input, is not restricted to whole words and categories. Instead, their study revealed that already with the first syllable of a word, all possible lexical hypotheses within the range of the visual scene are activated and anticipated.

Moving from the more general hypothesis of linguistic and visual interplay, we now consider some studies that focus on examining in a more principled and general way what people attend to in a visual (real-world) scene. Henderson and Ferreira (2004); Henderson (2003) present various approaches and findings about what in a visual scene naturally draws our attention to it (bottom-up) and how it is influenced by world-knowledge and goals (top-down). For the former, the stimulus-based gaze control, several features gathered through 'scene statistics' seem to catch our attention, e.g. high spatial frequency and edge density as well as high local contrast. These insights need to be taken into account when doing visual scene analysis in the first place. For knowledge-based gaze control Henderson and Ferreira (2004) again distinguish among several types: a) episodic scene knowledge, i.e. remembering certain spatial arrangements of a scene such as the photograph that always stands on the left corner of a colleague's desk, b) scene-schema knowledge, i.e. certain acquired generic knowledge about a scene, e.g. the chair that is typically in front of the desk, and finally c) task-related knowledge, which very quickly helps to filter information that may be relevant for fulfilling a task from irrelevant information in a scene. This is further evidence for the hypothesis that gaze relies on an deeper understanding of the situation, integrating experience, world knowledge and contextual knowledge from the current situation.

So far we have only considered the role of gaze in individual perception and comprehension without interaction partners. Since we are interested in HRI, i.e. in communication and gaze as communicative modality, it is necessary that we also take studies on gaze in dialogue into account. Hadelich and Crocker (2006) and Pickering and Garrod (2004), for instance, have shown gaze to be automatically aligned in simple collaborative interaction. The time intervals between eye-fixations during production and comprehension of a referring expression are shorter than in monologue. This is further evidence for the relevance of visual common ground of interlocutors and how that accelerates the activation of jointly relevant categories.

The studies presented in this section reveal that there exists a tight interaction between vision and language. They further indicate that this interaction is not *direct*, but *mediated* (Altmann & Kamide, 2004). There appears to be some categorical structure in-between that allows the projection of content from one modality into the other, and further from the current scene into a (possible) future scene. In the next section, we present further studies that support the idea of category systems that mediate between perceptual modalities and language. Moreover, it will be examined what properties these category systems may have and just how this mediation process could be modelled.

2.2 What do we know about categories?

Apparently, categorical understanding plays an important role in the sensori-motoric grounding of language. This is underlined by studies like (Glenberg & Kaschak, 2002;

De Vega, Robertson, Glenberg, Kaschak & Rinck, 2004), showing how categorical understanding gives rise to expectations based on affordances, influencing comprehension of spatial or temporal aspects of action verbs. Below we examine categorical understanding in more detail and investigate the nature of categories in humans, how they are acquired and what properties they may have. Further, we want to know how categories can be represented and combined to form category systems. And finally, considering particularly affordances, we relate the characteristics of category systems back to their purpose of mediating content.

First, we consider the questions what categories are and why we think it is essential to analyse them. Just like any other living being humans categorise. For instance, as Lakoff and Johnson (1999) point out, people need to categorise food and non-food, dangerous situations and enemies, friends and family. It is in the nature of every living being to distinguish and evaluate stimuli from the outside world, simply because it is of essential relevance whether a stimulus is vital or dangerous. It is also rooted in our biology that humans categorise, since their main information processing apparatus is a neural system that has sparse connections. There is no one-to-one mapping, instead many active neurons need to be mapped to few neurons and thereby a classification into similar patterns is the inevitable result, cf. Kandel, Schwartz and Jessel (1991). The same applies for the light-sensitive cells of the retina, where many thousand cells that receive light activate few retinal ganglion cells. Categorisation happens on all levels up to very high-level conceptual classification which may even be introspected consciously. What humans may not be aware of, however, is that the way the world is perceived and divided into categories is by no means the only way and certainly does not reflect the objective structure of the world. As Lakoff and Johnson (1999) observe rightly, human beings perceive wind and sun and green trees because they have skin with haptic sensors and light- and colour-sensitive cells in the eye. Humans perceive up- and down-movement because they sense gravity and balance and have a very complex sense for orientation. Possibly only because they have muscles and can move autonomously, movement is perceived at all.

Consequently, everything that is perceived is somehow influenced by how the perceiver's body is shaped and interacts with the environment. This is important to bear in mind because it explains why category systems cannot be investigated in isolation, without taking the whole embodied system into account. It also means that categories are formed through experience, through perception and interaction and therefore through sensorimotoric interplay. This in turn means that categories are grounded in situations and evolve from modal interaction, i.e. they are not abstract symbols which can be arbitrarily rearranged. It does not explain what properties these categories may have but it strongly suggests that they are generally *embodied*.

The next question is how these (modal) categories are exactly acquired. This may be interesting and give further indications on what categories may reflect. For instance, Brown (1958) and Rosch, Mervis, Gray, Johnson and Boyes-Braem (1976) (as cited by Lakoff

(1987)), have studied the acquisition of linguistic categories a long time ago. Their work on *basic-level categories* has been considered as very significant in learning about how humans categorise and is still relevant today. Here we want to briefly introduce some results of studies and for that partially refer to Lakoff (1987). Brown coined the term *basic-level categories* and meant distinctive actions and objects that children learned to distinguish and name first and that have the shortest names. Rosch et al. elaborated on this idea and found that *basic-level categories* are high level such that all members have similar properties (although some are more 'prototypical' than others). Furthermore, a single image can reflect all members of such a category and similar motor actions are used for interacting with its members. Thus, they specified the notion of 'basic level' to be basic for humans in perception (image), function (motor interaction), communication (short names, easiest to learn, most frequent) and knowledge organisation, e.g. relating to other categories. Moreover, Rosch et al. showed in experiments with children that *basic-level categories* are formed spontaneously and mostly with respect to what functional parts are perceived (cf. also Tversky and Hemenway (1983)). This again is based on the specific motoric capabilities and what is perceived as a possible function. Therefore, the formation of such categories is to a large extent depending on their "interactional properties" (as opposed to 'objective' properties) which again emphasises the importance of situated experience, and affordances in particular, on understanding and communication skills.

The acquisition of taxonomic logics in language, e.g. forming super- and subclasses, takes longer and can be observed with older children. Whether this is transferable to categorisation in general has been questioned by Mandler and colleagues (cf. Mandler and McDonough (1998)) who have shown that in non-linguistic tasks children rather form global categories first that then help them to acquire more specific (linguistic) categories. Borghi, Parisi and Ferdinando (2005) support this with a study of neural networks learning to categorise simulated objects and find that the networks acquire (non-linguistic) superordinate categories earlier than basic-level ones. However, they also support the claim that categorisation is action-based, i.e. it is primarily influenced by the functionality one assigns to it.

Summing up, there appears to be some kind of categorical structure organising human perceptual and cognitive processes and it seems like the forming of categories is crucially influenced by their "interactional" properties, i.e. affordances. Hence, these categories are somehow linked to the experience of the perceiver and are apparently interacting with each other. It remains to be seen how the categories are combined into larger systems. In the following we go into further detail on what properties such category systems may have and how they can provide the mentioned mediation function. There exist several theories on category systems, some of which are quite contrary with respect to what the requirements for category or "symbol" systems are and how they are met. The general issues one needs to address when discussing requirements of category systems are:

- Can the system distinguish types and tokens, i.e. classes and instances?
- How are categorical inferences dealt with?
- Can symbols be combined generically?
- Does the system represent propositions, i.e. applications of symbols to situations?
- Are abstract categories, e.g. *freedom* or *idea*, represented?

Classical a-modal symbol systems are considered to meet these requirements. But they have other short-comings when used in an embodied system. Their meaning is not grounded in the specific interactional properties and capabilities of the system. Therefore they are not adaptive (and in fact not meaningful) to the system and its perception in any given real-world situation. This also inhibits the activation of perceived categories and hence what is being paid attention to. The approach we follow is argued for by empiricists like Barsalou, Glenberg, Prinz or Damasio. Barsalou (1999), for instance, claims perceptual symbol systems (henceforth PSS) can meet the above stated requirements and, furthermore, that they do not have the same short-comings as classical a-modal systems. He claims that

> "perceptual symbols are modal and analogical. They are modal because they are represented in the same systems as the perceptual states that produce them. [...] Because perceptual symbols are modal, they are also analogical. The structure of a perceptual symbol corresponds, at least somewhat, to the perceptual state that produced it."

Barsalou suggests six core properties for a PSS that ensure it can meet the above requirements.

1. **Neural representation:** A perceptual symbol is a record of a neural state that underlies perception, i.e. sensori-motoric experiences.

2. **Schematic perceptual symbols:** Such a symbol does not record the entire neural state of the brain but rather comprises a schematic aspect. This schematic nature of the symbol is due to the isolating influence of selective attention. He further states that perceptual symbols a) are dynamic, b) are componential, c) need not represent specific individuals and d) can be indeterminate.

3. **Multimodality:** Each symbol is established in the brain area that is responsible for its perception type, i.e. the modality, and is therefore always modal. (see e.g. Damasio et al. (2004) for neural studies on this issue)

4. **Simulators and Simulations:** In long-term memory, perceptual symbols are organised into simulators, i.e. they are related spatially and temporally according to the experience that caused their storage. The simulators allows later complex and dynamic simulations.

5. **Frames:** The symbols are organised into simulators by frames that integrate them across category instances

6. **Linguistic Control:** Words can be associated with simulators and thus enable the control of simulation construction.

Property one states that symbols are unconsciously processed perceptions represented by neural activation patterns. It also implies that instances of perception are ordered into classes (categories) since this is one of the core properties of neural networks. According to property two they are typically partial representations and can be combined to form more complex symbols/representations (properties four and five). These properties meet the third requirement, for instance. Property four also meets requirement number four by facilitating the creation of situation-based simulations. Properties five and six yield more complex and often cross-modal symbols that could be used to construct abstract categories as required by the last issue. Glenberg (1997) similarly suggests that the meaning of categories is based on experience.

> "[...] perceptual systems have evolved to facilitate our interactions with a real, three-dimensional world. To do this, the world is conceptualised (in part) as patterns of possible bodily interactions, that is, how we can move our hands and fingers, our legs and bodies, our eyes and ears, to deal with the world that presents itself? That is, to a particular person, the meaning of an object, event, or sentence is what that person can do with the object, event, or sentence." (p.3)

He proposes the concept of "meshing" that relates categories like associations do. This relation, however, is more subtle than simple associations in the common sense. The related categories modify each other to obey posed constraints on bodily action. To illustrate this mutual modification imagine a piece of paper. The sheet is painted with red and blue stripes. If one imagines crumbling it the pattern on the mental sheet of paper will look just as crumbled, even after unfolding the sheet. This example shows how a category of a physical object is conceptually combined with categories for colour and shape in a way that they form a new entity with possibly additional or modified properties. This sort of grounding of categories and the interaction between categories is what we need to understand in order to model a system that provides content mediation across modalities and across linguistic categories like nouns and verbs. Further evidence and further elaboration on this approach comes from Glenberg and Kaschak (2002). They propose a specific account of how words could be grounded specifically in action. They discovered a phenomenon called the "action-sentence compatibility effect" (ACE) which describes the interference of the implied action of a sentence and the action that the subject is required to execute. The sentence "Close the drawer.", for instance, requires a movement away from the body. The implied direction interferes with a requested response task of

the subject. That means when the subject is supposed to make a movement towards the body this would be slightly delayed in comparison to a movement along the direction of the implied movement by the sentence. This effect generally supports the idea that meaning of language is grounded in perception and particularly that it is grounded in action. In agreement with this data is also the *indexical hypothesis* by Kaschak and Glenberg (2000) which states that *"the meaning of a sentence is constructed by (a) indexing words and phrases to real objects or perceptual, analog symbols; (b) deriving* affordances *from the objects and symbols; and (c) meshing the affordances under the guidance of syntax."*. The mediating effect of categories is further underlined by studies like (Barsalou, 2005; Altmann & Kamide, 2004; De Vega et al., 2004) showing how categorical understanding gives rise to expectations based on affordances, influencing comprehension of spatial or temporal aspects of action verbs. It is our goal to grasp and exploit these expectations for evoking predictive gaze behaviour. Hence, we take a closer look at affordances and what they can offer with respect to associating behaviour of categories.

The term *affordance* has been coined by Gibson (1979), and was refined and empirically supported by Phillips and Ward (2002) and Tucker and Ellis (1998). It denotes the *action* that a human typically associates with a certain type of object. In other terms, it is a function assigned to the object by the perceiver and of course these functions can vary enormously. This variation is to a large extent due to the embodiment of the perceiver. What you can do with an object depends, first of all, on what you "can do", i.e. your potential motoric competence. For a human being the primary affordance of e.g. a chair is to sit on it. This is the case because a human being can sit and knows that this type of object is typically being sat on. Returning to Glenberg et al., one may additionally have a certain goal in a given situation, e.g. repairing the lighting of a room, which takes influence on the perceived affordance. Assume, the state of the room is such that it contains a broken light bulb hanging from the ceiling (affording to grab it) and a chair standing in the corner (affording to be sat on or, slightly weaker, to be stood on). Meshing the affordances with respect to one's overall goal to change the light bulb will result in the actions of standing on the chair in order to reach for the light bulb and grab it and exchange it with the new one. Moreover, it is possible to assign new affordances to perceptual symbols by simply imagining a new functionality of the corresponding object. For example, it is perfectly normal to imagine that a wooden spoon may be used to stabilise a small plant, or that a closed umbrella may serve as a walking stick. According to Glenberg, this is only possible because perceptual symbols are modal and non-arbitrary. The idea of assigning affordances, and intentionally creating affordances, has also been promoted by Donald Norman. Norman introduced the notion of (perceived) affordance to design and explains that affordances are not necessarily inherent to an object (e.g Norman (1999)). Rather they are often created purposely by the designer such that the user is likely to perceive the affordance that reflects a possible relationship between object and actor. This is not to be confused with conventions that reflect arbitrary but conventional mechanisms for everyday interaction.

Affordances are a phenomenon reporting the interconnectivity between modal category systems, i.e. motor action and visually perceived objects. They also seem to play an important role in grounding words, perceiving our environment and planing our own next action steps. Ellis and Tucker (2000) have taken an experimental approach to prove their existence and to show that our intentions to act are based on already existing motor representations of possible actions retrieved from the visual scene. They hypothesised that the perception of an object activates the actions that can be made towards or with the object and this again activates the motor schemata needed to do so. But because there are so many possible actions, the associations must be somehow restricted to those that are most highly activated by the situation. They found out that position of the viewer and the perceived object, e.g. in terms of reachability, plays a role in whether an action schema is elicited or not. Furthermore, they discovered that an action-association is not restricted to high-level actions and objects such as writing with a pen or sitting on a chair. They can also occur on a level as low as object parts and a certain type of grasping action requiring a specific hand-shape, so-called micro-affordances.

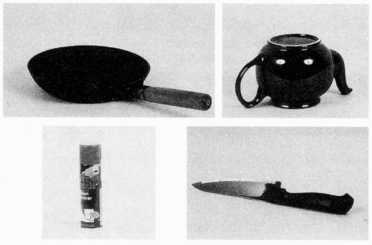


Figure 2.4: Experimental items from Ellis and Tucker (2000)

Ellis and Tucker give evidence for their hypothesis, among others, in the following study. Photographs of graspable objects as in Figure 2.4 are presented individually to the subjects. The objects can be distinguished with respect to upright and inverted position and to the orientation of the handle either towards the right hand or the left hand. The subjects were told to push a button at sight of the object as fast as possible, choosing the left or right hand depending on the upright/inverted position of the object. As expected, the left-right orientation of the handle facilitated or inhibited the reaction time of the hand that was used to give the push-response, i.e. when the handle was oriented such that it was easiest to grasp it with the right hand, a push-button-response with the left hand (e.g. for upright position) was slightly slower then the reaction with the right hand (correspondingly, for inverted position). This study shows that the sight of an object potentiates an action that is irrelevant for response determination. This action is generally highly associated with the object and in particular facilitated by the position of perceiver and perceived towards each other. The perception of affordances as typical functions of an object also influences the way we understand language. Carlson and Kenny (2005) argue that recognising a relevant function of an object constrains the wide range of interpretations of spatial language. The utterance "Put the cup *under* the tea pot" when the speaker is about to provide tea to the hearer, is usually interpreted as holding the opening of the cup underneath the muzzle of the tea kettle.

There are, of course, many more experimental approaches to examining affordances and human categorisation in general, partly with very distinct methods. Some of these

methods may involve recording the state of the brain of a subject in order to establish a neural record of a perceptual states. For instance, Helbig, Graf and Kiefer (2006) found evidence for affordances that challenge the classical view of the two separate neural pathways (*ventral* and *dorsal*). According to their studies information from both streams is integrated continuously and much earlier than assumed so far. This just emphasises the need for an interconnection mechanism between category systems when modelling those. Similarly, Rizzolatti and Arbib (1998) found a strong connection between observing an action and performing that same action on the neural level. In fact, the neurons they identified in monkeys (so-called "mirror neurons") fired equally upon perception and execution of a specific action. Since these neurons are located in the apes' analogue to humans' Broca area (that is mainly involved in language processing), Rizzolatti and Arbib concluded that this match between observation and execution provides the basis for communication. They argued that such a mechanism enabled humans to start exchanging messages expressing something about joint observable objects or events. Essential for us is that their studies reveal a cross-modal link between the motoric pattern of an action and the visually perceived action. Hence, it is important to investigate not just *that* categories mediate, but also *how*. How is a certain kind of interconnectivity achieved such that categories can activate each other (e.g. as in the case of affordances, where a perceived object triggers an action category) and, thus, mediate content in order to achieve situational awareness?

2.3 Related work on Interconnectivity

In this section, we briefly introduce some work on connectivity between modalities on the neural level, particularly focussing on language processing.
There seem to be quite different mechanisms involved in combining and coordinating multi-sensory input. Engel and Singer (2001); Singer (2003) present studies that suggest two distinct types of connectivity, convergence and coordination/association. Features basically are encoded by convergence cells, i.e. there are modality-specific cells that fire upon sensing very specific properties. Then there are cells which react on a specific combination of the 'feature' cells, so-called convergence cells. They are assumed to represent a certain common feature or object component. Whole objects, however, are represented through synchronous firing of the according convergence cells that may belong to any modality. This temporal (and possibly spatial) proximity is a flexible method for coordinating a large number of possible features that easily allows, for instance, new combinations for new objects. Calvert (2001) even report of multi-modal integration on a neural level with respect to higher cognitive processes such as speech comprehension and visual perception. Their study revealed that brain activity indicating integration of multimodal input takes place while seeing and hearing a speaker. In contrast, hearing a speaker but seeing incongruent lip movement expectedly drops integration activity. Calvert et al.

leave open whether the integration at that level is achieved by convergence or association via (spatio-)temporal binding as suggested above. Hickok and Poeppel (2004), on the other hand, emphasise the distinction of (mainly visual) perception processing into dorsal and ventral streams, also called "where" and "what" pathways respectively. Traditionally, in the ventral stream object identity is processed, whereas the dorsal stream processes spatio-temporal organisation and according to Hickok and Poeppel also manages visuo-motor integration. They argue that such a distinction can also be made for auditory perception and that the dorsal stream is responsible for audio-motor integration. They suggest that there are bi-directional connections between both pathways in vision and language. These connections are suggested to be implemented by networks mediating between identified entities and their spatio-temporal insertion.

Another phenomenon providing us insights on connectivity on higher, more categorical level, are affordances. Mecklinger, Gruenewald, Weiskopf and Doeller (2004) have found neural evidence for affordances and their effects. They have investigated their role for visual working memory. Their subjects had to view pictures of objects (manipulable and non-manipulable) while their brain activation was recorded through functional magnetic resonance imaging (fMRI). Initially, the subjects had to memorise the objects. The result was that the retainment of manipulable objects in working memory caused activation in the hand region of the ventral premotor cortex. Other objects, however, caused activation in a different part of the brain. This again is evidence for a tight connection between objects and their associated actions and it seems that the activated motor schemata are integrated with their cue which obviously determines storage in working memory.

The presented work in this section indicates several connection mechanisms that apply at different cognitive levels to integrate perceptual information from different modalities. On a low level we assume convergence to a certain degree such that single cells react to specific features. On a higher level one observes integration of features from the various modalities. This includes, e.g. visual object features and components, but also features specifying spatial properties or properties on the auditory or lexical level. How this integration takes place is yet unclear. Suggestions include spatio-temporal synchronicity as binding mechanism, e.g. to compose objects from features or to associate actions to objects, and convergence as suggested for the lower level.

2.4 Discussion

Most psycholinguistic studies use depicted not real natural scenes which may influence what is considered to be relevant and, therefore, where one looks. Real scenes contain even more sources of information than just the 2-D visual one. What humans can do with their body and how they learned to relate to particular objects (experience, convention) at least plays a role in determining relevance of objects and events for a given task. In the above setting, where an object is referred to and required to be put some-

where, reachability is a feature which may influence the saliency of the possible target objects. Furthermore, the experiments are typically performed with subjects having eye-trackers mounted to their heads, this may also alter the natural saccades. Nevertheless, these studies show a tight semantic and temporal interplay of modal-specific input, particularly vision and language, in order to compose meaning. The psychological studies have helped us to conceptualise what a category is, what it denotes, that it needs to be grounded and how it can be used in a category system to mediate content. The insights into the nature of affordances, in particular, were of great interest to us and from both, psychology and neuro-science, provided the fundamentals for our model of a category system.

2.5 Conclusions

In this chapter we have presented studies that provided us with insights into perceptual processing underlying gaze in humans. Essentially, the findings contribute to understanding gaze and its functions for relating utterance comprehension and visual scene analysis and for further refinement of both. We have further introduced a number of studies observing the mediating nature of categorical understanding and presented their resulting hypotheses about grounded category systems. Together with the need for an interconnection mechanism between the categorical structure, we obtain an integrated hypothesis about which principles human gaze production underlies. The purpose of the following chapter is to identify the requirements for a model of robot gaze production that is based on the integrated hypothesis outlined in this chapter.

Chapter 3

Requirements and Model

Summary

If we want to create a model of how a robot can produce anticipatory gaze during utterance comprehension, then what requirements should such a model address given the observations in Chapter 2? Below we first discuss the requirements, then the model we propose to address these requirements.

3.1 Requirements

Studies like (Kamide et al., 2003; Altmann & Kamide, 2004; Glenberg & Kaschak, 2002; De Vega et al., 2004) all point to the need for *mediating category systems*, underlining the more philosophical considerations of Lakoff (1987); Glenberg (1997); Barsalou (1999). These category systems mediate in that they *connect* sensorimotoric signals and higher-level cognitive processes such as language processing, and *raise expectations* on the basis of categorical content that gets activated by virtue of being connected to other modalities. Following Endsley (2000), category systems thus play a core role in situational comprehension and projection.

For category systems to fulfill this role, we need to address the issues of *activation* and *cross-modal interconnectivity*. Category systems need to model how different types of categories, like actions, objects, features of objects, are related. For example, a put-action typically has an animate object performing the action on another object, possibly with a destination specified for where the affected object should be put. Once we activate one concept, we need to have mechanisms for *spreading activation* to activate (or inhibit) related categories – on the basis of which we can then formulate mentioned expectations; cf. also (Anderson & Lebiere, 1998). Activation of categories, and its effect, depends on the interconnection between category systems, and cognitive and sensori-motoric modalities. For one, categories get activated based on how we can connect content across modalities. Second, the effect of category activation on selective attention in other modalities can only arise when content is connected to, i.e. *grounded in*, the category system.

Thus, cross-modal interconnectivity underlies how comprehension and projection can arise in situational awareness - and, more fundamentally, how we can achieve "*symbol*

grounding" in a mediated sense. There are several important functions that cross-modal interconnectivity should provide. It should facilitate the exchange of content, so that we converge or associate content across modalities (Calvert, 2001; Singer, 2003). In addition, cross-modal interconnectivity needs to facilitate different types of cognitive control: the spreading of activation from one modality to another so that activated categories can influence selective attention in other modalities, and monitoring and resolution of conflicts between content in different modalities. (We leave open whether cognitive control should be seen as a set of separate mechanisms, or as an emergent phenomenon, cf. Botvinick, Braver, Barch, Carter and Cohen (2001).)

If we want to use preferences (through spreading of activation across modalities, modulo cognitive control) during on-line utterance comprehension then this requires an incremental process. In this process, we gradually build up one or more representations, maintaining them in parallel but with a preference order over them to indicate which is (currently) the preferred interpretation. We would like to argue that the requirement of parallel processing for gradual refinement does not only hold for *linguistic* representations, but for *most if not all* content representations maintained in a system. This follows from the observation that we not only have situational awareness influencing how we comprehend an utterance, but also that what we talk about influences how we try to understand the situation – the interplay as observed in the experiments in Chapter 2.

Finally, based on the graded, partial representations we have built up so far of the linguistically conveyed meaning, we need to be able to derive where to turn our gaze. This requires the partial representations to indicate *types of open arguments*, and the *order* in which argument positions can be expected to be filled. For example, after "The woman puts .. " we first expect a mention of the object being put somewhere, before a mention of the destination. Furthermore, if possible, we need to determine where in the scene candidate references would be (modulo following deictic gestures made by the speaker), and direct gaze to the most likely candidate(s). This brings us back to cross-modal interconnectivity: content should not only be associated if it has already been seen or mentioned, but also if it concerns possible extensions (content-wise, but also spatio-temporally - cf. the observations in e.g. (Kamide et al., 2003; Altmann & Kamide, 2004; Knoeferle & Crocker, 2006).

3.2 The Model

The model we propose here addresses several of these requirements. Because we focus in this thesis primarily on grounding in categorical aspects of situational awareness, we only mention spatio-temporal aspects in passing. The overall model is inspired by functional-biomimetic concepts of architectures for situated language processing proposed in e.g. (Calvert, 2001; Hickok & Poeppel, 2004), discussed in Chapter 2. We discuss the three principal components here: *category systems*, *incremental utterance*

processing, and *cross-modal interconnectivity*. Finally, we explain how this model is used for robot gaze production.

Category Systems

We model category systems as a collection of *distributed ontologies*. The ontologies model the world knowledge and contain modality-specific categories. An ontology inherently provides convergence mechanisms through its taxonomic nature. Therefore, within an ontology, categories can principally be related using *sub/super-concept* relations (IS-A) and *part-whole* relations (HAS-PART). The former relation is slightly different from what we call convergence, it yields more general categories rather than a composition to another, possibly more general, category. The latter can be achieved by the (HAS-PART) relation that we employ to obtain a hierarchical organisation of visual objects. However, this is only one possibility to compose entities. Another is simply feature composition (Singer, 2003) that we also provide by including anonymous classes in the ontologies that are characterised by the presence of certain properties or features.

Across ontologies (and modalities) we can associate categories through *thematic relations*. This takes general cross-modal association mechanisms into account as proposed by Calvert (2001) and is in accordance with Glenberg's hypothesis on the complexity of associations (Glenberg, 1997). These relations enable us to provide a basic model of *affordances*. For example, by relating a type of object with a manipulative action we can indicate that (a) an instance of the object allows for interactions using this type of manipulative action, and (b) the action affords manipulation involving this type of object. (At the level of basic-level categories, this can be seen as modelling *micro-affordances* in the sense of (Ellis & Tucker, 2000).) However, this relation need not be symmetric. As Ellis and Tucker's studies reveal, an object can elicit a very specific action. That same action in a different context may not necessarily activate that very same object or object part. More often, one would associate a whole class of objects that could generally be involved in the action process. This also depends on how "universal" the action is, e.g. compare "see" and "fry" and their range of target objects. The associations between categories are weighted, and affect which categories get activated. We handle activation using a working memory, in which we store the activated category, its activation levels and a pointer to the cue and target category definitions in the ontologies. That pointer also contains the information on thematic roles if actions are involved. The design of the system meets the requirements discussed by Barsalou as described in Chapter 2. For instance, the ontologies handle the type-token distinction and we can draw inferences from categories to their properties or related categories. Furthermore, some categories can be combined to form more complex categories, e.g. via *part-whole* relations. The categories can always be mapped onto the situation or used indirectly to influence acting in that situation, e.g. associations priming a linguistically processed action command. However, we do not deal with abstract categories so far.

Incremental Parsing

We use *Combinatory Categorial Grammar* (CCG) (cf. Steedman (2000)) to model incremental utterance comprehension. CCG is a grammar framework in which we specify, for each word, how it can syntactically combine with other words (its arguments) to form a grammatical expression with a particular meaning. Consider the example below.

(1) put :-
$s_{t1}/pp_{x1}/np_{x2}$:
$@t1 : action(put \wedge$
$\langle$Mood$\rangle$**imp** $\wedge$
$\langle$Actor$\rangle(r1 : hearer \wedge$ **robot**$) \wedge$
$\langle$Dir:WhereTo$\rangle x1 : location \wedge$
$\langle$Patient$\rangle x2 : thing)$

Example 1 illustrates the lexical entry in the grammar for the word "put." First, we have the *category* that states that "put" can yield a sentence (s) if we combine it first with a noun phrase np to its right, and then with a prepositional phrase pp to its right, to yield the order "put np pp". Furthermore, we have a specification of the meaning of "put." "Put" is specified as an action, that has an Actor ($r1$) performing the action, a Patient ($x2$) which is the object being taken, and a location ($x1$) specifying where the object should be taken to. By marking the syntactic arguments (pp, np) with the indices of the semantic arguments ($x1, x2$), we *link* the two levels of description. Thus, the np provides the meaning for the object, and the pp the meaning for the location (Baldridge and Kruijff (2002)). Previous approaches to incremental parsing include work by Hepple (1991) proposing the dependency calculus as most appropriate framework and earlier work by Steedman et al. using CCG. Steedman (2000) describes how we can incrementally parse an utterance with CCG grammars. We gradually build up a representation of the syntactic structure and the meaning of an utterance by analysing it in a "left-to-right" fashion. The partial parses are stored in a chart and may contain several (ultimately ranked) variants. At each step, argument positions are saturated in the direction as specified by combining each of the existing parses with the parse of the argument. E.g. compare to Example 2 so far being unambiguous with "put" now having a saturated patient role.

(2) put the ball :-

s_{t1}/pp_{x1} :
$@t1 : action(put \wedge$
  $\langle$Mood$\rangle$**imp** $\wedge$
  $\langle$Actor$\rangle(r1 : hearer \ \wedge$ **robot**$) \wedge$
  $\langle$Dir:WhereTo$\rangle x1 : location \wedge$
  $\langle$Patient$\rangle(x2 : thing \ \wedge$ **ball** $\wedge$
    $\langle$Number$\rangle singular \wedge$
    $\langle$Delimitation$\rangle unique \wedge$
    $\langle$Quantification$\rangle specificsingular))$

Example 2 also illustrates that we still have unsaturated argument positions in the phrase. The type of $x1$ indicates what kind of expression to expect next, and through its linking with the semantics, what the expected kind of meaning of that expression is. If the next phrase is e.g. "on the cup", then this would be an appropriate filler and we get a saturated phrase corresponding to the sentence: "Put the ball on the cup." However, it can also be interpreted as a modifier of the noun phrase "the ball" which would result in a different parse and a yet uncomplete sentence. The parses are both represented in the chart and may be ranked according to other available information from e.g. the visual scene. Note, that in our analysis, we primarily focus on utterance-level comprehension here making little or no use of its (simultaneous) integration with discourse-level comprehension. This, of course, would provide us with more information but also pose more problems. Referent resolution, for instance, becomes more complex given that more options are provided to chose from. Definite noun phrases, for instance, may identify a discourse referent ("the ball I just talked about") or a uniquely identifiable visual object ("the only visible or visually most salient ball"). There, one needs to deal with visual and linguistic saliency and preference over both.

Interconnectivity

We use *ontology-based mediation* to connect content across modalities (Gurevych et al. (2003); Kruijff, Kelleher and Hawes (2006)). First of all, each modality directly links into its own conceptual structure. However, patterns across these structures may be quite similar, e.g. we often visualize an object when we hear the name of it. Or we vaguely feel the meaning of an action (e.g. grab) when we read the according term. Many concepts have variants in each or at least some modalities, depending on whether the concept is perceivable by the corresponding sensors. Instead of assuming that these concept variants converge to one amodal concept that is no longer grounded, we propose a way of cross-modally interconnecting these concepts (Calvert, 2001; Glenberg, 1997). There are two mechanisms achieving this: a) ontological relations link one concept to another, e.g. the property hasShape for any visual object ranges over concepts in the features ontology; b) there are arbitrary associations between any two concepts that have often occurred

together in some way or another. The former can be considered a variant of convergence providing a compositional mechanism for creating and characterising objects as well as actions to a limited extent. The latter typically are formed between actions and objects, and thus concern affordances.

3.3 Gaze production

We argued that gaze is based on situated awareness which shows in many psycholinguistic studies. We further looked into the mechanisms that create situated awareness and identified a special type of category system as the key component that has to mediate contents between modalities. Given the model we proposed for such a system, what are the effects on gaze production? Essentially we predict gaze behaviour that is similar to human's gaze in settings like the 'visual world'. These simple scenarios typically contain few objects as referents and an action that involves some of the referents. We would like to evaluate the produced robot gaze in that kind of scene. However, we use and test the robot in a real 3-D environment instead of constraining it to 2-D pictures. We create a table top scenario where the robot needs to interact with a human in order to learn, act and talk about a given situation. The two main gaze effects we wish to produce are the following. When the robot processes an ambiguous (partial) sentence, e.g. with respect to a PP-attachment where the prepositional phrase can be interpreted either as modifier or as location/destination, it should look for potential referents in the scene. The visual information on potential referents should be used to disambiguate the sentence structure if possible and produce a preference over the partial parses. This, in turn, can raise an expectation of what is to be perceived next. That is the case if the preferred parse still has at least one open argument. The lacking information for making sense of the utterance has a specific type, e.g. *patient*. Since in many of the psycholinguistic studies two competing sources of information for raising expectations for the completion of linguistic utterances have been identified: world knowledge and the visual scene (e.g. Knoeferle and Crocker (2006)), the system has two options. It can retrieve a typical patient for the action from its association collection if possible and it can scan the scene to extract a potential patient from it. Thus, the predicted argument type directs the robot's gaze in an anticipating manner to find a visual match for the expected linguistic completion. Which source, either world knowledge or visual information, is generally preferred by humans cannot be told with certainty yet. We have decided for the robot to give preference over visual stimuli since its associations are still hand-crafted and small in coverage.

3.4 Conclusions

The model we have presented in this chapter integrates the properties and issues we have raised in Chapter 2 to a large extend. It comprises a category system in the style of percep-

tual symbol system proposed and supported by Barsalou, Glenberg and others. It enables the representation of affordances as relations between actions and objects involved in these. Thus, we follow suggestions by Gibson, Ellis and Tucker, Carlson and other studies on functionality of objects and their implications for situated awareness. The design of the system allows content mediation within and across modalities based on category connections. This mechanism is the prerequisite for higher level interaction, the effects of which can be observed in the use of gaze. In the next chapter we present an implementation of the model. Finally, to show that the model accounts for some aspects of gaze in the way we have predicted here, we present a sample run of the system explaining the interactions step-by-step in Chapter 5.

Chapter 4

Implementation

Summary
In Chapter 2 we described several studies illustrating how humans perceive and integrate visual and linguistic information and presented various studies that attempt to explain these mechanisms in more detail. On the basis of that, Chapter 3 reported the requirements and the model for robot gaze productions that draws on the findings in human gaze production. In this chapter we propose an implementation of the model that integrates a multimodal category system with visual scene analysis and incremental utterance comprehension to produce gaze. We begin by providing an outline of our implementation of the category system. Our main tools are an OWL-ontology, a reasoning system called "Racer" to handle assertions and draw inferences within the ontology, and a collection of weighted pointers between categories of the ontology. Furthermore, crossmodal activation patterns like affordances are shown to be modelled in principle. Subsequently, we introduce the visual scene analysis we employ and the incremental utterance parser. We explain how the category system mediates between language and vision and thus produces gaze behaviour that is grounded in situational awareness.

4.1 Tools

4.1.1 Knowledge Representation

To represent knowledge about the world such that an artificial system can access and use it for recognising and reasoning about its environment, knowledge needs to be structured and formalised. An ontology is a system that captures knowledge in one domain (due to feasibility) in a formalised way. It provides descriptions of categories and relations between them, typically these are taxonomic relations that give an ontology its hierarchical character. We have decided to use OWL [1] to represent our world knowledge because it is a rich ontology language that provides a larger vocabulary than other languages like Resource Description Framework (RDF) Schema [2] for the characterisation of classes and

[1] http://www.w3c.org/TR/owl-features
[2] http://www.w3.org/TR/rdf-schema/

properties and the relation between classes such as disjointness of classes. OWL comes in three different types, ordered according to their expressiveness:

- OWL Lite, being the most restricted and least expressive version,
- OWL DL, which corresponds to description logics (hence its name) and is most powerful while still being complete and decidable,
- OWL Full, which has maximum expressiveness but is currently not automatically reasoned over in a complete way (McGuinness & Harmelen, 2004).

OWL DL is the best suited version for our purposes since an automatic reasoning mechanisms can be applied while still having as much expressiveness as possible to try and model the infinitely complex (real) environment.

Since OWL DL is based on description logics (DL) [3] we will briefly dive into its characteristics. The name of OWL DL refers to the use of concepts to describe a domain and to the logic-based semantics that is obtained by the translation into first-order predicate logic. The syntax is composed of a set of unary predicate symbols that denote concept names (i.e. the `classes` in OWL), a set of binary predicate symbols that denote role names (i.e. the `properties` in OWL) and a recursive definition for defining concept terms from concept names and role names using constructors. A concept (or as we call it, a category) denotes a set of individuals that belong to it and a role denotes a relation between two concepts. It is important to note the distinction made between individuals and concepts as they lead to the distinction in DL between the so-called **TBox** (terminological box) and the **ABox** (assertional box). The TBox contains information about the taxonomic structure of the ontology, i.e. what is a class of what and how are they related. The ABox, on the other hand, contains assertions on individuals, i.e. it provides the connection of the principled knowledge to real world situations.

In OWL terms (cf. Horridge (2004)) we are dealing with `classes`, `properties` and `individuals` and keep the TBox and ABox distinction to model the robot's knowledge about the environment in general and situations in particular. To compose and edit our OWL ontology we have used the Protegé editor [4] that provides a graphical user interface to all OWL functionalities, and among others a plugin for inferencing with automatic reasoners (cf. below). The classes we work with are also described below in detail and illustrated in graphs. What cannot be depicted so easily is the internal structure of the ontology. As already mentioned, OWL provides a variety of properties, essentially `DatatypeProperties` and `ObjectProperties`, and several ways of relating classes. The former are distinguished according to their value, i.e. `DatatypeProperties` point to primitive data types such as integers, strings or

[3] http://dl.kr.org/

[4] http://protege.stanford.edu/overview/protege-owl.html

booleans and `ObjectProperties` establish a relation to another class of the ontology. For some examples, consider the constructs based on and similar to RDF Schema below (the prefixes in angle brackets indicate that the terms have been introduced before OWL by RDF or RDF Schema).

```
<owl:DatatypeProperty rdf:ID="isReachable">
   <rdf:type rdf:resource="http://www.w3.org/2002/07/
                    owl#FunctionalProperty"/>
   <rdfs:domain rdf:resource="#Reachability"/>
<rdfs:range rdf:resource="http://www.w3.org/2001/
                      XMLSchema#boolean"/>
```

This creates a data type property called isReachable which is functional, i.e. it can only be applied once and yields only one value. Its domain is the Reachability-class, meaning it can be applied to instances of that class, and its range is a boolean value such it evaluates to either true or false. Below is an example of an ObjectProperty showing the property isPartOf that is also functional and further has an inverse property hasPart. The property is a sub-property of Part. Its domain is the visual object-class Handle and the range contains Mug and Cup (that are currently the only categories involved in part-whole relationships of objects due to our very restricted scenarios with automatic object recognition).

```
<owl:FunctionalProperty rdf:about="#isPartOf">
   <rdfs:domain rdf:resource="#Handle"/>
   <rdfs:range>
      <owl:Class>
        <owl:unionOf rdf:parseType="Collection">
          <owl:Class rdf:about="#Mug"/>
          <owl:Class rdf:about="#Cup"/>
        </owl:unionOf>
      </owl:Class>
   </rdfs:range>
   <rdf:type rdf:resource="http://www.w3.org/2002/07/
                    owl#ObjectProperty"/>
   <rdfs:subPropertyOf rdf:resource="#part"/>
   <owl:inverseOf rdf:resource="#hasPart"/>
</owl:FunctionalProperty>
```

Some possible connections among classes are illustrated in our next example showing the complex class Cube. This explicit class is equivalent to the anonymous class created by the intersection of a number of restrictions. These restrictions determine that if an

instance a) has the functional property hasCubicShape with the range CubeShape and b) it belongs to the class VisualThing and c) it is not a sub class of Containershape, then it is also an instance of the explicit class Cube. This and other conclusions can be drawn by an automatic reasoner on the basis of the given assertions. It is also for reasoning purposes that we have included very specific roles like hasCubicShape but this shall be discussed in Chapter 5.

```
<owl:Class rdf:ID="Cube">
   <owl:equivalentClass>
     <owl:Class>
        <owl:intersectionOf rdf:parseType="Collection">
          <owl:Restriction>
            <owl:someValuesFrom rdf:resource=
             "#Cubeshape"/>
            <owl:onProperty>
              <owl:FunctionalProperty rdf:about=
               "#hasCubicShape"/>
            </owl:onProperty>
          </owl:Restriction>
          <owl:Class rdf:about="#VisualThing"/>
          <owl:Class>
            <owl:complementOf>
              <owl:Class rdf:about="#Containershape"/>
            </owl:complementOf>
          </owl:Class>
        </owl:intersectionOf>
     </owl:Class>
   </owl:equivalentClass>
</owl:Class>
```

4.1.2 Reasoning

To make use of the captured knowledge, a system needs to be able to evaluate assertions given a number of already existing facts in the knowledge base. It needs to be able to draw inferences combining grounded facts (assertions from the ABox) and axiomatic or prototypical knowledge (contained in the TBox) and update the knowledge base accordingly. In summary, a system has to be capable of reasoning over the provided or acquired knowledge otherwise it is lost in a dynamically changing environment. To execute this task we are using the OWL reasoner RACER [5] in our system (cf. Haarslev and Moeller

[5] http://www.racer-systems.com/

(2001)). To explain how RACER works, we are again considering some examples: Racer is used to deal with queries of the kind

```
(individual-instance?  mugXYZ Containershape)
```

which asks whether the specified mug is an instance of the concept Containershape. Since *mugXYZ* is an instance of the concept Mug that is a sub-concept of Containershape, Racer returns true. Another possibility is to reason over relationships between categories, e.g. when the system recognises a specific instance of a category then it makes sense to check wether typically instances of that category have certain roles to other categories. That means, one first has to know what roles there usually are for the category and second what is the range, i.e. the target categories, for each role. Since prototypical reasoning is difficult with RACER we use the following strategy assuming the role in question is the hasPart-relationship. First we check whether the category in question, e.g. Mug, is contained in the domain of the role by retrieving the whole list of cue categories.

```
  (role-domain hasPart
& optional (<TBox name> (tbox-name *current-tbox*)))
```

If Mug is contained in that, one can continue to check what categories are typically associated with the source via this role by asking for its role range.

```
  (role-range hasPart
& optional (<TBox name> (tbox-name *current-tbox*)))
```

The resulting categories can themselves activate more associations and are stored in working memory. However, the role range in OWL is not specifiable for one single category (as opposed to a single instance) and, thus, always yields more categories than initially desired.

4.2 The categories

Our ontologies encode world knowledge that the embodied system is assumed to have gained somehow. We use several ontologies, basically for each modality one, that are all subsumed under the OWL general type *Thing*. Thus, it may appear as one ontology but, in fact, they are more or less separate and distinct. The advantage of having them all inherit from the same OWL-Thing is that we can easily employ roles to relate between categories. Since our sample scenarios will be very simple and restricted to start with, the ontologies are equally simplistic, containing only objects found on a table top as well as actions typically applied to such objects. The ontologies for language and vision strongly resemble each other which is due to the fact that usually one knows what an object or action one sees is called and, vice versa, humans often employ mental imaging upon hearing

a word referring to an object. (Do you picture anything in particular when I talk about a blue elephant with yellow dots on it?) This, of course, implies that I know generally what is being talked about. If I have never seen a mug, for instance, then the word 'mug' does not necessarily evoke a mental image. However, in the case of the blue elephant, one has enough conceptual knowledge about elephants and colour that one instance can be pictured even though such a creature has never been encountered. In our ontologies we want to focus on concrete things and only associate a visual category with a linguistic term if the system has experienced it that way. Furthermore, we feature a very small ontology for motoric action schemata and a taxonomy for multimodal features. The reason we created a separate branch for features is the additional complexity of information for the other categories. Since OWL does not support anything like primitive properties one needs to instead create roles that always need a target concept as range filler. In the following we take a closer look at the formal details of our ontologies. For practical reasons and clarity, category names from vision start with a "V", those from the motor system start with an "M" while linguistic category names are just plain.

4.2.1 Language

It is not obvious at all that there have to be separate ontologies for vision and language. But as already mentioned, we need some means to distinguish between the word for an object and the image of it because it is possible to only know either of it. Additionally, we have evidence for the vision and language ontologies having separately stored categories from split-brain patients revealed by Roger Sperry (winner of the nobel price for medicine in 1981). Confronted with an image only visible in their left visual field, the patients can find matches with similar pictures but cannot retrieve the name of the depicted object. Obviously there are two modal category systems that, in these cases, are no longer properly connected and can therefore not be used interchangeably. Consequently, the language ontology contains all sort of categories for objects, actions, properties etc. We exemplarily included the names of some motoric action schemata but really only make use of the object names as indicated below.

As depicted in Figure 4.1 the objects contained in the linguistic ontology are names for manipulable objects. At this point we revert to basic-level categories and claim that the type of categories used in our ontology are chosen according to Rosch's principles, i.e. they are basic and uniform for humans in perception (mental image for whole category is possible), function (same motor interactions), communication (short, frequent names) and knowledge organisation (Rosch, 1978). We further claim that, at this stage, this also applies for robots and therefore we do not go into more detailed or specific linguistic object names. For the generation of referring expressions we include properties and spatial relations of the object in question and do not need more specific names.

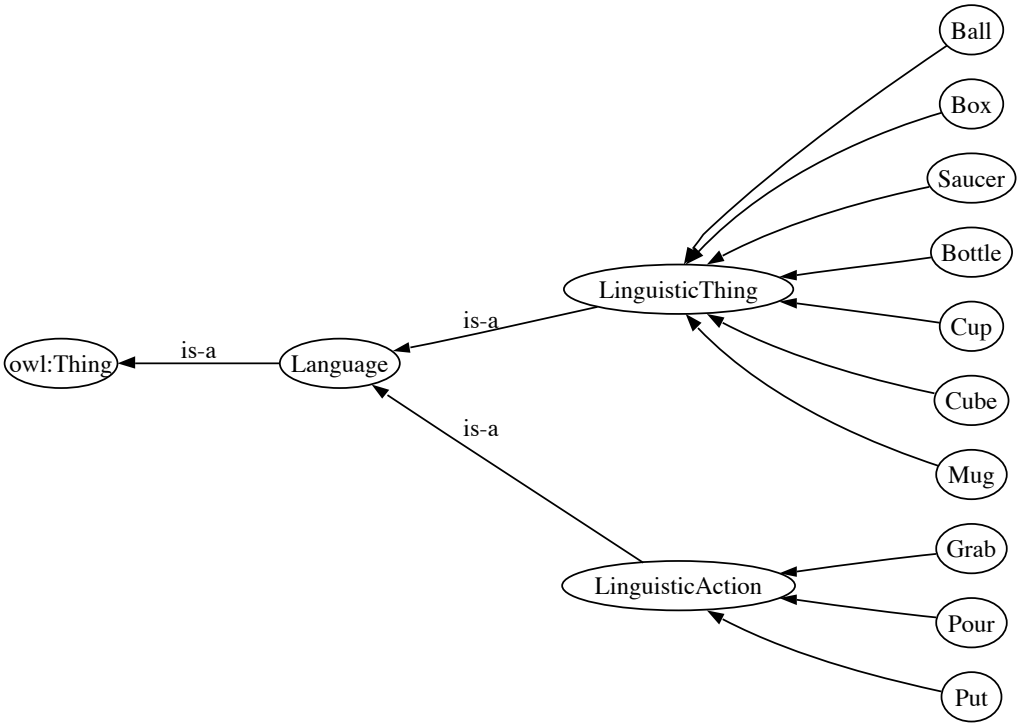


Figure 4.1: Linguistic Ontology

4.2.2 Vision

The visual ontology is slightly more complex than the linguistic one. In fact, it should contain much more information! We presume that most of the (category) learning is done pre-linguistically and, moreover, that our visual category system is much more refined and complex than what one can express about the state of the environment verbally. We can hardly account for this complexity and richness of visual perception because neither our technical means for visual scene processing are sufficiently advanced nor do we know enough about how humans perceive and store 3-D visual entities. For practical purposes we have thus constrained ourselves to the recognition and representation of the same simple manipulable objects as before. Merely few additional roles such as "partOf"/"hasPart", composing a little more complex hierarchical objects, and "contains"/"isContainedIn", indicating a slightly more refined configuration, pay tribute to these circumstances.

The handle, for example, is part of a mug and a cup. Of course, the part-relation is symmetric and features the inverse isPartOf. Boxes, mugs, cups and bottles, on the other hand, may contain some visual thing themselves. However, a bottle, mugs and

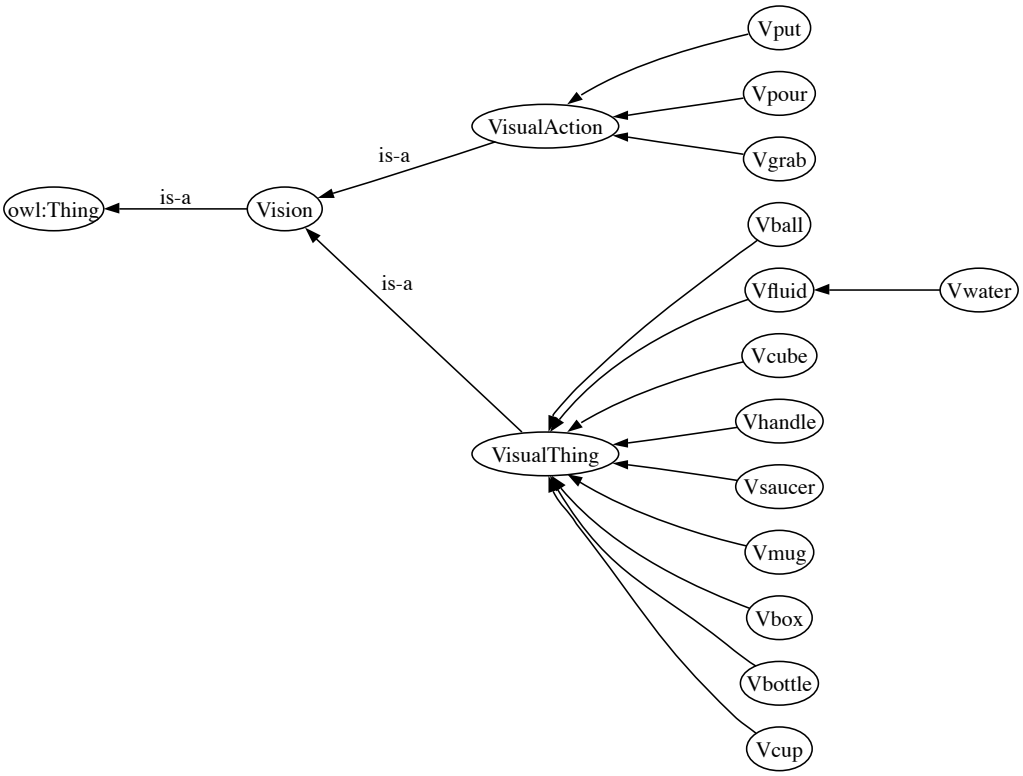


Figure 4.2: Visual Ontology

cups may contain only fluids and boxes may even contain things of larger size - whatever the relative size is precisely. When reasoning about categories and associations we are dealing with prototypical relations only. A cup or mug would be considered of large size such that it can be placed and retrieved from a box. This relation is realised via the container-role that subsumes the symmetric contains- and isContainedIn-roles. There again we distinguish between container roles for fluids and larger visual things. This strategy ensures that we do not get the full range of a container-role which would yield more categories than wanted. Since there seems to be no other way of specifying the role range for a specific category, we have decided to go with more specific roles instead.

Our visual objects additionally have certain shape properties expressed by the general role hasShape that ranges over all visual features. The features are introduced below. Here it is only relevant to note that visual objects need to be describable, and to some extend uniquely describable, by their visually perceivable shape such that inferences can be drawn as towards what an unknown object may be. Of course, these ontologies (vision and features equally) are subject to constant deformation and extension and because they

heavily rely on learning in and experience with the real world. They will naturally be very valuable when it comes to intelligent computer vision and object classification that is influenced by cultural and linguistic constraints. Again, in our case we are lacking the technical means and have to fall back on exemplary visual categories and very simple object recognition.

4.2.3 Features

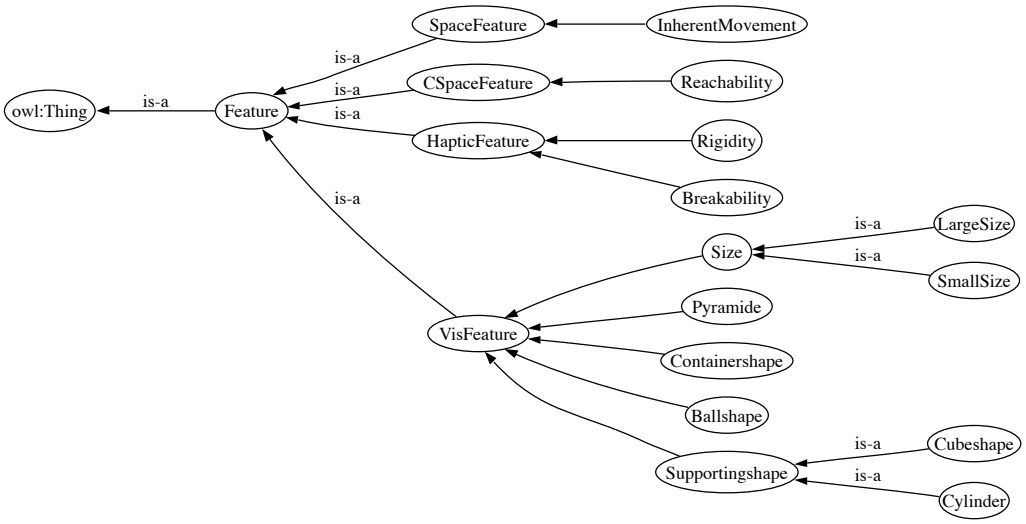


Figure 4.3: Feature Ontology

The feature system has been introduced because classes (representing categories) in OWL are flat and trivial. But obviously objects have many properties humans perceive through several sensors. Apparent are the visual features that you perceive by merely looking at an object, e.g. size and colour and shape and that it contains something or is a part of another object. The distinction of parts, however, is closely related to what functionalities one perceives, as we have described in Chapter 2. Humans have also learned to judge by sight whether an object is suitable to support another object so we included a feature determining whether something can be put on top of the corresponding object. Moreover, one can visually estimate the distance to an object and it may very well be of relevance whether an objet is within reach. We assigned the reachability property to the class of configuration space features instead of visual features because it is, although perceived visually, dependent on the body of the perceiver. Further distinctions with respect to the degree of reachability, e.g. *direct, through own movement, through using helping devices, not at all*, are conceivable. Other features refer to haptically sensing an object, thus specifying the material and therefore the rigidity and breakability of the

object. Ideally, our robot will be able to explore an entity in such a way using its arm and gripper in the near future. Moreover, we included a feature class that determines how an objects behaves in space with respect to its position. Imagine a pyramid and the numerous positions it can be put in. Some are more stable than others, in the worst case it will even fall into a more stable position right away. A ball, on the other hand, is never in a quite stable position but is inherently moving since its only touching surface is a point - which is very small.

4.2.4 Actions

The action ontology contains categories for rather abstract motor schemata of varying complexity. There are so-called action primitives like Grab and Release or Reach that consist of a single action like closing and opening the gripper and approach an object with the gripper. Other actions like Put are rather complex and imply the execution of several actions in a specified order. In our OWL-ontology we can specify the decomposition of actions to a certain extent. For instance, Put implies Reach,Grab,Lift,Move,Release. Further details with respect to how precisely the actions have to be executed are a matter of motor planning where concrete action schemata precisely fitting the situation are generated. That module can create a sensible sequence of actions to fulfill a given goal like "put the cup onto the cube". The categories used in the ontology, on the other hand, are currently abstract and used purely for mediation. Ideally, there will be a closer connection via the motor planning module into the respective modalities (arm, hand, legs/wheels) once these are physically available. The sample command above ("put the cup onto the cube") illustrates an additional aspect of actions. The way people linguistically treat actions - in terms of verbs - is closely related to the structure of an action in the real world. That is, syntactically verbs have sub-categorisation frames providing slots for objects and semantically they get assigned thematic roles to them (cf. Fillmore et al./ Baker, Fillmore and Lowe (1998) and the FrameNet approach). The thematic roles evolve from experience and reflect what one really needs to know for executing an action: who shall execute it; to which object shall the action be applied to; who or what is the recipient; and what is the original location and what the goal destination? Particularly the last question is very much dependent on a profound spatial interpretation of the scene, possibly some understanding of causal relationships and choosing the right frame of reference.

Since the knowledge on thematic roles seems to be essential for truly understanding an action we have enriched the associations between objects and actions with information on their interaction, e.g. agent-role, patient-role, goal-role etc. The latter is an example for encoding spatial relations between the relevant objects typically expressed by prepositions in linguistic terms. In modelling complex associations between categories we partly follow Glenberg (1997). We depend on a mechanism, however, to extract the specific situated meaning of a preposition to grasp the specific meaning of the action. Currently, we have to assume this type of information in order to demonstrate the functionality of

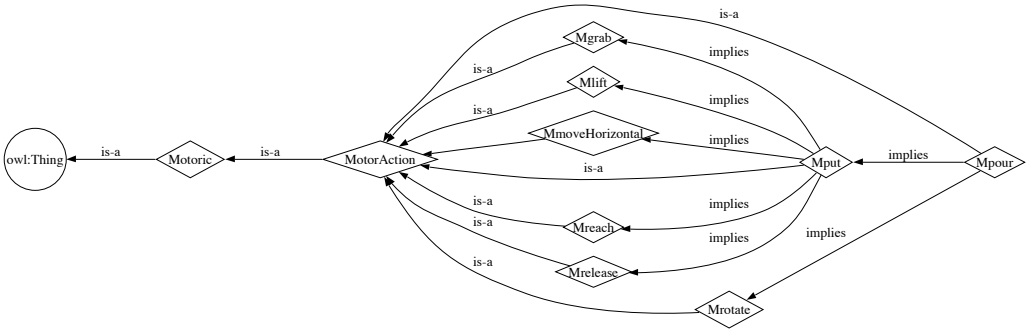


Figure 4.4: Action Ontology

the module.

4.3 The associations

The idea of associations is based on what has been found out on affordances (cf. Ellis and Tucker (2000); Glenberg (1997)) and on multi-modal integration for the purpose of mutual disambiguation and prediction, e.g. Knoeferle and Crocker (2006). First of all, we principally use associations between the linguistic and the visual counterparts of a category, i.e. the category mediates the content between the modal complements. This enables e.g. mental imagery upon hearing a word or to just facilitate the visual search for a predicted object or event upon hearing (part of) a sentence. Second, we introduce associations between objects and action schemata that are prototypical in the sense that a person would consider the action to be a functionality of the object. Note, that the object can be a complex, like a mug, in which case we deal with affordances in Gibson's sense, possibly also in Norman's sense if the object is designed such that it enables some relationship between object and interactor. Or it can be a simple object (part), like a handle, in that case we would call the connection a micro-affordance according to Ellis and Tucker.

The associations are a collection of complex data structures organised in an undirected weighted graph. The vertices represent the categories as they are employed in the ontologies and the (weighted) edges are the associations between these categories. The labels of the edges contain all the information on the nature of the connection between both categories. The details thereof are given below. The whole construct is represented by an adjacency matrix that is a common and reasonably efficient format for representing graphs [6]. The only difference to commonly known adjacency matrices is that the connection between two vertices, i.e. categories, may be manifold since two categories can

[6] From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/AdjacencyMatrix.html

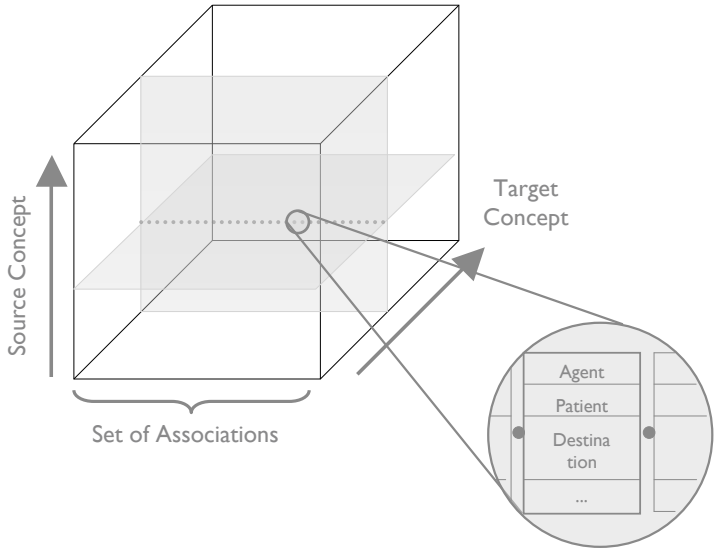


Figure 4.5: The association data structure

be connected via different types of associations of varying strength. Therefore, we do not deal with zeros and ones to indicate whether two vertices/categories are adjacent but use complex objects instead. As shown in Figure 4.5, the adjacency matrix is therefore a 3-dimensional construct rather than 2-dimensional with only one edge being the intersection. In our case, one dimension is constructed of the source categories and the second of the target categories. For each concept, however, we can have several types of associations putting the source and potential target concept into different relations. Therefore, the third dimension is the set of edges implementing associations between two categories that differ in their thematic role values and/or weights. In the paragraph below we explain what exactly that means and why they have to differ.

4.3.1 Thematic Roles

The concept of thematic roles is a very fuzzy one. According to Dowty (1991) it has been inroduced into modern linguistics in the 1960s by Jeffrey Gruber and further elaborated by Ray Jackendoff. It was considered to reflect the conceptual structure of an event and to be found by simply looking at lexical and syntactic patterns and their relation to meaning. Examples such as "to butter the bread" indicate one of the problems of thematic roles : they may exist even though no noun phrase can be identified to realise it, e.g. the patient of "to butter" is fully expressed by the verb and takes no additional syntactic argument. Verbs vary considerably with respect to how they assign the semantic roles to their arguments and the notion of these roles do so likewise. Hence, it is difficult to

Source Concept:	v(isual)mug
Target Concept:	m(otoric)pour
Agent:	*<agent>*
Patient:	v(isual)fluid
Location:	<location>
Destination:	v(isual)mug
Source:	<source>
Weight:	0.38

Source Concept:	v(isual)mug
Target Concept:	m(otoric)pour
Agent:	*<agent>*
Patient:	v(isual)fluid
Location:	v(isual)mug
Destination:	<destination>
Source:	<source>
Weight:	0.3

Figure 4.6: Sample associations

identify criteria that uniquely determine the concrete thematic role of an argument. In any case, to do so, one needs to agree on what linguistic evidence defines the types of thematic roles first. A lot of effort has been put into that task but according to Dowty this is not necessary and not possible. In (Dowty, 1991) he introduces the notion of *thematic proto-roles*, in the style of Rosch's category prototypes. (McRae, Ferretti & Amyote, 1997), on the other hand, suggested that the wide-spread view of thematic roles as slot/filler mechanism (where the slot is a kind of semantic argument), be kept in principle but is extended by the world knowledge that people possess about who typically does what to whom in specific situations. Thus, thematic roles should be considered verb-specific and feature-based categories. They supported their view by a study where subjects had to first generate features for a role (e.g. for agent: someone can be frightening), then selected typical fillers for that role (e.g. a monster can frighten) and finally they let other subjects rate the importance of the features for the given role fillers. From that a similarity measure could be computed to predict typicality for the combination which in turn has been used to bias (locally ambiguous) sentences to see whether this had an impact on disambiguation. The use of a typicality rating suggests that McRae et al. at least do not deny the existence of prototypes.

We stick to the most widely known and commonly used thematic roles like agent, patient, goal, location, and source. We use them to represent the structure of an event and encode details about the who, where and what. How the candidate role fillers are recognised and classified into the various role types is another issue that needs to be dealt with elsewhere. We would like to point out, however, that the amount and type of thematic roles used for describing an association is principally flexible and can be extended or changed according to the model one wishes to follow.

Knowing what thematic roles reflect, it is now easy to see why there are several options for one association between an action and an object. The categories alone can be part

of many different events and we need to distinguish them by their conceptual structure, i.e. by thematic roles. As shown in Figure 4.6 simply exchanging destination and location yields a quite different action to be executed or predicted, it would be directed in the contrary way. In our sample association, the source concept Mug is associated with the target Pour in a way such that it is the 'destination' of the action and something fluid is supposed to be poured **into** the cup. If the Mug was the 'location' than it would yield a pouring-action **from** the mug into some other container. Because we store these associations in a list, they can be accessed by their order of weights such that the most salient association is retrieved first.

Weights

The weighting of the associations is rather difficult to determine as there are no studies or other evidence that may suggest a relative estimate for the strength of the various associations. Of course, it is impossible to give a measure for that since it is very difficult to make associations made by humans such as affordances explicit at all. Furthermore, they evolve from a life-long experience of each individual person and account for how often she has encountered this object-action-combinations or how useful she judges it. So naturally the weights are learned individually and again depend on the body of the perceiver and the experience. For that reason, hand-crafted weights can only be crude indications that reflect our intuition on that. For instance, we set the weights between visual and linguistic correspondents of a category to very large values, indeed to 0.5, because upon experiencing one usually both occur to one if both exist. The weights between categories of the motor action ontology and object-categories in vision are initialised with a value representing their default character to us, i.e. Handle and the Grab-action are prototypically associated by strong weights whereas Bottle and Put is just one out of many possible associations of Bottle and of Put and is therefore represented by a lower weight. When retrieving associations we want to get the most likely ones, that is, the ones with the highest weight values. At the same time, we want to retroactively enhance associations the system actually encounters such that the weights stay "up to date". This is achieved by a simple mechanism similar to hebbian learning in neural nets: the more often a connection is used the higher the weights become (Russell & Norvig, 2002). There the new weight is determined by multiplying the new activation values (input and output vector *a* and *b*) and adding the old weight(*w*), i.e. $w[i,j] = a[i] * b[j] + w[i,j]$. In our case we do not deal with input and output vectors and only increase the weight such that it steadily increases but asymptotically approaches 0.5 as maximum chance for being chosen. This is ensured by the following function *p*:

$$p(w) = \frac{w}{w+1} \text{ with } \lim_{w\to\infty} p(w) = 0.5$$

where *w* is double the weight. This way, the weight may steadily increase according to usage and yet the likeliness that results from it and that determines which association is

selected first is still bounded by 0.5.

4.4 Incremental utterance processing

The OpenCCG framework[7] provides means for parsing and realizing with CCG grammars. The parser is a chart-parser, maintaining a *chart* in which it stores all partial analyses, with each analysis indexed to the string positions in the utterance it covers. Parsing proceeds by trying to combine partial analyses (at the lowest level, individual words) on the basis of their categories, until one or more analyses are found that span the entire utterance. An agenda acts as a scheduler, by storing what partial analyses are still *active* in the sense that we can try and combine them with other partial results. After each parsing step, we update the agenda; cf. (Sikkel, 1999). (Steedman, 2000) explains how we can obtain an incremental "left-to-right" parsing strategy for CCG by adjusting the order in which the agenda schedules partial analyses to be considered. Essentially, this comes down to first trying to extend partial analyses, starting from the 0-position in the utterance, in a left-to-right fashion. This is the strategy we adopt here, too. Consider Example 3 below.

(3) "The man pushes the box"

a. **Initialize:**
{0} the :- np_x/n_x
{1} man : n_{p1} : $@_{\{p1:person\}}$**man**
{2} pushes : $s_{t1}/np_{x2}\backslash np_{x1}$:
$@t1 : action(push \wedge$
$\langle$Mood$\rangle$**ind** $\wedge$
$\langle$Actor$\rangle p1 : person \wedge$
$\langle$Patient$\rangle x2 : thing)$
{3} the :- np_x/n_x
{4} box : n_{b1} : $@_{\{b1:thing\}}$**box**

b. {0,1} the man :- np_{p1} : $@_{\{p1:person\}}$**man**

c. {0,2} the man pushes : s_t1/np_{x2} :
$@t1 : action(push \wedge$
$\langle$Mood$\rangle$**ind** $\wedge$
$\langle$Actor$\rangle p1 : person \wedge$ **man** $\wedge$
$\langle$Patient$\rangle x2 : thing)$

d. {3,4} the box : np_{b1} : $@_{\{b1:thing\}}$**box**

e. {0,4} the man pushes the box : s_t1

[7] http://openccg.sf.net

$@t1 : action(push \wedge$
$\langle$Mood$\rangle$**ind** $\wedge$
$\langle$Actor$\rangle p1 : person \ \wedge$ **man** $\wedge$
$\langle$Patient$\rangle x2 : thing \ \wedge$ **box**)

Example 3 illustrates parsing "The man pushes the box" left-to-right. Step (a) initializes the chart with the lexical entries for the individual words. Step (b) starts from the 0-position, combining "the" and "man" together by *applying* the category of the determiner to that of the noun. In step (c) we then can then apply the category for the verb to the category for the expression "the man", to yield the expression "the man pushes". The meaning for this expression specifies the man as the actor of the pushing action. Subsequently, in step (d) we form the noun phrase "the box" and combine it then with the expression "the man pushes" to yield an analysis of the entire utterance.

We try to connect the content from the analysis at each step with content in other modalities. In the best case, we find conceptual structures in the working memory with filled argument slots. The possible parses are matched against the available information there, e.g. on thematic roles, which produces a preference of one parse over another.

4.5 The visual scene analysis

The method we use for visual scene analysis is based on (Kelleher et al., 2006). The representation format is a logical form, as in the linguistic analysis. The logical form contains information on the type of objects that are recognised and on their spatial arrangement from egocentric point of view. An example of a simple visual scene (in Blocksworld style) is provided in next chapter. The scene is analysed on the basis of a stereo-camera frame (still without 3-D interpretation though). Furthermore, we employ a head-gesture recognition software called "Watson" [8] that allows us to track the direction of gaze of the human interaction partner. This does not currently influence the robot's use of gaze to comprehend the scene and act in it accordingly. However, the next step will be to integrate situated gaze with "conversational" gaze to provide visual feedback and facilitate turn-taking (cf. Sidner et al. (2005)).

4.6 Gaze production

We now look at how grounded gaze evolves from our system. In any situation the robot has to fuse linguistic and visual information with already existing knowledge. This is best explained by a snapshot of such a situation. E.g the incremental parser produces partial parses for the command "Put the ball on the box" with either all saturated arguments yielding a complete sentence or with an unsaturated location-argument since "on

[8] http://groups.csail.mit.edu/vision/vip/watson

the box" could be interpreted as a modifier. The visual setting provides the robot with additional information on what is being talked about. So it interprets the visual scene with respect to basic object classification and spatial arrangement and detects one ball-type object being on a box-type object. The category system produces category activations for each partial parse and for each new scene analysis that are merged and stored in a working memory. In our snapshot this working memory contains categories like Put, Mug, Cup, Ball, Put, Box and thematic role information for some of the associations. The association between Put and Ball still contains open slots for location and destination. However, when the visual information is merged with the contents of the working memory the empty location-slot in that association is filled up. Having this version of an association in the working memory primes the linguistic analyses with respect to thematic role assignment. This means, the partial parses from the incremental parser are matched against already available information on thematic roles from the working memory. The result is a preference for the parse with one yet unsaturated argument, namely the location argument. Unfortunately, for the specified action, the system does not find any information with respect to typical locations nor can it uniquely identify a potential location from the scene since the whole table top is a potential location area. However, at this stage the robot should already start looking at the table for a referent with the expected thematic role before it hears the complete sentence "Put the ball on the box on the mug".

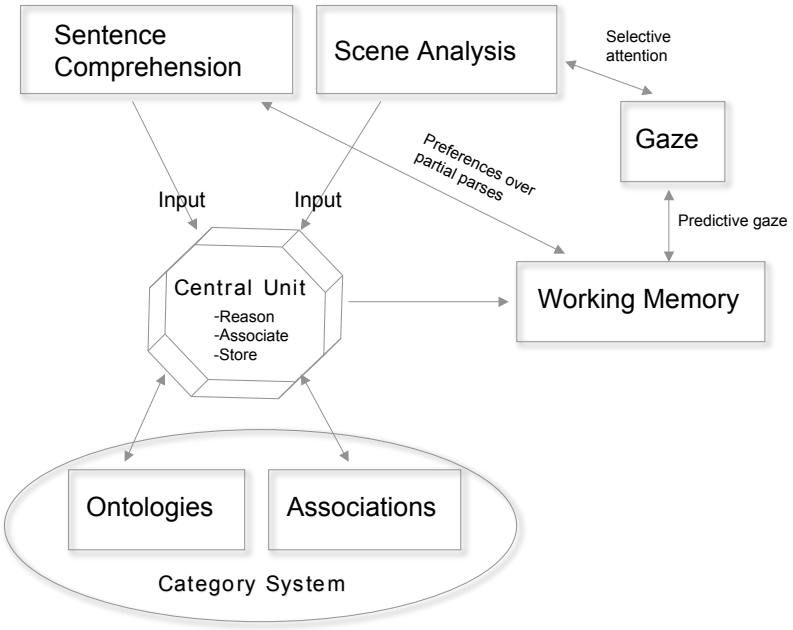


Figure 4.7: Sketch of Module Interaction

4.7 Discussion

Besides the details given above, there are still many open questions. We have repeatedly emphasised the fact that we need several ontologies each being grounded in its modality because that is where the meanings of categories evolve from. It is exactly this kind of grounding that is still an unsolved problem. Possible approaches include work by Bailey, Feldman, Narayanan and Lakoff (1997); Feldman and Narayanan (2004); Roy (2005). These are options for grounding action-categories in the motor competence. In our case, we also need to ground the feature ontology, let alone the visual ontology. The latter is a very delicate matter. To explain why, we briefly return to the example of mental imagery and the blue elephant with yellow dots on it. It is conceivable that people store images of places, persons and specific objects in a kind of (visual) episodic memory. If one is asked to visualise one's grand-mother one can immediately retrieve mental images of her. The elephant in question, however, is not something one would have stored in memory because one has never seen it. And yet, it is a simple task to visualise it. In our opinion, this is so, because (instead of or additionally to whole (3D-)pictures) humans store single features and properties, such as shapes, colour, frequency, position etc. and they can creatively recombine them in our mind to produce new images or to simulate anticipated images and even whole events. The decomposition of percepts into features has been promoted among others by David Marr and his model of visual processing. This, however, poses the next difficulty. It leads to the notorious problem of binding, that is, re-combining perceived features to recognize more complex entities like whole objects (cf. Engel and Singer (2001); Singer (2003)). At this point we return to our feature ontology and want to once again emphasise how important the grounding of features is. Our feature ontology includes visual properties like shape, colour and size. But it also contains properties relating to configuration space like Reachability. It is necessary that we build a direct connection from the category into the configuration space to evaluate the feature Reachability for a particular object. Other features relate to haptic sensors and should be grounded in the specific sensory input pattern that leads to the category. Haptic features indicating the nature of a material are, of course, not determined by merely sensing the object. World knowledge about combining the sight of something and what it feels like is needed to construct a deeper understanding of the object and to make a judgment on e.g. Breakability or Rigidity. Figures 4.8 and 4.9 illustrate that there are several overlaps in classes that allow us inferencing with respect to both superordinate categories of a class. That means, some categories like mugs and boxes are also containers and therefore can contain other things. This kind of inheritance reasoning will become more and more important as inferencing over related categories (except for the IS-A-relation) is very restricted with OWL and Racer.

4.8 Conclusions

In this chapter we have presented a possible implementational design for a multimodal category system that, in combination with incremental utterance processing and visual scene analysis produces robot gaze that is grounded in a categorical understanding of the current situation. The main modules of the system and their interactions are depicted in Figure 4.7. We have pointed out some short-comings of the current implementation throughout the chapter and particularly in Section 4.7. However, the designed system principally allows reactive as well as anticipating use of gaze - as will be illustrated by some examples in the following chapter.

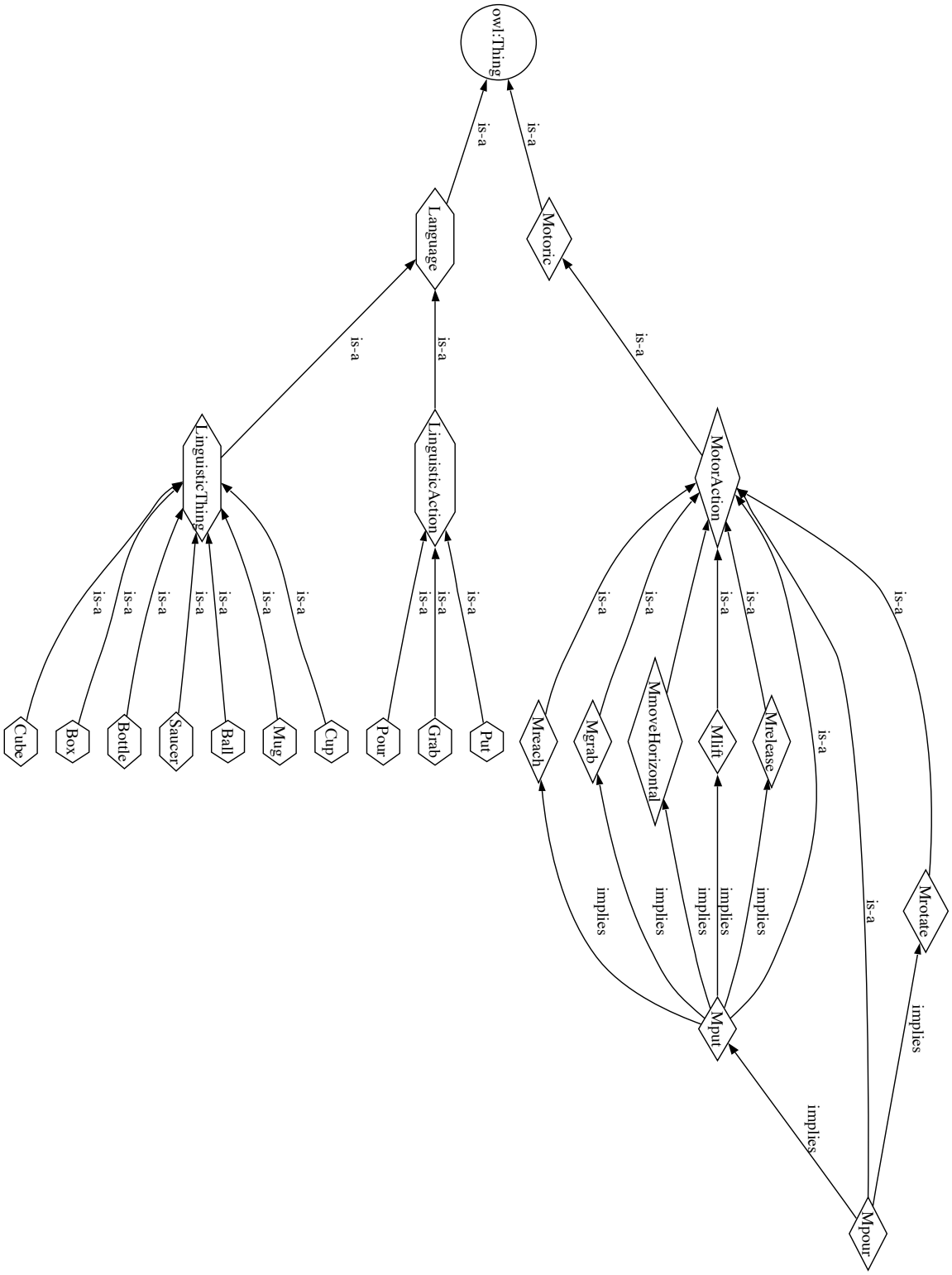


Figure 4.8: Language and Action Ontologies

Figure 4.9: Vision and Feature Ontologies

Chapter 5

Modal Integration and Evaluation

Summary
This chapter exemplifies the interactional processes of our system. We explain our predictions made in the previous chapter about robot gaze behaviour concerning eye-saccades and resolving ambiguities. Simultaneously, we provide a sketch of how we plan to fulfill these predictions. We validate the implementation in sample interactions and describe the expected effects step-by-step. Additionally to the engineering evaluation, we plan to compare the system's performance (on a functional level) to psycholinguistic experimental results which would allow us an evaluation of the system as cognitive model.

5.1 Preliminary issues and challenges

There are mainly two competing approaches to explaining how people use and make sense of the information being perceived via different modalities. The *common format* theories, on the one hand, assume that all modal stimuli are transformed into some kind of more abstract representational format such that they can then be processed. Theories supporting the concept of 'association', on the other hand, assume that stimuli are and remain *modality-specific* and that they are processed within their specific networks. The results of such processing are then associated with other modal-specific results. Although the former approach is more widely accepted, there exist experimental results supporting both, e.g. Bernstein, Auer and Moore (2004); Singer (2003). The idea is that some information is integrated early across modalities while other information remains separate, modality-specific over longer processing. We have adopted a mixture of both, the *common format / convergence-* and the *modality-specific/association* theories, for our implementation with a stronger tendency to association across all levels. However, different approaches are conceivable here. One aspect influencing this design decision is the representation of abstract categories. We have not yet dealt with categories that have no direct connection into any modality like "division" in maths or "freedom". It remains unclear whether this kind of abstract(ed) knowledge is achievable by grounded categories and association only. Moreover, the knowledge base should ideally be fed back into

the sensori-motoric perception modules, in our case these are the incremental parser and the visual scene analysis. This would facilitate the abstraction process from perceived instances to categories thereof.

Further limitations of the implementation are posed by the reasoning agent RACER and the OWL knowledge base. Apart from some problems with "prototypical" reasoning mentioned earlier, we are also restricted to monotonic reasoning. That means, upon experiencing a fact that conflicts with the current knowledge base the system can do nothing but reject the new fact, so there is little space for correction.

Nevertheless, already with the current implementation we can show that the principles of our model are valid. We can make predictions on gaze movement with respect to the current scene and an incrementally parsed utterance. For very limited scenarios both, predictive use of gaze driven by the linguistic analysis and the disambiguation function of gaze during sentence processing, are shown below to be possible within the current implementation.

5.2 Examples for Evaluation

In this section we present some simple examples showing how the modules need to interact. We include samples for visual input, the linguistic analysis and the interconnected category system. When a modality perceives a certain individual, it first needs to be able to classify this as instance of a particular concept. E.g. vision may report a red-mug-individual which is an instance of the concept mug. The activated concept mug is pushed forward by a `retrieveConcept(mug)` query which then triggers an activation process. Each concept retrieved from the ontologies by Racer is also checked for associations. An association object for mug may look as depicted below. Each as-

Target	Weight	Agent	Patient	Location	Destination
Pour	0.37499	null	Fluid	null	Mug

sociation specifies the potential event by stating who, what, where and how an action is (potentially) being executed. If a slot is *null* the role is underspecified and to be determined by the concrete situation. The weight slot specifies popularity and therefore preference of one association over another, and is determined by frequency of perception and use. The figure above is a sample association capturing *"typically something fluid can be poured into a mug"*. The results of the whole activation process are merged into a collection, the working memory, storing the most prominent concepts. For a mug, the following is obtained from our interconnected category system :

1. Parent Concepts: VisualThing, Containershape
2. Association for initial concept: Pour, 0.3749999, null, Fluid, null, Mug
3. Association for initial concept: Pour, 0.2857142, null, Fluid, Mug, null
4. Concept in part-whole relationship: Handle

5. Association for part concept: Grab, 0.16666669, null, Handle, null, null
6. Concept in containment relationship: Fluid
7. Association for contained concept: null
8. Concept that is implied (e.g. composes) : null
9. Association for implied concept: null

This, in turn, can then be used as a filter for further perceptual processing. On one hand, the linguistic analysis can be disambiguated by using the information to develop a preference over parses. Thus, the retrieved concepts help to anticipate the semantics of an utterance. On the other hand, the vision module can use this information to filter the recognition process and search the scene in a more goal-oriented way for what the robot expects to be relevant. Consider, for instance, a setting comparable to the experimental setup in (Tanenhaus et al., 1995). We use the visual scene analysis in (4) as produced in (Kelleher et al., 2006), our category system and incrementally parsed linguistic input of "Put the ball on the box on the mug."

(4) Scene Analysis :-
$@s1 : abs - obj(scene \wedge$
$\langle$Property$\rangle(b1 : thing \wedge ball \wedge$
$\langle$Number$\rangle singular \wedge$
$\langle$x-Coord$\rangle 100 \wedge \langle$y-Coord$\rangle 100 \wedge \langle$z-Coord$\rangle 100 \wedge$
$\langle$Property$\rangle(r1 : color \wedge red)) \wedge$
$\langle$Property$\rangle(b7 : thing \wedge box \wedge$
$\langle$Number$\rangle singular \wedge$
$\langle$x-Coord$\rangle 100 \wedge \langle$y-Coord$\rangle 100 \wedge \langle$z-Coord$\rangle 0 \wedge$
$\langle$Property$\rangle(g3 : color \wedge green)) \wedge$
$\langle$Property$\rangle(b8 : thing \wedge mug \wedge$
$\langle$Number$\rangle singular \wedge$
$\langle$x-Coord$\rangle - 200 \wedge \langle$y-Coord$\rangle 100 \wedge \langle$z-Coord$\rangle 0 \wedge$
$\langle$Property$\rangle(o1 : color \wedge orange)))$

The above represents a scene with a ball lying on top of a box in front of the robot, and a mug to its right.

(5) "Put" :-
$@t1 : action(put \wedge$
$\langle$Mood$\rangle$**imp** $\wedge$
$\langle$Actor$\rangle(r1 : hearer \wedge$ **robot**$) \wedge$
$\langle$Dir:WhereTo$\rangle x1 : location \wedge$
$\langle$Patient$\rangle x2 : thing)$

This analysis triggers the following related and associated concepts for put:

1. Parent Concepts: MotorAction
2. Association for initial concept: Mug, 0.3749999, null, Mug, null, null

3. Association for initial concept: Cup, 0.2857142, null, Cup, null, null
4. Association for initial concept: Ball, 0.3749999, null, Ball, null, null
5. Concept in part-whole relationship: null
6. Association for part concept: null
7. Concept in containment relationship: null
8. Association for contained concept: null
9. Concept that is implied (e.g. composes) : Grab
10. Association for implied concept: Handle, 0.1666666, null, Handle, null, null

The robot would then use gaze to search the scene for any of the predicted "puttable" visual objects and will try to further analyse matched instances, i.e. the found mug and the spotted ball. Furthermore the linguistic analysis yields a thing as patient and a location as destination. Both can prime expected types of categories in principle.

(6) "Put the ball" :-
$@t1 : action(put \wedge$
$\langle$Mood$\rangle$**imp** $\wedge$
$\langle$Actor$\rangle(r1 : hearer \wedge$ **robot**$) \wedge$
$\langle$Dir:WhereTo$\rangle x1 : location \wedge$
$\langle$Patient$\rangle(x2 : thing \wedge$ **ball** $\wedge$
$\langle$Number$\rangle singular \wedge$
$\langle$Delimitation$\rangle unique \wedge$
$\langle$Quantification$\rangle specificsingular))$

The concept ball is activated and retrieved concepts and associations are:

1. Parent Concepts: Visual Thing
2. Association for initial concept: Put, 0.1666666, null, Ball, null, null
3. Concept in part-whole relationship: null
4. Association for part concept: null
5. Concept in containment relationship: null
6. Association for contained concept: null
7. Concept that is implied (e.g. composes) : null
8. Association for implied concept: null

The result would be merged with the content of the working memory and the robot's gaze could be directed towards the ball, the most prominent category. Thus, the robot would already know that the ball is lying on top of the box. It should fill the location-slot of the association structure between put and ball above with the visually recognised box and store that in working memory. However, the linguistic analysis suggests two different options.

(7) "Put (the ball on the box)" :-
$@t1 : action(put \wedge$
$\langle$Mood$\rangle$**imp** $\wedge$
$\langle$Actor$\rangle(r1 : hearer \wedge$ **robot**$) \wedge$
$\langle$Dir:WhereTo$\rangle x1 : location \wedge$
$\langle$Patient$\rangle(x2 : thing \wedge$ **ball** $\wedge$
$\langle$Number$\rangle singular \wedge$
$\langle$Delimitation$\rangle unique \wedge$
$\langle$Quantification$\rangle specificsingular \wedge$
$\langle$Location$\rangle(o1 : region \wedge$ **on** $\wedge$
$\langle$Positioning$\rangle static \wedge$
$\langle$Proximity$\rangle proximal \wedge$
$\langle$Dir:Anchor$\rangle(b2 : thing \wedge$ **box** $\wedge$
$\langle$Number$\rangle singular \wedge$
$\langle$Delimitation$\rangle unique \wedge$
$\langle$Quantification$\rangle specificsingular))))$

(8) "Put (the ball) (on the box)" :-
$@t1 : action(put \wedge$
$\langle$Mood$\rangle$**imp** $\wedge$
$\langle$Actor$\rangle(r1 : hearer \wedge$ **robot**$) \wedge$
$\langle$Dir:WhereTo$\rangle(x1 : location \wedge$ **on** $\wedge$
$\langle$Positioning$\rangle static \wedge$
$\langle$Proximity$\rangle proximal \wedge$
$\langle$Dir:Anchor$\rangle(b2 : thing \wedge$ **box** $\wedge$
$\langle$Number$\rangle singular \wedge$
$\langle$Delimitation$\rangle unique \wedge$
$\langle$Quantification$\rangle specificsingular))$
$\langle$Patient$\rangle(x2 : thing \wedge$ **ball** $\wedge$
$\langle$Number$\rangle singular \wedge$
$\langle$Delimitation$\rangle unique \wedge$
$\langle$Quantification$\rangle specificsingular))$

The first version interprets "the box" as a modifier of the ball, i.e. that it specifies the location of the ball, while it still has an open argument for the destination. The second version interprets "the box" as the destination of the put-action. Filtering these results against the already accessible information in the working memory, the robot should develop a clear preference for the first parse and anticipates the sentence to yet provide the destination of the demanded action. Hence, the robot could start looking around for an object or a region that is a typical or at least possible filler for the destination-role. Since it does not know of any particular filler object of that kind it cannot yet predict the subsequent argument of the sentence. However, the role filler gets further specified upon hearing the determiner and possibly adjectives etc. (e.g. "the red...") of the noun phrase which facilitates the visual search by reducing the search space. It is this kind of anticipating use of

gaze that was reported in studies by e.g. Knoeferle and Crocker (2006). Upon hearing the next noun phrase the complete sentence is accordingly parsed as:

(9) "Put the ball on the box on the mug" :-
$@t1 : action(put \wedge$
$\langle$Mood$\rangle$**imp** $\wedge$
$\langle$Actor$\rangle(r1 : hearer \ \wedge$ **robot**$) \wedge$
$\langle$Dir:WhereTo$\rangle(x1 : location \ \wedge$ **on** $\wedge$
$\langle$Positioning$\rangle static \wedge$
$\langle$Proximity$\rangle proximal \wedge$
$\langle$Dir:Anchor$\rangle(b3 : thing \ \wedge$ **mug** $\wedge$
$\langle$Number$\rangle singular \wedge$
$\langle$Delimitation$\rangle unique \wedge$
$\langle$Quantification$\rangle specificsingular))$
$\langle$Patient$\rangle(x2 : thing \ \wedge$ **ball** $\wedge$
$\langle$Number$\rangle singular \wedge$
$\langle$Delimitation$\rangle unique \wedge$
$\langle$Quantification$\rangle specificsingular \wedge$
$\langle$Location$\rangle(o1 : region \ \wedge$ **on** $\wedge$
$\langle$Positioning$\rangle static \wedge$
$\langle$Proximity$\rangle proximal \wedge$
$\langle$Dir:Anchor$\rangle(b2 : thing \ \wedge$ **box** $\wedge$
$\langle$Number$\rangle singular \wedge$
$\langle$Delimitation$\rangle unique \wedge$
$\langle$Quantification$\rangle specificsingular))))$

The robot should look from the ball location to its proposed destination to confirm the linguistic analysis and its internal understanding of the required task. This in turn provides the non-verbal feedback to the human interaction partner that the robot has understood.

5.3 Conclusions

We have shown how the partial implementation of the model can be used to produce gaze that is influenced by integrating crossmodal information. And vice versa, gaze is shown to take influence on the information fusion across modalities. On one hand, the general situated awareness allows the robot to use visual information for disambiguation during sentence processing. On the other hand, the predictive use of gaze has been explained to work in principle. That means our system should be able to reproduce some of the effects observed with humans in psycholinguistic studies by modelling crossmodal mediation of contents. From already linguistically processed verb-argument structure, the robot can anticipate the last thematic role-filler. Whether it detects a match between objects of the scene and the required role filler then also depends on the internal (coverage of the) representation of thematic roles.

Chapter 6

Conclusions

In human communication, gaze plays a fundamental role in providing non-verbal feedback and in guiding gradual refinement of situational awareness on the basis of what is being talked about. In this thesis we propose a model for robot gaze production during utterance comprehension, which is grounded in situational awareness to provide the above functions as observed for human gaze production. We have developed a category model that is grounded in perception and which connects the communicative modalities with each other. This provides a deeper understanding of the situation and it allows the fusion of modal information and world knowledge in order to predict information across modalities. The anticipating nature of the model enables the robot to focus early on what it believes to be relevant. This way, it simultaneously provides feedback to the communication partner about what it is paying attention to. Besides the engineering effort, with this model we hope to have also made a contribution to further cognitive investigations by providing a possible model for gaze production and testing it in a real-world environment.

6.1 Future Work

Of course there are still many open questions and many issues that need to be tackled, some being more difficult than others. Several implementational as well as conceptual details will have to be addressed in advance to using this system in further applications. First, the lack of true prototypicality reasoning is one short-coming of the tools OWL and RACER. It is essential, however, that the system can inference over categories and not just instances. The method we currently employ can overcome this deficit only partially and will have to be revised for applications in larger and more dynamic domains/scenarios. Furthermore, the ontologies representing world knowledge and the associations collection as the key component for modelling affordances reflect embodied experience of the robot. As such both need to be extendable dynamically. Therefore much future work will have to be spend on combining learning mechanisms with these structures and on improving the weighting mechanism. Related to that is the question of storage in a working memory. The best capacity measure, for instance, will have to be determined experimentally. A third major issue is the representation of categories directly as sensori-motoric patterns. That means, we will have to revise the connection of the symbols into their modalities

carefully and find out what exactly we need each category to reflect. Particularly action categories have been neglected in this thesis with respect to their grounding. Finally, we intend to integrate gaze production that is merely grounded in situated awareness with the additional functionality of gaze during dialogue such as its role in turn-taking.

6.2 Applications

The first application of the system should be to use it towards integrating the two Cosy scenarios [1] . Due to the compositional nature of ontologies it is easily conceivable that we replace - or even better: extend - our ontological system with a similarly organised knowledge base that comprises large scale entities such as rooms and larger objects contained in rooms and the functions thereof. This would enable a kind of zoom between small manipulable objects as being dealt with within the Playmate Scenario of the CoSy project and the environment in which a dynamic embodied system would have to navigate as is the case in the Explorer Scenario of CoSy. This means, that depending on the current task, the system would use different scales of entities (rooms and doors etc. versus bottles and mugs) and the same mechanisms to merge modal information and process functionalities in order to better understand its environment and predict or anticipate further events.

Moreover, any application that employs anthropomorphic characters could make use of a model for gaze production that facilitates the interaction with a human user. Designer of computer games, for instance, strive for ever so realistic characters that appear more and more human. Gaze certainly has the potential to increase naturalness of an agent's behaviour to a large extent. The same holds for other virtual agents, especially when real communication with the human is intended. These agents may be guides in a public institution or a tutor in some educational software. Of course, the ultimate goal is to provide a robot with situated understanding and gaze reflecting this, such that interaction during learning, collaboration and etc. becomes easier and more natural. This is in accordance with the CoSy project's ultimate goal to investigate methods for creating cognitive systems for cognitive assistants.

[1] http://www.cognitivesystems.org

List of Figures

Bibliography

Allopenna, P. D., Magnuson, J. S. & Tanenhaus, M. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419-439.

Altmann, G. & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247-264.

Altmann, G. & Kamide, Y. (2004). Now you see it, now you don't: Mediating the mapping between language and the visual world. In J. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (p. 347-386). New York NY: Psychology Press.

Anderson, J. & Lebiere, C. (1998). *The atomic components of thought*. Lawrence Erlbaum.

Bailey, D., Feldman, J., Narayanan, S. & Lakoff, G. (1997). Modeling Embodied Lexical Development. In *Proceedings of CogSci97*. Stanford.

Baker, C. F., Fillmore, C. J. & Lowe, J. B. (1998). The berkeley framenet project. In *In proceedings of the coling-acl*. Montreal, Canada.

Baldridge, J. & Kruijff, G. (2002). Coupling CCG and hybrid logic dependency semantics. In *Proc. acl 2002* (pp. 319–326). Philadelphia, PA.

Barsalou, L. (1999). Perceptual symbol systems. *Behavioral & Brain Sciences*, *22*, 577-660.

Barsalou, L. (2005). Abstraction as dynamic interpretation in perceptual symbol systems. In L. Gershkoff-Stowe & D. Rakison (Eds.), *Building object categories* (p. 389-431). Majwah, NJ: Erlbaum: Carnegie Symposium Series.

Bernstein, L., Auer, E. & Moore, J. (2004). Audiovisual Speech Binding: Convergence or Association. In G. Calvert, C. Spence & B. Stein (Eds.), *The handbook of multisensory processes* (p. 203-223). Cambridge, Massachusetts: MIT Press.

Borghi, A. M., Parisi, D. & Ferdinando, A. di. (2005). Action and hiercharchical levels of categories: A connectionist perspective. *Cognitive Systems Research*, *6*, 99-110.

Botvinick, M., Braver, T., Barch, D., Carter, C. & Cohen, J. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624-652.

Breazeal, C., Hoffman, G. & Lockerd, A. (2004). Teaching and working with robots as a collaboration. In *Proceedings of third international joint conference on autonomous agents and multi agent systems (aamas'04)* (pp. 1028–1035). New York, NY.

Brown, R. (1958). How Shall a Thing Be Called? *Psychological Review*, *65*, 14-21.

Calvert, G. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cerebral Cortex*, *11*, 1110-1123.

Carlson, L. & Kenny, R. (2005). Constraints on spatial language comprehension. In D. Pecher & R. Zwaan (Eds.), *Grounding cognition. the role of perception and action in memory, language and thinking* (p. 35-64). Cambridge University Press.

Chambers, C., Tanenhaus, M. & Magnuson, J. (2004). Actions and affordances in syntactic ambiguity resolution. *Jnl. Experimental Psychology*, *30*(3), 687-696.

Damasio, H. et al. (2004). Neural systems behind word and concept retrieval. *Cognition*, *92*, 179-229.

Desimone, R. & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Cognition*, *92*, 179-229.

De Vega, M., Robertson, D., Glenberg, A., Kaschak, M. & Rinck, M. (2004). On doing two things at once: Temporal constraints on actions in language comprehension. *Memory and Cognition*, *32*(7), 1033-1043.

Dowty, D. (1991). Thematic Proto-roles and Argument Selection. *Language*, *67*, 547-619.

Ellis, R. & Tucker, M. (2000). Micro-affordance: The potentiation of components of action by seen objects. *British Journal of Psychology*, *91*, 451-471.

Endsley, M. (2000). Theoretical underpinnings of situation awareness: A critical review. In M. R. Endsley & D. J. Garland (Eds.), *Situation awareness analysis and measurement.* Lawrence Erlbaum.

Engel, A. & Singer, W. (2001). Temporal binding and the neural correlates of sensory awareness. *Trends in Cognitive Science*, *5*(1), 16-25.

Feldman, J. & Narayanan, S. (2004). Embodied Meaning in a Neural Theory of Language. *Brain and Language*, *89*, 385-392.

Gibson, J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Glenberg. (1997). What memory is for. *Behavioral & Brain Sciences*, *20*, 1-55.

Glenberg & Kaschak. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, *9*(3), 558-565.

Glenberg et al. (2005). The Case for Emotion: Grounding Language in Bodily States. In R. Zwaan & D. Pecher (Eds.), *The grounding of cognition: The role of perception and action in memory, language, and thinking.* Cambridge: Cambridge University Press.

Griffin, Z. (2004). Why Look? Reasons for Eye Movements Related to Language Production. In J. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (p. 213-248). New York NY: Psychology Press.

Gurevych, I., Porzel, R., Slinko, E., Pfleger, N., Alexandersson, J. & Merten, S. (2003). Less is More: Using a Single Knowledge Representation in Dialogue Systems. In *Proceedings of the hlt-naacl ws on text meaning.* Edmonton, Canada.

Haarslev, V. & Moeller, R. (2001). *RACER User's Guide and Reference Manual* (Tech. Rep.). University of Hamburg, Computer Science Department.

Hadelich, K. & Crocker, M. (2006). Gaze alignment of interlocutors in conversational dialogues. In *Proc. 19th cuny conference on human sentence processing.* New York, USA.

Helbig, H., Graf, M. & Kiefer, M. (2006). The role of action representations in object recognition. *Experimental Brain Research*.

Henderson, J. (2003). Human gaze control in real-world scene perception. *Trends in Neurosciences*, *7*, 498-504.

Henderson, J. & Ferreira, F. (2004). Scene Understanding for Psycholinguists. In J. Henderson & F. Ferreira (Eds.), *The Interface of Language, Vision, and Action: Eye Movements and The Visual World* (pp. 1–58). New York NY: Psychology Press.

Hepple, M. (1991). Efficient Incremental Processing with Categorial Grammar. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL-29).* Berkeley, CA.

Hickok, G. & Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition*, *4*(4), 67-99.

Horridge, M. (2004). *A Practical Guide To Building OWL Ontologies With The Protege-OWL Plugin Edition 1.0* .

Kamide, Y., Altmann, G. & Haywood, S. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye-movements. *Journal of Memory and Language*, *49*(1), 133-156.

Kandel, E., Schwartz, T. M. J. & Jessel, T. M. (Eds.). (1991). *Principles of neural sciences* (fourth ed.). New York: Elsevier.

Kaschak, M. P. & Glenberg, A. M. (2000). Constructing meaning: The role of affordances and grammatical constructions in sentence comprehension. *Journal of Memory and Language*, *43*(3), 508-529.

Kelleher, J., Kruijff, G. & Costello, F. (2006). Proximity in context. In *Proc. ACL 2006.*

Knoeferle, P. & Crocker, M. (2006). The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking. *Cognitive Science*.

Kruijff, G., Kelleher, J. & Hawes, N. (2006). Information Fusion For Visual Reference Resolution In Dynamic Situated Dialogue. In E. André, L. Dybkjaer, W. Minker, H. Neumann & M. Weber (Eds.), *Perception and Interactive Technologies (PIT 2006).* Spring Verlag.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal About The Mind*. Chicago IL: University of Chicago Press.

Lakoff, G. & Johnson, M. (1999). *Philosophy in the Flesh*. New York NY: Basic Books.

Mandler, J. & McDonough, L. (1998). On developing a knowledge base in infancy. *Developmental Psychology*, *34*, 1274-1288.

Mayberry, M. R., Crocker, M. W. & Knoeferle, P. (2005). A Connectionist Model of

Anticipation in VisualWorlds. In *Lecture Notes in Computer Science (Proceedings of IJCNLP)* (Vol. 3651, p. 849-861).

McGuinness, D. L. & Harmelen, F. van. (2004). *OWL Web Ontology Language Overview. World Wide Web Consortium (W3C).* At http://www.w3.org/TR/owl-features/.

McRae, K., Ferretti, T. R. & Amyote, L. (1997). Thematic roles as verb-specic concepts. *Language and Cognitive Processes*, *12*(2-3), 137-176.

Mecklinger, A., Gruenewald, C., Weiskopf, N. & Doeller, C. F. (2004). Motor Affordance and its Role for Visual Working Memory. *Experimental Psychology*, *51*(4), 269-280.

Miyauchi, D., Sakurai, A., Makamura, A. & Kuno, Y. (2004). Active Eye Contact for Human-Robot Communication. In *Proceedings of CHI 2004* (p. 1099-1104). ACM Press.

Norman, D. (1999). Affordance, conventions, and design. *Interactions*, *6*(3), 38-43.

Novick, D., Hansen, B. & Ward, K. (1996). Coordinating turn-taking with gaze. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP-96)* (p. 1888-1891). Philadelphia, PA.

Phillips, J. & Ward, R. (2002). S-R compatibility effects of irrelevant visual affordance: Timecourse and specificity of response activation. *Visual Cognition*, *9*, 540-558.

Pickering, M. & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*, 169–225.

Rizzolatti, G. & Arbib, M. A. (1998). Language within our grasp. *Trends in Neurosciences*, *21*(5), 188-194.

Rosch, E. (1978). Principles of Categorization. In E. Rosch & B. Lloyd (Eds.), *Cognition and categorization.* Hillsdale, NJ: Lawrence Erlbaum.

Rosch, E., Mervis, C. B., Gray, W., Johnson, D. & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382-439.

Roy, D. (2005). Grounding words in perception and action: computational insights. *Trends in Cognitive Science*, *9*(8).

Russell, S. J. & Norvig, P. (2002). *Artificial Intelligence: A Modern Approach*. Upper Saddle River, New Jersey: Prentice-Hall.

Sidner, C. L., Kidd, C. D., Lee, C. H. & Lesh, N. (2004, January). Where to Look: A Study of Human-Robot Engagement. In *ACM International Conference on Intelligent User Interfaces (IUI)* (p. 78-84).

Sidner, C. L., Lee, C., Kidd, C., Lesh, N. & Rich, C. (2005). Explorations in engagement for humans and robots. *Artificial Intelligence*, *166*(1-2), 140-164.

Sikkel, K. (1999). *Parsing schemata*. Springer Verlag.

Singer, W. (2003). Synchronization, binding and expectancy. In M. Arbib (Ed.), *In: The handbook of brain theory and neural networks* (Second ed., p. 1136-1143). Cambridge, MA: The MIT Press.

Steedman, M. (2000). *The syntactic process*. Cambridge MA: The MIT Press.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K. & Sedivy, J. (1995). Integration of

Visual and Linguistic Information in Spoken Language Comprehension. *Science*, *268*, 1632-1634.

Tucker, M. & Ellis, R. (1998). On the Relations Between Seen Objects and Components of Potential Actions. *Journal of Experimental Psychology*, *24*(3), 830-846.

Tversky, B. & Hemenway, K. (1983). Categories of environmental scenes. *Cognitive Psychology*, *15*(1), 121-149.

Yoshikawa, Y., Shinozawa, K., Ishiguro, H., Hagita, N. & Miyamoto, T. (2006). Responsive Robot Gaze to Interaction Partner. In *Proc. Robotics: Science and Systems II.*