Grounded Situation Models for Situated Conversational Assistants

by

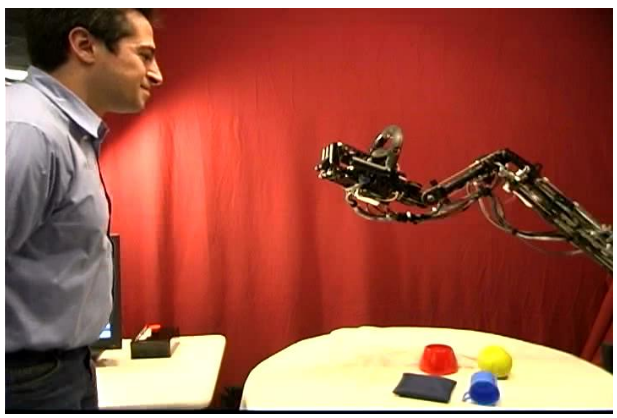
Nikolaos Mavridis

Towards Intelligent Conversational Robots

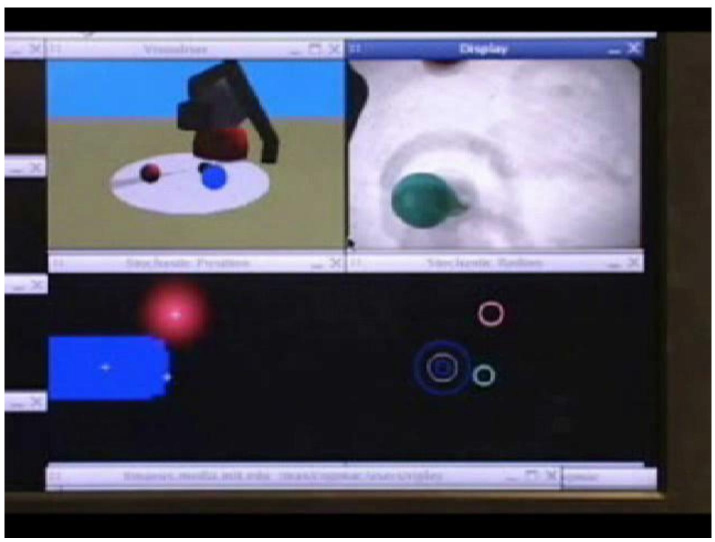
φέρε δὴ ἴδωμεν, ὦ Ἑρμόγενες, πότερον καὶ τὰ ὄντα οὕτως ἔχειν σοι φαίνεται, ἰδίᾳ αὐτῶν ἡ οὐσία εἶναι ἑκάστῳ, ὥσπερ Πρωταγόρας ἔλεγεν λέγων "πάντων χρημάτων μέτρον" εἶναι ἄνθρωπον - ὡς ἄρα οἷα μὲν ἂν ἐμοὶ φαίνηται τὰ πράγματα [εἶναι], τοιαῦτα μὲν ἔστιν ἐμοί: οἷα δ' ἂν σοί, τοιαῦτα δὲ σοί;

But would you say, Hermogenes, that the things differ as the names differ? And are they relative to individuals, as Protagoras tells us? For he says that man is the measure of all things, and that things are to me as they appear to me, and that they are to you as they appear to you. Do you agree with him?

From Plato's Cratylus



Out there, in the world:
Human, Robot, Objects, Events, Words...



Inside the Mind of the Robot:
Partial Visualisation of Grounded Situation Model

Grounded Situation Models

for

Situated Conversational Assistants

by

Nikolaos Mavridis

Towards Intelligent Conversational Robots

Foreword

A Situated Conversational Assistant (SCA) is a system with sensing, acting and speech synthesis/recognition abilities, which engages in physically situated natural language conversation with human partners and assists them in carrying out tasks. This book addresses some prerequisites towards an ideal truly cooperative SCA through the development of a computational model of embodied, situated language agents and implementation of the model in the form of an interactive, conversational robot. The proposed model produces systems that are capable of a core set of situated natural language communication skills, and provides leverage for many extensions towards the ideal SCA, such as mind reading skills.

The central idea is to endow agents with a sensor-updated "structured blackboard" representational structure called a Grounded Situation Model (GSM), which is closely related to the cognitive psychology notion of situation models. The GSM serves as a workspace with contents similar to a "theatrical stage" in the agent's "mind". The GSM may be filled either with the contents of the agent's present here-and-now physical situation, or a past situation that is being recalled, or an imaginary situation that is being described or planned. Furthermore, the GSM contains descriptions of both physical (such as objects) as well as mental aspects of situations (such as beliefs of others). Most importantly, the proposed GSM design enables bidirectional translation between linguistic descriptions and perceptual data / expectations.

To demonstrate viability, an instance of the model was implemented on a manipulator robot with touch, vision, and speech synthesis/recognition. The robot grasps the semantics of a range of words and speech acts related to cooperative manipulation of objects on a table top situated between the robot and human. The robot's language comprehension abilities are comparable to those implied by a standard and widely used test of children's language comprehension (the Token Test), and in some directions also surpass those abilities. Not only the viability but also the effectiveness of the GSM proposal is thus demonstrated, through a real-world autonomous robot that performs comparably to those capabilities of a normally-developing three-year old child which are assessed by the token test.

Acknowledgments

First of all, I would like to thank my thesis supervisor Prof. Rosalind Picard as well as Whitman Richards and Yiannis Aloimonos, for all of their help: their patient and careful reading of the thesis, their most constructive feedback, all of the interesting and fruitful discussions we have had throughout the years. All three of them have been very important scientific mentors for me, and through them I have been taught and inspired in multiple ways. Thus, they have been for me exemplars of thinkers, scientists, professionals, as well as -most importantly- well-rounded and deeply human beings. Apart from the above three, the other person that has contributed quite significantly towards this thesis is Deb Roy, the head of the Cognitive Machines group of the MIT Media Laboratory, whom I certainly admire for his novel ideas and the contributions of his group towards the important growing field of language grounding.

Furthermore, I would like to extend a warm thank you to all of the people that had contributed towards Ripley-related projects throughout the years: most importantly Kai-yuh Hsiao, and also Peter Gorniak, Niloy Mukherjee, as well as all the UROPs that have worked with me: Steve Liu, Stephen Oney, Alexander Patrikalakis, Brian Wong, Marjorie Cheng. Apart from all of the above students, I would also like to thank all of the members of the Affective Computing and the Cognitive Machines group - with whom I have enjoyed interacting with throughout the years - and most importantly Stefanie Tellex, for her most helpful comments and support.

Here it is also worth acknowledging that this work could not have easily materialized in an environment other than MIT - and thus I would like to thank all the other professors with whom I had the pleasure and honor to interact with, most notably Ted Selker, as well as Patrick Henry Winston, Pawan Sinha and Wheeler Thackston (from Harvard). And most importantly, those who through their vision and hard work have made the Media Lab a reality: such as Nicholas Negroponte, the co-founder of the Media Lab, and those that currently are the prime movers for sustaining and expanding this unique place: such as Frank Moss, the director of the Media Lab. Also, the administrative personnel of the Media Lab, with whom I have interacted and that have assisted me throughout he years - most notably Pat Solakoff and Linda Peterson.

Furthermore, I always remember the shapers of my academic personality start from well before MIT, and thus I would like warmly thank all of my previous teachers and professors, at UCLA, at the UK Open University, at the Aristotle University of Thessaloniki, at Anatolia College (with a special thank you to Vasilis Adam and Yiannis Lalatsis), as well as all those that had taught me indirectly through their writings.

Finally, all of my dearest and closest ones: my father Lysimachos for having nurtured my natural curiosity and having taught me how to find the next question to be asked, my mother Valia for her sweetness and support, my grandmother Chrysoula and her sister Vagelitsa for their loving care, and my son Lysimachos for patiently waiting for me, to show him the latest robots. And all my friends in whichever corner of the world they might be in.

Last, but most importantly, comes the person that stood next to me during difficulties and happy moments - my significant other, Layli, for all of the warmth, beauty, and support. Finally, an open thank you, to all those, who have helped or will help, towards creating a better future in which humans, robots, and other artificial entities, will live together in harmony and peace.

Contents

List of Figures

Chapter 1

Introduction

This chapter starts with an introductory section, where at first very briefly five main aspects of this thesis are presented[1]:

Vision - which main ultimate purpose this work serves

News - the performance and achieved novelties of the implemented system

Contribution - what are the main contributions of this thesis in terms of their type: theories, architectures, methods, implemented systems etc.

Salient Idea - what is the main salient idea underlying this thesis (Grounded Situation Models - GSMs), and what are some defining characteristics of GSMs

Steps - A summary of the main steps of this piece of work

Then, the chapter proceeds, with a section where the ultimate vision of this thesis: cooperative Situated Conversational Assistants (SCA), is discussed in depth. Subsequently, three highly relevant notions are introduced, which form a contextual background from the existing literature: Symbol Grounding, Situated Language, and Situation Models. Then, a discussion of the reasons why current NLP subsystems fail to simply "plug-and-play" with vision and motor control subsystems and build SCA's follows in this chapter; this discussion will also provide motivation for the subsequent introduction of Grounded Situation Models (GSMs) in chapter 4. Finally, we will conclude this chapter with a roadmap to the structure of this thesis document.

1.1 Vision, Contribution, News, Salient Idea, Steps

1.1.1 Vision of this thesis

The ultimate vision of this thesis is simple to state:

[1]This enumeration of aspects of a piece of scientific work into Vision, Steps, News, Contributions, and Salient Idea is one of the central themes taught by Patrick Henry Winston in his "Human Intelligence Enterprise" course at MIT.

- Building truly cooperative Situated Conversational Assistants (SCAs)

- and I will soon explicate the meaning of this term and discuss it in depth in section 1.2.1.

1.1.2 Contributions of this thesis

This ultimate vision has demonstrably become closer to realization through the main proposal of this thesis, the proposal of:

- Grounded Situated Models (GSMs)

- and the biggest part of this text will be devoted to the detailed description of this proposal, which contains:

- Motivation and Requirements for Grounded Situation Models (chapters 1 and 4)
- An explicit computational model: A customizable Representation, together w associated Processes (chapter 5)
- A modular Implementation Architecture (chapter 5)
- An Implemented System: Ripley the Robot (chapter 6)
- Quantitative and Qualitative Evaluations and a proposal for multi-level evaluation of SCAs (chapter 7)
- A discussion of the relation and side-benefits of GSMs to prominent AI, cogsci, and semantics theories - including a proposal for empirically sound semantics (chapter 8)
- A discussion of future extensions which could form an outline of a multi-year research program (section 9.1)

Furthermore, this text also contains, among others:

- A rough method for designing GSM-based systems (appendix D)

A more detailed listing and discussion of the contributions of this thesis can be found in section 9.2.1.

1.1.3 News - What have GSMs proven they can do?

So - even before we see what GSMs exactly are - what have we proven they can do? In short - GSMs have enabled Ripley the Robot (figure 1-1), a manipulator robot living on a tabletop populated with objects and equipped with vision and speech synthesis and recognition, to exhibit *three novel abilities* as compared to the state-of-the-art of conversational robots (see section 6.5.1), and to exhibit language comprehension abilities that are comparable to those implied by a standard and widely used test of children's language comprehension (the Token Test for Children - see section 7.4), and in some directions also surpass those abilities. In layman's terms, GSMs have enabled Ripley to go

Figure 1-1: Ripley the robot, human partner, and objects on a table

from being a somewhat "dummy" robot, that could only handle simple action requests about object that are currently visible from its viewpoint, to a more "intelligent" robot - that can imagine, remember, and entertain uncertainty, which is closer to becoming a truly cooperative Situated Conversational Assistant.

But, how dummy was Ripley before, and how much better is he now - and most importantly, what do we mean by imagining, remembering, and entertaining uncertainty? Again, nothing serves better as an introduction, than a real-world example:

Consider Ripley in his infancy - quite some years ago - as demonstrated in the video which can be found at [Hsiao2003]. Back then, all that Ripley was capable of doing was responding to commands such as: "Find the green beanbag" or "Find the top beanbag" - and tapping the requested object with his mouth. Furthermore, if an object was on the table but out of his field of view - it could not be referred to, even if he had seen it there before. Ripley had no clue in what bodily configuration it was or where the human was or what he was doing - it just knew about the objects on the table, and only those that met its eyes. Ripley did not remember anything about the past, and Ripley could only respond to simple commands - no question answering, no entertaining of uncertainty, no possibility of getting information about the situation through a human utterance such as: "Hey, Ripley, there is a blue ball on your left!".

Now consider Ripley recently - after his system was redesigned on the basis of GSMs - as demonstrated in the video which can be found at [Mavridis and Roy2006b]. For example, Ripley is now able to perform in the following way:

PART I: Answering questions

[Ripley, human user, table with a red object.Ripley looks towards the top of his table, where a red object resides]

(H1) Human: Where is the red one?

(R1) Robot: At the top.

(H2) Human: Look at the center.

(R2) Robot. OK!

[The robot looks at the center. The red object is not in view for the robot anymore. The robot sees a green object at the center of the table, which it had not seen before]

(H3) Human: Where is the red one?

(R3) Robot: Most probably at the top.

[The robot has reduced its confidence]

(H4) Human: What color are the objects?

(R4) Robot: One is red, and the other is green.

PART II: Imagining and verifying/updating through the senses

(H5) Human: Imagine a blue object on the left!

[Note that here this has the semantics of: "There is a blue object on the left!" The blue object is not visible yet - the robot cannot view the area where the imagined object should reside. However, we can still refer to it, answer questions etc.]

(R5) Robot: OK

(H6) Human: Where is the blue one?

(R6) Robot: Most probably at the left.

(H7) Human: How big is the blue one?

(R7) Robot: I have no idea.

(H8) Human: Look at the left!

[Now, the previously imagined blue object becomes visible - and the robot can now know more precisely where it is, how big it is, and what shade of blue it is. Notice, that there exists also the case that there was no object there or that there was an object with an unfitting description - this would have been noticed too]

(R8) Robot: OK

(H9) Human: How big is the blue one?

(R9) Robot: Small.

[Before the robot had no idea, now it knows it is small]

PART III: Remembering, referring to events

(H10) Human: How big was the blue one when your head started moving?

(R10) Robot: I had no idea

[The robot goes back to "when his head started moving", remembers, and answers]

(H11) Human: Where was the red one when the green one appeared?

(R11) Robot: Most probably at the top, but maybe not.

[Yet another gradation of certainty]

(H12) Human: Thank you.

(R12) Robot: You're welcome! ;-)

[good robots have good manners]

Thus, indeed Ripley can now:

a) *Imagine* parts of situations when hearing appropriate utterances, and later verify / update the imagined parts through sensory information,

b) *Remember* past situations and answer questions about them, and

c) Entertain and express multiple *gradations of uncertainty* and possible alternative situations.

Of course, the robot can arguably do so at a rudimentary level - but a significant part of this text will be devoted to descriptions of the achieved extent, the generality, and the scalability of the proposed approach. Notice that all three of the above capabilities are *novel as compared to other state-of-the-art conversational robots*, as will be discussed in section 6.5.1. Also notice that in our system, the meaning models of the adjectives ("red", "small") and verbs ("pick up") can in principle be learnt through examples - and we will again explicate further later (section 8.3.1).

Currently, the robot even surpasses the abilities demonstrated by the above real scenario - it can also detect human hands, indexical pointing, human interactions with objects, speak two languages, and more. And most importantly, the central claim that follows is that: The GSM proposal has been strategically instrumental in attaining all these capabilities, and can furthermore be applied to other SCAs (not only Ripley), and also provide a systematic method and a path for further surpassing the current capabilities (see chapter 9).

1.1.4 The Salient Idea - Grounded Situation Models

Here I will try to introduce the main salient of idea of Grounded Situation Models through a comic interlude, through which the three notions of *external reality*, *situation model*, and *embedded situation model* are introduced. This ideas will be discussed and defined in more depth later - here, a gentle and hopefully comic conceptual introduction is all that is desired. Then, I will provide a list of defining characteristics for Grounded Situation Models - i.e. what characteristics should a Grounded Situation Model minimally possess, in order to be called so.

Consider the following example of a situation, depicted by the comic in figure 1-2:

Imagine that in a couple of years, after Ripley the Robot becomes melancholic due to prolonged loneliness, we decide to build a significant other for him: Ms. Ripley. The first robot wedding takes place, and a dreamy life starts for the couple. Of course, together with rights always come responsibilities - and one of those responsibilities which no husband should ever forget is to remember his wife's birthday - and express his remembrance through a present worthy of her value for him (anyone of the readers that has forgotten it in the past should need no further explanation of its importance ;-)).

The situation that has arisen on this very important day (Ms. Ripley's birthday), is depicted in figure 1-2. Notice that are three depictions of the situation within the figure: first, the unbubbled version at the bottomleft - a depiction of external reality, or "the situation as is" in imprecise layman's terms. Fortunately for him, Ripley being equipped with a RAID array, nttp, and a prioritized multitasker, he has not forgotten this most important day, and thus apart from the couple and the table, we can also see a small (but precious!) present that he bought that lies on the table, waiting to be offered and opened. However, notice that, because yesterday the couple was playing battleship, a

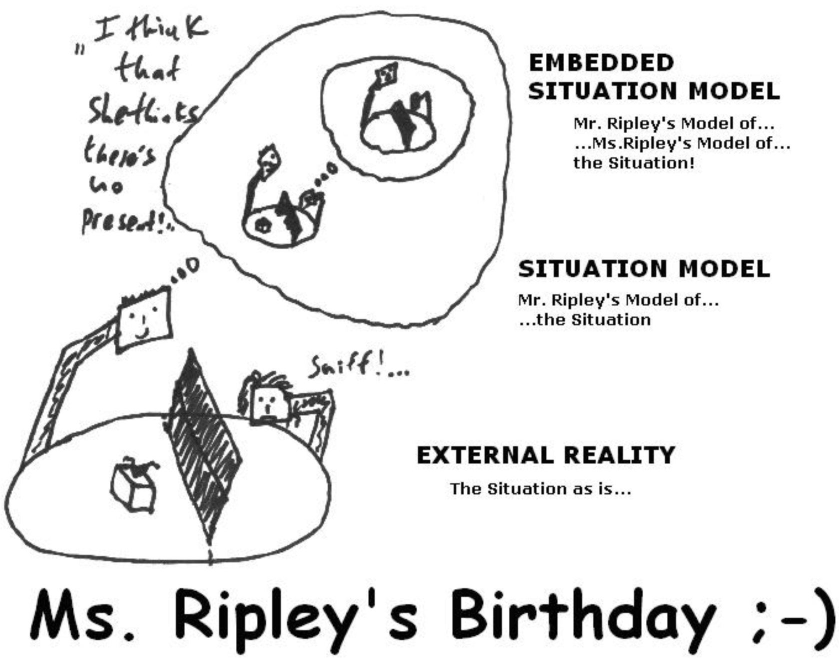


Figure 1-2: Ms. Ripley's birthday - depictions of: external reality, Mr. Ripley's situation model, and Mr. Ripley's estimate of Ms. Ripley's situation model (embedded within his situation model)

barrier exists on the table, which enables visibility of the present for Mr. Ripley, but obstructs visibility of it for Ms. Ripley, who seems to be quite unhappy. But enough about "the situation as is" - now let us ask: what is the subjective model of the situation that Mr. and Ms. Ripley possess? Traditional comic semiotics have reserved the "bubbles" for such models - and indeed, within the main bubble, we can see a partial depiction of the situation model possessed by Mr. Ripley. In very imprecise layman's terms, the information depicted within the bubble, is similar to what we call a "situation model" - which we will define more precisely soon. Notice that within the "bubble" there is also another "bubble": it is Mr. Ripley's model of Ms. Ripley's model of the situation - an estimated model, based on his knowledge of her and the situation. Such "situation models within situation models" we shall term "Embedded Situation Models" in this thesis. Notice the difference in contents between Mr. Ripley's model and his estimate of her model: knowing that she cannot view the present because of the occlusion by the barrier, he has not included the present in his estimate of her situation model's contents. Such differences in content are extremely useful: they might motivate inform statements - and thus, Ripley has good motivation to announce that "There is a present for you behind the barrier!" and turn her tears into a kiss.

So far we have introduced External Reality, Situation Models, and Embedded Situation Models. The next thing to notice is that situation models can be fed through multiple modalities: through vision, touch, proprioception, as well as through speech - for example, when listening to fairy tales. Having introduced these terms, one can proceed and ask many relevant questions: What form and contents do human situation models have? What is a suitable representation for situation models for robots, and associated processes that operate on them, accomplishing for example updates on the basis of sensory or linguistic evidence? How can we build cooperative conversational robots around situation models? Is there a systematic way to design such things? What are the implications of the GSM stance for prominent theories in semantics and AI?

In this thesis, we will try to provide answers to some of the above questions, through what we shall term the Grounded Situation Model proposal, which consists of all the parts listed previously in this foreword. But before we proceed to the main text, we will try to attempt to clarify one central question: what are GSMs? What are some desirable characteristics / requirements for terming something a GSM?

Desirable Characteristics of Grounded Situation Models

Below follows a list of desirable characteristics of Grounded Situation Models, which can be thought of as a loose definition for them. These characteristics are not independent, and in many cases possessing one will be dependent on possessing others in the list:

- GSMs should be customizable amodal representations[2], together with associated processes, which (in layman's terms) act as a "theater in the head" filled with descriptions of present, past or imagined situations.
- GSMs should be able to be fed through information arising from multiple modalities, including vision and speech, and their contents should be able to provide parameters for both speech and motor actions.
- GSMs should enable bi-directionality between sensing and language: sensory- derived situations should be translatable to linguistic descriptions, and linguistic descriptions to sensory expectation.
- GSMs should be accompanied by a set of explicated standardized processes, which perform basic operations on them: content updates on the basis of sensory evidence, matching of sensory evidence with sensory expectations etc.
- GSMs should break down situations into the ontological types that are implied by the worldview of natural languages (the "folk ontology"): situations should break down to agents and objects, the agent descriptions should have both a physical and mental part, the physical part can be broken to subparts, subparts have properties etc.
- GSMs should contain representations for "me, you, others": i.e. the self, the situational partners, and the relevant passive objects (whenever these exist in the situation under consideration)
- GSMs should also contain "embedded GSMs" within the mental part of the descriptions of agents.
- GSMs should be able to handle uncertainty and represent multiple possible alternative hypothesis about the state of a situation.
- GSMs should be able to handle past situations, by storing moments of past situations and parsing moment sequences into discrete events.

[2]For an explication of the sense with which the term "amodal" is used here, see section 4.3.

As we shall see, the GSM proposal presented in this thesis has produced models that fulfill the above desirable characteristics. Due to the interdependencies among the above characteristics, during the derivation of the GSM (later in chapter 4), we will later start by focusing on two desiderata, which correspond to the "folk ontology" and "bi-directionality" characteristics listed above, as these two desiderata, effectively "generate" the above list.

1.1.5 Steps

Below follows a discussion of the main topics and steps taken in this thesis - a concise summary:

A Situated Conversational Assistant (SCA) is a system with sensing, acting and speech synthesis/recognition abilities, which engages in situated natural language conversation with human partners and assists them in carrying out tasks. Examples of such systems include robotic helping hands, intelligent car computers, or even interactive theater lighting controllers. The ideal SCA should be truly cooperative: it should not need a constant command stream but would be able to often behave autonomously, and thus it should be able to coordinate plans with the human partner and change plans in real-time. However, in existing systems, several prerequisites of fluid cooperation with humans have not been yet achieved. Such prerequisites include capabilities for natural interaction using situated natural language, as well as a capacity for context-dependent human intention recognition and real-time activity coordination.

This thesis addresses the above prerequisites through the development of a computational model of embodied, situated language agents and implementation of the model in the form of an interactive, conversational robot. The proposed model produces systems that are capable of a core set of natural language communication skills, and provides support for extensions enabling intention recognition and activity coordination skills. The central idea is to endow agents with a sensor-updated "structured blackboard" representational structure called a Grounded Situation Model (GSM), which is closely related to the cognitive psychology notion of situation models. The GSM serves as a workspace with contents similar to a "theatrical stage" in the agent's "mind". The GSM may be filled either with the contents of the agent's present here-and-now physical situation, or a past situation that is being recalled, or an imaginary situation that is being described or planned.

Two main desiderata drive the design of GSMs. First, they should parse physical situations into ontological types and relations that reflect human language semantics. Second, GSMs should enable fluid bidirectional translation from sensation-to-language and language-to-action. The proposed three-layer hierarchical GSM design satisfies both desiderata.

In order to satisfy the first desideratum, the GSM contains representations of the self, the agent's communication partner, and salient passive objects that form the physical common ground. The representations of the self and communication partner contain descriptions of both the physical (body parts and properties) as well as the mental (beliefs and desires) aspects of them. These representations are hierarchical - for example, the physical starts at the level of composite objects, which break down to simple objects and their relations, which in turn break down to a set of lowest-level nodes representing properties (color, shape etc.).

Towards the second desideratum, the contents of each lowest-level node "ground out" to three layers of representation. The first layer ("stochastic layer") contains a multi-dimensional continuous probability distribution over property values, encoding information in a form suitable for interfacing with sensory subsystems. The second layer ("categorical layer") contains a uni-dimensional discrete distribution over verbal categories, encoding information in a form suitable for interfacing with language subsystems. The third layer ("continuous layer") contains a single multi-dimensional continuous value, which can provide single-valued action parameters for motor control subsystems. These three layers enable the desired bidirectional translation: verbally described "imagined" objects and events can be augmented / verified through sensory information and vice-versa.

Apart from specifying the representation, a set of basic processes is proposed. These perform elementary GSM operations, such as matching sensory evidence with GSM contents, updating the contents with recent sensory or linguistic information etc.

To demonstrate the viability of the proposed GSM design, an instance of the model has been implemented on an interactive manipulator robot. The robot integrates haptic and visual perception with motor control to engage in situated spoken conversation with human communication partners. The robot grasps the semantics of a range of words and speech acts related to cooperative manipulation of objects on a table top situated between the robot and human. Compared to state-of-the-art conversational robots, the robot exhibits three novel abilities: handling of inform speech acts regarding the situation, verbalization of gradations of uncertainty, and handling of references to past events. The robot's language comprehension abilities are comparable to those implied by a standard and widely used test of children's language comprehension (the Token Test for Children), and in some directions also surpass those abilities. Not only the viability but also the effectiveness of the GSM proposal is thus demonstrated, through a real-world autonomous robot that performs comparably to those capabilities of a normally-developing three-year old child which are assessed by the token test.

1.2 The Vision: Situated Conversational Assistants (SCA)

Here we start with a definition of SCA's, followed by some examples. Later, the ideal SCA is envisioned, and some prerequisites towards such an SCA are discussed. Finally, our proposal towards fulfilling some of the prerequisites is briefly touched upon.

1.2.1 What is an SCA?

A Situated Conversational Assistant (SCA) is a system with sensing, acting and speech synthesis/recognition abilities, which engages in physically situated natural language conversation with human partners and assists them in carrying out tasks.

Figure 1-3: Some helper robots - existing or under development: Roomba, Elvis, Robonaut

1.2.2 Examples of SCAs

Numerous applications await for the development of adequately capable SCAs. One might envision robotic lab assistants, that can assist with assembly tasks: for example, the robotic platform Ripley [Roy et al.2004] which is used in this thesis was designed with that ultimate purpose in mind. Household helpers that can do more than current systems such as the vacuum-cleaner "Roomba" are another domain of application. Robotic helpers for space assembly tasks, such as NASA's "Robonaut" project would benefit from advances in SCA technology. Finally, many more systems with less anthropomorphic "bodies", such as cars, theater lights [Juster and Roy2004], affective robotic computers , and even virtual agents could use similar SCA "engines".

1.2.3 Marginally qualifying SCAs

If one relaxes the requirements of the given definition of SCAs, then numerous examples of such systems are already in wide commercial use. By relaxing the requirement of speech recognition, while limiting sensing to push-buttons and acting to simple servos/valves, then various ATMs / ticket selling-machines as well as gas stations could marginally be called "SCAs". However, the point to notice is that such machines most often have very limited domains of application, a very inflexible and machine-centric cycle of interaction (usually having single-initiative dialogues), and no capabilities of learning / adaptation.

1.2.4 The ideal SCA

Moving from the existing to the ideal, one would expect an ideal SCA to perform comparably to a human assistant, for a wide variety of skills. Here, the notion of a "cooperative assistant" as contrasted to a "dummy tool" becomes central. But what are some main dimensions of difference between the two?

Here, we propose three main dimensions of difference: autonomy, user-adaptation, task-relevant as well as situated knowledge (figure 1-4). As a first illustrative example, consider a human faced with the task of writing his name on a paper, and two possible tools in order to accomplish it: a *calligraphy pen* (passive, dummy tool), and a *trained calligraphy artist* (human cooperative assistant). The passive tool, the *calligraphy pen*, has no autonomy - it needs continuous control by the human - every minute movement has to be dictated by manipulating your hand posture and pressure. Furthermore, the calligraphy pen does not adapt to the human - in contrast, the human has to learn how to use it, and adapt to its particular needs. Regarding knowledge, the calligraphy pen

Dummy Tool vs. Cooperative Assistant:

Tool:

- No autonomy – Needs continuous control
- Human adapts to tool
- Tool does not know about human, objects & task

Cooperative Assistant:

- Coordinated autonomy – Takes initiative
- Assistant adapts to human
- Assistant knows about human, objects & task

Figure 1-4: Dummy tool vs. Cooperative assistant

does not possess any dynamically updated information about the human, the objects it interacts with, as well as the task: it might only epiphenomenally "know" through the tool designer's choices. Now consider using a *trained calligraphy artist* (human cooperative assistant) for the same task: you just have to say "please write down my name on this paper" and the human will autonomously execute your command, adapting to your name and to the surface of the paper, while having task-relevant knowledge (through calligraphic training) as well as situated knowledge (regarding the state of his own body, the commanding human, the paper in front of him and the pen he has at hand etc.).

As a second example, let's examine a slightly different case: a human helping me out with cooking - an example which compared to the first offers a more balanced cooperative scenario, requiring continuous interaction and shared re-planning. Here, the assistant achieves coordinated autonomy: when my friend and I are preparing lentil soup, I do not need to command every move he makes: I might be washing the lentil while he might be cutting the onions, without me supervising / commanding any move he makes. He might even have decided to deal with the onions just by seeing me take the lentil, without any explicit "overall" command from my side. Also, he would have adapted to quite a variety of choices of order or style of task execution from my side. Furthermore notice that the human assistant is dynamically monitoring me, the objects involved in cooking, and keeping track of the current state of the task - in total contrast to a "dummy" screwdriver. Finally, he can change plans in case of an unexpected outcome or in case I change my mind.

Of course, between the *calligraphy pen* (dummy tool) and the *kitchen helper friend* (truly cooperative assistant - the ideal SCA), there exists a whole continuum of intermediate cases. The architecture and implementation presented in this thesis lies somewhere in between in the tri-dimensional space created by the difference dimensions- and as we will see (in chapter 9), provides significant leverage for extending further towards the direction of the ideal SCA.

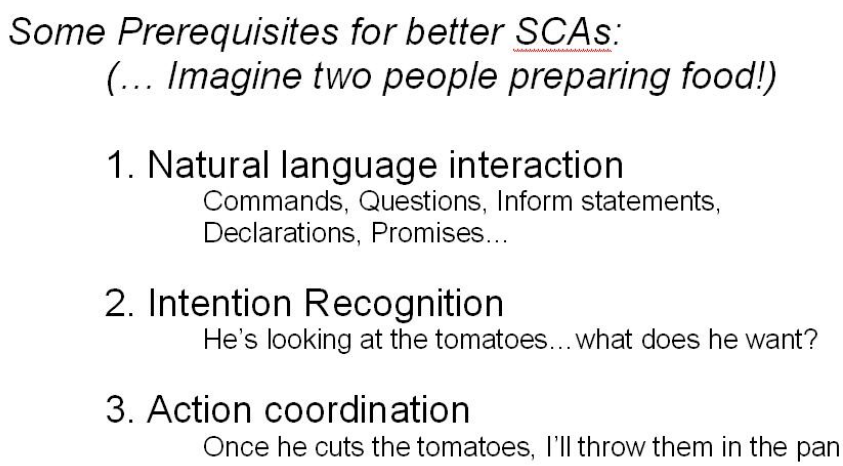


Figure 1-5: Some prerequisites for better SCAs

1.2.5 Prerequisites for better SCAs

But what are some prerequisites for creating SCAs that exhibit true cooperation? Here, we propose that such prerequisites include capabilities for natural interaction using situated natural language, as well as a capacity for context-dependent human intention recognition and real-time activity coordination.

Again, consider the cooking assistant case as an illustrative example:

First, numerous different types of natural language interactions take place (figure 1-5). Commands are given ("pass me the dish") or questions asked ("where is the lentil?"): both are traditionally viewed as REQUESTS under speech act theory [Searle1969] (The first being a request for a motor action to be taken, the second for a speech action). Apart from REQUESTS, INFORM acts also take place ("the lentil is behind you") as well as PROMISES ("I will prepare the sauce"), and even DECLARATIONS ("We call this sauce béchamel"). The ideal SCA, should be able to produce and comprehend all the above species of natural language interactions.

Furthermore, often non-linguistic devices are also used for similar purposes. For example, indexical pointing is often employed in conjunction with speech ("There!" or "This one").

Also, apart from explicit statements of declaration of intention ("Now I want the dishes to be cleaned"), intention is often silently inferred by observing actions that naturally belong to specific plans given the situational context. For example, when I see my assistant looking at the tomatoes that are within my reach, I might easily have guessed that he needs them for the salad he will prepare.

Finally, my assistant and I fluidly orchestrate our moves and take turns, with or without verbalised action coordination. (For example, I might be awaiting for the pieces of garlic that my assistant is cutting, and keep adding them to my mix whenever they arrive).

REQUEST for Motor Action:
"Pass me the dish", "Cut the potatoes", "Find a knife"

REQUEST for Speech Action:
"Have you seen the blender?", "What should I do after cutting them?"

INFORM:
"The blender is in the top drawer", "I don't know where it is"

PROMISE:
"I will prepare the sauce", "I will go and get some"

DECLARATION:
"We call this sauce béchamel", "This is Mary"

Figure 1-6: Some different types of natural language interactions, classified by speech act

1.2.6 How to fulfill the prerequisites?

Having started from the definition of an SCA, having envisioned an ideal SCA, and having proposed some prerequisites for such an SCA, what is needed now is a starting point for the design of such a system.

The approach taken here will rely on the previously presented central salient idea, namely the notion of a "Grounded Situation Models" (GSM). In short: Grounded Situation Models are "structured blackboards" representing current / past / imagined situations. The GSM serves as a workspace with contents similar to a "theatrical stage" in the agent's "mind". The GSM may be filled either with the contents of the agent's present here-and-now physical situation, or a past situation that is being recalled, or an imaginary situation that is being described or planned. In layman's terms, the GSM is very roughly similar to the "bubble in the head" that one finds in comic strips:

In subsequent chapters, we will motivate GSM's through multiple pathways as well as derive their overall architecture through two explicit desiderata (chapter 4), and propose specific representations, processes, as well as a modular architecture for their implementation (chapter 5), which will are also demonstrated in an implemented real-world robotic assistant (chapter 6).

But how does this help in designing SCAs? A rough sketch of a design method will be described later (appendix D), and an example derivation of a design will be discussed. The method starts by translating a behavioral specification of an SCA into a specific structure of a GSM, essentially specifying what aspects of the situation in which the SCA is embedded have to be noticed. The method also specifies which processes must minimally exist in order to "connect" the GSM to senses and actions. Later, the whole SCA is built around the GSM. Thus, there are two main stages in the design method: from behavioral specification to GSM, and from GSM to SCA by "building around" the centrally-located GSM. More details in appendix D.

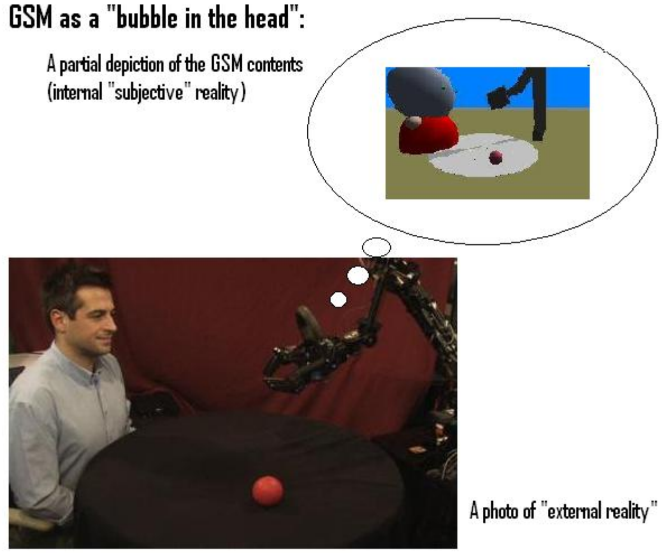


Figure 1-7: GSM as a "bubble in the head": Photo of "external reality" containing Robot, Human and Object, and depiction of corresponding GSM "bubble" with partial depiction of GSM contents

1.3 Relevant Notions: Symbol Grounding, Situated Language, Situation Models

In this section, three key background concepts are going to be introduced: *symbol grounding, situated language*, and *situation models*. All three of these concepts are crucially involved in the proposed method of designing SCAs. In chapter 2, we will revisit these concepts in view of the semiotic triangle, that will also provide a unifying picture.

1.3.1 Symbol Grounding

Any system of symbols, if isolated from the physical world, cannot escape becoming circular. For example, consider a modern dictionary of English. Each word is defined in terms of other words. The entry for "apple" will be built up from other words such as "fruit", "tree" etc. But then, what does "apple" mean to me, if I have no knowledge whatsoever of "fruit" or "tree"? This question lies exactly the heart of the symbol grounding problem [Harnad1990], and is also central to the highly-relevant Chinese-room argument [Searle1980]. Unfortunately, most traditional NLP systems, even those aiming towards semantic representations, suffer from this problem.

For example, consider one classic method for representing meaning in traditional NLP: semantic networks, such as wordnet [Fellbaum1998]. Semantic networks are essentially graphs, where the vertices roughly represent concepts (synonym sets in wordnet), and the edges represent typed relations, such as category membership (IS_A), HAS_A etc. Then, a small part of the giant graph containing all concepts, centered on the vertex "apple", might look like (figure 1-8):

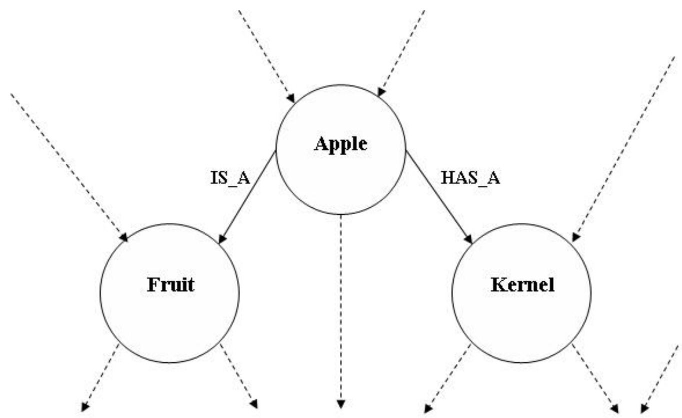


Figure 1-8: Part of a semantic network - "apple" IS_A "fruit", "apple" HAS_A "kernel"

So, what is the meaning of "apple" here? First of all, we have cut down numerous other edges that might be attached to the "apple" vertex. Also, an important question arises: should we not include edges originating from "fruit" or "kernel" in the meaning of apple? How far should we go? Where do we stop? The similarity with the dictionary definition case should be obvious. Again, in this case, one can only define "apple" in terms of other concepts: "fruit", "kernel". Even worse, the total meaning of apple contains all other concepts in the graph[3]. Furthermore, any such attempt towards expressing meaning is bound to be circular and enclosed within the system: one only gets the relation of the meaning of "apple" with the meaning of other concepts within the system, and one can never escape the system in order to define the meaning of "apple" in terms of other, non-conceptual or non-linguistic stuff.

The SCAs described in the proposal presented in this thesis, overcome this problem, by connecting the spoken or heard words to sensory data and actions, effectively "grounding" the meaning of their symbols to the world. In this respect, this thesis forms part of a wider body of research, which is not centered on SCA's, but which attempts to solve the symbol grounding problem by grounding words in perception and action, and for a review the interested reader is referred to [Roy2005]. Furthermore, notice that here, for the case of the SCA's proposed in this thesis, the words do not directly connect to the world: in most cases the intermediary is the GSM of the SCA. For example, the word "apple" will first connect with the appropriate object within the situation model of the agent (GSM), and then secondly the GSM object will connect to the external world, mediated by the senses. More on this "indirect" grounding will be given in the context of the semiotic triangle in chapter 2.

Having discussed the motivation behind language grounding, a number of questions naturally follow: first, if we have to ground the meaning of words to something external to words (i.e. not words themselves), what should this external thing be? Second, what are the limits of this approach? Can we cover the whole of natural language, and if yes, how? We will delve into a deeper consideration of these questions in chapter 8, where we will introduce various possible "meaning spaces" (section 8.1.1), and also propose a

[3]This is essentially a consequence of the phenomenon known as semantic holism.

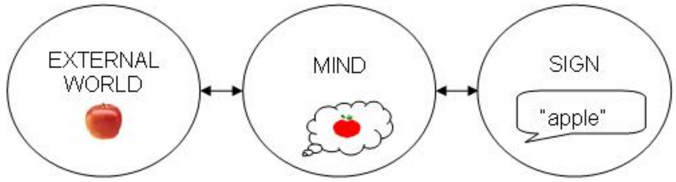


Figure 1-9: Example of "indirect grounding": the word "apple" first connects with the appropriate object within the mind (the situation model of the agent), and then secondarily the mental object connects to the external world indirectly, mediated by the senses

three-step long-term plan for covering the whole of natural language (section 8.1.3).

1.3.2 Situated Language

Although we may eventually want robots that can converse about a range of abstract topics, a natural starting point is to develop means for robots to talk about their immediate physical and social context. This parallels the development of semantics from concrete to abstract domains in child language acquisition [Snow1972]. Thus, in the proposed systems, language about directly observable and manipulable objects (situated language) preceeds language about distant and imagined situations or abstract concepts. For example: Understanding the utterance "this apple" which are being uttered by somebody sitting on a table on which there is an apple, falls within the realm of situated language. On the other hand, understanding "imagine a democracy without voting" definitely falls outside - it is concerned with imaginary situations (not current) and involves abstract concepts such as "democracy". Humans indeed comprehend the first of the above statements much before the second during their development.

This *gradual progression of linguistic abilities*, i.e. starting from utterances about the here-and-now, and moving towards more spatiotemporally detached and abstract utterances (about the past, about imagined situations etc), is not only *evident in humans*, but was also *manifested in the stages of our development of the robotic system Ripley* that is presented in chapter 6 [4]. In its initial stages [Hsiao2003] the robot was only able to understand language about the here and now, and specifically language that was furthermore constrained by direct accessibility to its senses (no internalized "map" of the environment containing persistent objects was available at that stage of the development of the robot - words connected directly to sensory signals). Later, in the stage described in [Roy et al.2004], the robot became able to speak about persistent objects that were recently seen but were not directly visible any more because the robot was currently looking elsewhere (through a representation which we back then had called a "mental model", a predecessor of the GSM, which provided indirect grounding). Finally, in the currently implemented GSM-based system [Mavridis and Roy2006b], the robot is able to understand and speak about the present (accessible or not), the past, and even about imagined situations. In this way, the discrete and designer-imposed

[4] of course, the development here was discrete and designer-imposed and not continuous and natural as is the case in humans.

stages of the development of the robot Ripley, roughly aligned with the empirically evidenced progression of linguistic abilities of humans from the "here-and-now" to the further spatiotemporally detached, and the imaginary[5].

But what are the implications and requirements of situated language? In order to get a better understanding of these, consider partners engaged in a collaborative building task. The partners will often refer to physical objects that are available to them, their locations, and their properties. Each person knows about the objects through perception, can act upon the objects through manipulation, and can talk about the objects that are present. But then perception, action, and language are aligned by the structure of the physical environment. However, when trying to build an artifact such as a conversational robot, we must address a basic challenge: how should different subsystems such as visual perception, motor control, and language interface with each other? This question will re-appear as one pathway providing motivation for the GSM proposal in chapter 4.

1.3.3 Situation Models

In the cognitive science literature, various experiments have tried to address the nature of the representations that are created during human story understanding ("situation models" in [Zwaan and Randvansky1998]), as well as mental inference ("mental models" in [Johnson-Laird1983]). Also, Marvin Minsky has briefly touched upon "mental models" in his well-known "Society of Mind" [Minsky1988], and also in the more recent "Emotion Machine" [Minsky2006].

Basic evidence for the existence of situation models

One of the classic results that has led to the kind of research reviewed in [Zwaan and Randvansky1998] is that the representations involved are neither image-like nor lists of verbal propositions, but are organized along dimensions corresponding to story space, time, protagonist, causality etc., and not their real counterparts (space, time etc. of the actual narration - not the story).

For example: imagine that person S narrates a story to person H. The story is: U1: "After a long day of work, John entered his studio, and left his umbrella next to the door." U2: "He ran to the bathroom, to perform his ritual nightly cleansing of his face, in order to remove all the coal and dirt that he had accumulated during his long day of work at the mines." U3: "He then went to his sofa, grabbed the remote, and turned on his TV"

Now imagine the following two cases:

C1) After U2, H is asked a question about the umbrella, and reaction time is measured.

C2) After U3, the same question is asked to H, and again RT is measured.

Interestingly enough, the availability of information regarding the umbrella is higher after U3, then it is after U2. But why is this the case? If the story was internally represented in H in some form where the time of the actual narration was a main axis, as

[5]Also look at section 8.1.4, where the levels of detachment from situatedness are discussed in more detail.

would for example be the case if the story was represented through strings of the actual words heard, then one would expect, that as the word "umbrella" was heard more recently at the temporal instant after U2 as compared to the temporal instant after U3, then the word "umbrella" should be more available after U2 then after U3. But if the story is represented in some different form, where the token corresponding to the umbrella is connected to other information not in terms of the actual time of the narration, but in terms of story space and time, then it might be the case that because in story space John is closer to the umbrella (in the same room) after U3 as compared to after U2 (where he is in the bathroom), the umbrella becomes more available for recall after U3 and not after U2.

Thus, through the body of research reviewed in [Zwaan and Randvansky1998], we have evidence that story space, time, protagonist and so on form the main organizational axis of the "situation model", i.e. the internal representation that is formed during story understanding. Furthermore, there is evidence that the same representations that are invoked during story understanding are also used for situated language comprehension, this time "filled in" through data derived not only through language, but also through the senses.

Definitions, components, and levels for Zwaan / Randvansky's "Situation Models"

In their latest overview of the theory, which is available at [Zwaan2006], the authors offer the following definition of a situation model:

> "A *situation model* is a mental representation of a described or experienced situation in a real or imaginary world"

In their listing of *situation model components*, they propose four primary types of information composing situation models:

1. A spatial-temporal framework (spatial locations, time frames)
2. Entities (people, objects, ideas, etc.)
3. Properties of entities (color, emotions, goals, shape, etc.)
4. Relational information (spatial, temporal, causal, ownership, kinship, social, etc.)

These components are roughly aligned with those proposed here, with the following though differences: First, in the GSM proposal, the spatial framework does not exist as a separate entity; positions are just one more property of entities[6]. Second, in our proposal, abstract "ideas" cannot directly populate the GSM as entities[7]; they can either populate the GSM through exemplary concrete situations, or can by analogy be "objectified". Also, notice that there is no explicit notion of "embedded situation models"[8]

[6]See section 5.1.3, 6.3.
[7]See section 8.2.3.
[8]See section 5.1.1.

in the Zwaan / Randvansky proposal: for example, goals of other agents become just properties of them, while embedded recursive models have special treatment and form a central aspect of the GSM proposal.

Having seen the *situation model components* proposed by Zwaan and Randansky, let us now briefly visit their proposal for *levels of representation* within situation models. They assume that there are three levels of representation:

1. Situation model (event-specific)
2. Episodic model (coherenent sequences of events)
3. Comprehensive model (a comprehensive collection of episodes)

In the GSM proposal, all three levels are addressed: the situation-type (termed "moment") addresses an instantaneous state of the situation, while the history-type[9] (consisting of a sequence of "moments" and a list of "events" timelining the moments) addresses the episodic model as well as the comprehensive model[10].

Thus, in conclusion, during its development, the GSM proposal has proceeded along similar lines of thinking about *situation model components* and *levels of representation* as the current proposal of Zwaan and Randvansky, with the differences mentioned above: most importantly the no explicit consideration of embedded situation models by Zwaan / Randvansky, and the different view on space.

Johnson Laird's "Mental Models"

On the other hand, Johnson-Laird in [Johnson-Laird1983] has focused on inference through counterfactual reasoning on similar representations, and not on formation of situation models during story understanding. He has also provided a rough taxonomy and overview of mental-models, which again are neither image-like not propositional, and which might be only be filled by tokens representing directly observable physical attributes, but also by token representing conjectured mental attributes of the situation under consideration. Thus, Johnson-Laird introduces "embedded" mental models, which are models of the mental models of others, or even self-reflective models of the self.

Minsky's "Internal Models"

Finally, Minsky in [Minsky1988] and [Minsky2006] introduces many ideas relevant to this thesis's GSM proposal and shares many common views. For example, he speaks about "world models" (a part of which is indeed the Grounded Situation Model of our proposal), about "mental models of other humans" and "mental models of the self" that arise within the "world model" (very much like the "embedded GSMs" proposed here (see foreword and section 4.2.1), and he also decomposes the models of beings in two parts (physical / mental realm), while our proposal as we shall see adds one more part

[9] See section 5.1.4.

[10] Albeit, without a pre-fixed "segmentation" of episodes within the history - as is assumed by the "comprehensive" model.

(physical / mental / interface). Most importantly, he also stresses the importance of the co-existence of a hierarchy of different representations which span multiple levels of description, from symbolic narrative stories to numerical classifiers - which is also the case in our proposal (event histories, three layers of GSM (section 4.2.2]), categorical classifiers etc.). More on the relation of Minsky's ideas with the GSM proposal can be found in section 8.5.9.

Conclusion

Thus, overall, the GSMs proposed in this thesis *share many attributes* with the models discussed in both of the above two sources - they are neither purely image-like nor propositional, they contain representations of both the physical as well as the mental, and they contain "embedded GSMs" representing the estimated GSMs that reside within the mind of others, as seen from the self, they consist of entities, properties, and relations, and both instantaneous ("moments") levels as well as episodic ("history") are provided. However, there exist also some small-scale differences explicated above between the GSM proposal and the cited work, and also, most importantly, *some big differences* between the above proposals and the GSM proposal:

- The focus on building *real-world situated conversational assistants* for the case of the GSM proposal vs. *human cognitive modelling* for the above cited work
- The level of *computational specificity* of GSMs and the availability of *real-world implementations* such as Ripley the Robot
- The fact that the above proposals provide no explicit consideration of the *processes* that deal with the *bidirectional interfacing* of sensory data and linguistic data to the GSM, which we do here

1.4 Why NLP cannot "plug and play"

Having discussed Situated Conversational Assistants (SCAs), and the three relevant notions of Symbol Grounding, Situated Language, and Situation Models, we will now to address an important motivation behind the introduction of Grounded Situation Models (GSMs): the fact that traditional Natural Language Processing (NLP) techniques cannot "plug and play" when creating conversational robots - and thus GSMs, as we shall see, come to fill in the void.

The development of natural language processing technologies and robotics has proceeded with relatively little interaction. NLP deals with the discrete, symbolic world of words and sentences whereas robotics must confront the noisy, uncertain nature of physically embodied systems with sensory-motor grounded interaction. Current computational models of semantics used in NLP are variants of "dictionary definitions", essentially structured networks of word-like symbols, and thus suffer from the symbol grounding problem that was mentioned above. It is impossible to directly apply these NLP approaches in any principled way to endow robots with linguistic skills since the underlying theories of semantics in symbol-based NLP provides no appropriate "hooks" for action and perception. Let us consider some specific examples of problematic cases.

As a first example, consider trying to hook a semantic network such as a subset of "wordnet" [Fellbaum1998] to the world. The entry for "apple" in wordnet looks like this: (without following all branches of the subtree)
Sense 1 apple – (fruit with red or yellow or green skin and crisp whitish flesh)

PART OF: apple, orchard apple tree, Malus pumila
What does "red" mean in terms of sensory expectations? What about "green" or "skin"? All of this information is *not* encoded in wordnet, it depends on the sensors of the agent, and has to be learned somehow. Anyway, suppose that we have a vision system that through exemplars has been trained to recognize "apples". Would this be enough? Notice that "apple" is a generic - the conversational agent must be able to deal with specifics, present or past, fictional or true, such as "this apple" or "the small one", "the apple that you saw yesterday" or even "the apple Eve gave to Adam". But to do so it must have *much more* than just a simple visual-classifier-augmented wordnet entry: it needs *structured models of the specific situations* within which the referred apple exists, a *way to update the models* through sensory or linguistic information, a *way to resolve temporal and object references*, a way to quantify and deal with *uncertainty*, and much more. All of these, and much more, the proposed GSM-based SCA architecture provides.

But let us continue the above example. Let us suppose now that we are informed that "a green sphere is behind you", and that we had wordnet and a classifier for "green spheres" or even maybe a classifier for "spheres" and one for "green things". Can we readily deal with the requirements of representing the information derived by this statement, in a way that will enable later sensory verification and augmentation, or question answering ("where is the green one") and physical actuation ("pick up the green one!"). We have wordnet and the two classifiers ("green" and "sphere"). Where do we start? All the problems stated in the preceeding paragraph still have to be solved, and even more - for example, one has to deal with the underspecification of language as compared to sensory evidence. I.e. if we hear that "a green sphere is behind you" we know something about the color - that it is "green" but we don't know the exact shade of green it is, which we would if we could see it. We also know something about the position ("behind you") but again much less precisely as compared to sensory evidence etc.

From the above examples, it should have become clear to the reader that no "obvious" way of interconnecting existing vision systems with the usual "semantic dictionaries" (such as wordnet-style semantic networks) and motor systems exists. This is exactly where the proposed GSM-centric methodology enters the picture: it provides a *systematic approach* towards building systems that fluidly integrate language with perception and action, which furthermore can also be designed in a way that provides testable quasi-minimality given a behavioral specification (as shown in appendix D). The proposed GSM design is able to provide a viable solution to the problems exposed above, and also deal with the problem of the underspecification of language as compared to sensory evidence its triple layered-structure that enables information of differing granularity (sensory vs. linguistic) to habitate the GSM.

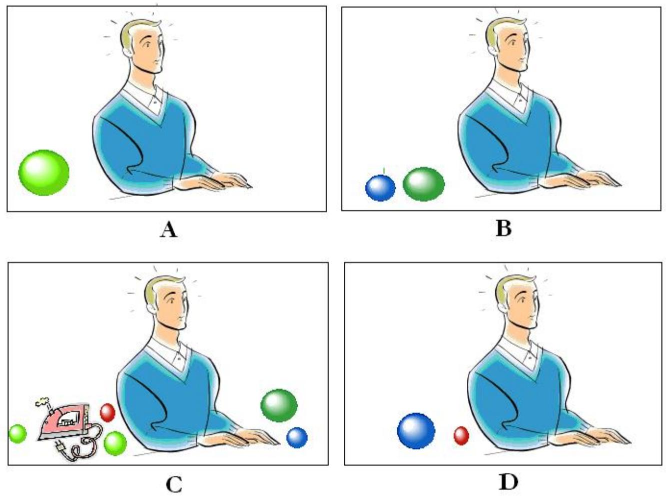


Figure 1-10: Example of the underspecification of language - all (a), (b) and (c) are valid situations given "a green sphere is behind you", while (d) isn't.

1.5 Structure of this document

Below follows an overview of the structure of this document. There are roughly three parts to this thesis: introductory (chapters 1-2), core (chapters 3-6, including agent model, GSM representations/processes, implemented system), and extended (chapters 7-9, covering evaluation, systematic design, semantics proposals and reflections, leveraged extensions, and conclusion).

In the opening chapter, we will first discuss the vision of truly cooperative SCA's, and then introduce three notions from the existing literature that are highly relevant to the GSM proposal: symbol grounding, situated language, and situation models. Then, we will address the question of why traditional NLP methods cannot "plug-and-play" with other subsystems in order to create conversational robots - one large void that GSMs propose to fill. In Chapter 2, a big-picture viewpoint will be taken, positioning our discussion within a version of the "Semiotic Triangle", and discussing signs, minds and worlds.

Having explicated some of the main positions implied regarding descriptions of external "objective" reality, the position of life and purposeful agents within such models, the notion of "minds" and Cognitive Architectures and the ultimate purpose of minds, the need for the emergence of internal representations etc. in chapter 2, we will proceed to the GSM proposal in the subsequent chapters. We will do so by progressively focusing from generals to particulars: starting from generic agent models (Chapter 3), moving on to a derivation and description of a customizable GSM (which can be thought of as covering three central representations of the agent model - past / current / future situation reps, and related processes) in Chapters 4 and 5, to the specific implementation of a GSM customized for Ripley the Robot and a given behavioral specification in Chapter 6.

In a little more detail: in Chapter 4, the notion of Grounded Situation Models will be motivated and introduced, and two most important aspects of the GSM proposal

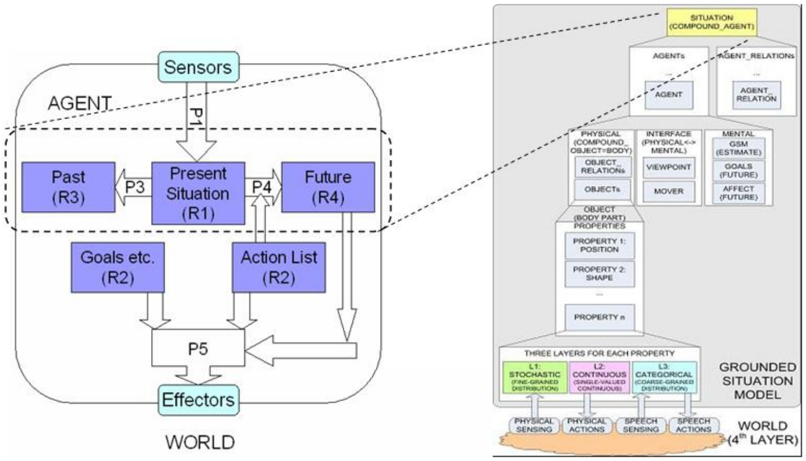


Figure 1-11: Relationship between the proposed agent-model (Chapter 3) and the proposed GSM representation (Chapters 4 and 5): Three central parts of the agent model also contain representations of "situation" type, and the GSM proposal also includes processes for keeping track of past situations and expecting the future.

(form of representational hierarchy, triple-layered descriptions) will be derived from two desiderata: NL-like parsing of situations, and bi-directionality between the senses and language. In Chapter 5, the specifics of the proposal will be explicated: the GSM representations, the associated processes, and the modular implementation architecture of real-world GSM-based SCA's. Later, in Chapter 6, Ripley the Robot will be presented: an operational real-world conversational assistive manipulator robot, based on GSM's.

Subsequently, in Chapter 7, the question of the evaluation of SCAs will be discussed. Also note that in Appendix D, a rough-sketch of a design method for quasi-minimal GSM-based SCAs will also be given, and an example of its application discussed. In Chapter 8, some reflections on current semantic theories and other topics will be given, on the basis of our experiences with Ripley - this discussion includes, among many others, an interesting proposal on empirically sound semantics. Finally, in chapter 9, an extensive discussion of future extensions readily leveraged by the current system takes place, together with a concluding section. We hope that reading this text will prove enjoyable, as well as illuminating and useful for the readers that will invest their time into studying it.

1.6 Recap

In this chapter, we first started by giving a brief description of five main aspects of thus thesis: Vision, Steps, News, Contributions, and Salient Idea. Then, we discussed the vision of truly cooperative SCA's in further detail. Later, we examined in detail three notions relevant to the GSM proposal: Symbol Grounding, Situated Langauge, and Situation Models. After that, we addressed the important motivational question of why traditional NLP methods cannot "plug-and-play" with other subsystems in order to create conversational robots. Finally, we provided a roadmap through the structure of the rest of this document.

In the next chapter (2), a big-picture viewpoint will be taken, positioning our discussion within a version of the "Semiotic Triangle", and discussing signs, minds and worlds, before we introduce a GSM-based agent model and then motivate and introduce Grounded Situation Models in the subsequent chapters.

Chapter 2

Signs, Minds and Worlds

In this chapter, we will take a big-picture viewpoint, and position our discussion within a version of the "Semiotic Triangle" - there, we will discuss models of signs, minds and worlds. We will start by introducing the semiotic triangle, and by re-examining the previously introduced notions of Symbol Grounding, Situated Language and Situation Models, within this new framework (section 2.1).

The purpose of the rest of the chapter will be to serve as a prelude to the proposed GSM-based Agent Model in the next chapter. The GSM-based Agent Model that will be introduced in the next chapter is a computational model of an artificial mind, which connects to the physical world through sensing and acting, and which includes an internal representation of current, past or imagined situations (the Grounded Situation Model) as a central component of the mind model. Thus, before introducing the proposed GSM-based Agent Model, we will here provide an initial discussion of related topics, such as: *purpose* and *goals*, *minds*, *sensing* and *acting*, *action selection*, types and purpose of *internal representations*, *prediction* etc.

We will do so by starting from descriptions of "objective reality" in the section named "models of the world" (section 2.2). Then, we have to relate minds to this picture; we will do so in two steps: first, we will introduce the concepts of "teleology" (purpose and goals) and "life" (section 2.3) and relate them to models of "objective reality", and then, we will move from the "physical" to the "mental", and will talk about "models of minds" (section 2.4), where the minds under consideration can be either natural or artificial (as is the case in AI and in our implemented system). Within our discussion of mind models, we will also specifically talk about internal representation, prediction, and action selection, and this lengthy section will close with a quick presentation of some existing models of minds - coming from a variety of directions and addressing differing needs. Overall, this chapter will provide a proper prelude for the introduction in the next chapter of the overall agent architecture framework within which the proposed GSM-based Agent Model will be positioned.

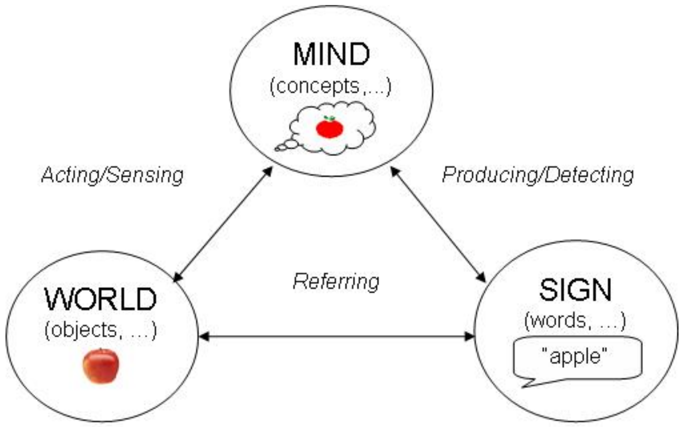


Figure 2-1: Signs, Minds and Worlds

2.1 A Higher View: The Semiotic Triangle

2.1.1 Introducing the semiotic triangle

Taking a higher-level view of the previous sections on Symbol Grounding, Situated Language and Situation Models, one can isolate three main realms that participate in the big picture:

R1) WORLD: The external reality, part of which is the situation in which the speakers are embedded (roughly equivalent to "physical realm" where real objects and events reside)

R2) MIND: The mind of the speaker, in which an internal "situation model" reflecting the external situation resides (roughly equivalent to "mental realm" where internal concepts corresponding to external objects and events reside, as well as other "mental stuff", such as intentions etc.)

R3) SIGNS: The communicative signs used (if mainly verbal, then roughly equivalent to the "linguistic realm" where words, phrases etc. reside)

This "big picture" is now reminiscent of the standard "semiotic triangle"[1], and can be easily aligned with it (figure 2-1).

Let us assume now that we take an idealistic "external observer" view of the world, being isolated from it but having full visibility (with infinitesimal informational granularity) of it. Then, we could decide to fix:

D1) a method of description of the world (external reality[2]

D2) a method of description of the mind of the speakers

D3) a method of description of the symbols used

[1]Usually attributed to Charles Sanders Pierce, although similar ideas go all the way back to Aristotle. For a wider discussion, see for example [Ogden and Richards1989].

[2]i.e. "objective" reality in layman's terms - although the notion of a totally objective reality is of course an idealized notion, and any realistic description of reality can just be more or less "subjective".

2.1.2 Symbol grounding, situated language, and situation models, viewed in terms of the semiotic triangle

Within this new framework, one can re-examine the previously introduced notions of Symbol Grounding, Situated Language and Situation Models:

Symbol Grounding

The relation between elements of D3 and elements of D1 ("direct physical grounding"). For example: the sign "apple" corresponds to a real object. Alternatively, one could talk about "indirect grounding mediated by the mind" (from signs to the world through the mind), i.e. primarily the relation between elements of D3 and elements of D2, and secondarily the relation between D2 and D1 (see figure 1-9 from chapter 1). In this case, "apple" corresponds to a thing in my mind, which might in turn have been caused by the image of an apple falling on my sensors. Notice that under the heavy simplifying assumptions that:

- there is complete knowledge of the local situation and
- any perspective-dependent subjectivities can be ignored,

then it follows that the relevant elements of D2 (mind) might be idealized as mirroring D1 (external reality) perfectly, and then "direct physical grounding" becomes almost equivalent to "indirect grounding mediated by the mind"[3].

Situated Language

Language about the spatio-temporally local physical situation in which the speakers are embedded, i.e. the case where we can pretty much talk about "direct physical grounding" in the here-and-now. For example: "this apple is red" referring to a physical apple in the here and now is a statement of situated language, while "yesterday the sky was blue" is not, and even more so: "the intricacies of democracy as well as its systematic abuse by the economic powers will pervert the meaning of this political system as it is represented in the average mind" is not an example of situated language.

Notice that non-situated language[4] is arguably always ultimately based on recall / combinatorial resynthesis of previous situated experience[5]. For example, consider: language describing a past experience ("yesterday I saw a blue bird"), or an imaginary situation ("how I would like to live in a spaceship made of chocolate!"). Obviously, what enables language to detach from the senses and transcend the limits of space, time, and actual existence is the mediation of the mind[6].

---

[3]Else, as we shall see, if we undo the simplifying assumptions of complete objective knowledge, the ExternalReality-to-SituationModel process can be broken down as proposed in section 3.2.1.

[4]For a proposal of a set of levels of detachment from situatedness, see section 8.1.4.

[5]For a lengthier argumentation supporting this statement, see section 8.2.5.

[6]Arguably, reversing the cause-effect relationship, it has been proposed that it might actually be *language* that enables the reshuffling / compositional resynthesis across sensory realms in humans - see for example [Hermer-Vazquez et al.1998].

Situation Models

The part of D2 (the description of the mind) that contains the "internally cached" versions of external reality, and which serves as the primary field for the intermediate representation required for "indirect grounding mediated by the mind". These models might be filled by present / past / imagined situations, for example when talking about the concrete here-and-now / recalling a past memory / imagining a fairy tale that we hear.

Having assumed the free choice of some description method for D1-D3, we have nevertheless not specified anything about the method. In order to proceed further, we must now ask: What are possible / convenient / suitable ways of modeling external reality (D1) / internalized situations (D2) / sign sequences (D3)? Also, a relevant cognitive science question is: how do humans model internalized situations, i.e. what is the structure of the situation models that humans posses[7]? Here, I will attempt an initial exploration of the first question, in order to derive comments useful for the purpose of this thesis from the answers given.

2.2 Models of the world

Assuming the existence of some at least partially objective "external reality" is a prerequisite for the higher view capacitated by the semiotic triangle. However, at this stage one might ask: what type of model should one choose for representing external reality in a computational theory of semiotics? Even before that, one could ask: what different types and species of such models exist? Here, we will consider some such models, arising from different directions: first, from *physics* - starting from a simplistic newtonian worldview. Then, we will consider a model coming from *semantics*: possible-world formalisms. Finally, we will discuss what the world model implied by the way that natural langauges describe the world is - the *"folk ontology"* of the world, in other words[8].

2.2.1 A simplistic Newtonian model

Let us start with a simplistic example: a simplified finite point-mass newtonian worldview - in which, only a finite number of point masses exist in the whole universe, and they are only under the influence of the newtonian gravitational forces that arise between them. Some questions and the corresponding answers that can be asked when considering such a model are:

- What exists?

Primarily, a finite number of point masses. Secondarily, their properties: their positions, their speeds, maybe also forces, maybe also trajectories if we extend beyond a single time instant. Due to the nomic interdependencies[9], unless one assumes an or-

[7] In a vein similar to [Zwaan and Randvansky1998].

[8] Notice that the possible ways of describing external "objective" reality are also options for describing internalized situations; and indeed in the next chapter we will choose this "folk ontology" of the world when presenting the proposed representations comprising Grounded Situation Models.

[9] For example: if we know mass and force, we can derive acceleration, but notice that also: if we know mass and acceleration, we can derive force etc.

dering of primary vs. secondary existence, this question is not precisely posed. Thus, a better question might be:

- What should one minimally know in order to completely specify the state of the world at a given instant?

In the case of a finite number of objects (N) with point masses, in order to completely specify the system in a single time instant, one such minimal set of required knowledge consists of: a set of N 3-D vectors specifying positions, as well as a set of N 3-D vectors specifying accelarations. Of course, this set is not the sole minimal set that would completely specify the state in a single time instant.

Moving on from a single instant to the whole lifetime of this universe, one might ask:

- What information is enough to uniquely specify the whole lifetime of this universe?

Because of the dependencies caused by the newtonian laws, if one knows completely the kinetic state of this universe in a single time instant and also knows the masses of the objects, then he can in principle integrate forward and backward towards all time in order to predict / reconstruct the whole lifetime. Thus, the set consisting of the positions, accelarations and masses contains enough information to uniquely determine the whole unfolding of this universe.

Thus, we have seen a first simplistic example of a model of external reality (Newtonian point-mass universe), and touched upon two questions for this model: what exists? and what should one minimally know in order to completely specify the state of the world at a given instant / at all times?

2.2.2 Other models of external reality

Now, let us try to briefly move on to other models of external reality, by either relaxing some assumptions, and/or exploring different modeling directions. One possible direction would be to move from a finite to a potentially infinite number of objects. Another would be to extend point masses to mass distributions, or extend to different force fields apart from gravitation. Yet another would be to abandon the idealized "complete" knower that knows precisely the required information, over all space and all time, and create a hierarchy (partial ordering) of "incomplete" knowers, who possess partial, ambiguous and / or noisy information[10]. Here, some notion of "informational granularity" might also enter the picture: one might know more "grossly" than another, who knows more "finely"[11]. Also, when complete knowledge is not always possessed, or once one abandons universes with deterministic laws (due to practical or theoretical reasons), one is forced to entertain multiple possibilities for the present / future. Furthermore, notions of theoretical or practical computability might impose further restrictions.

Model-theoretic semantics, possible-worlds formalisms

Now let us consider a different tradition: semantics. What are the characteristics of the world models used in semantics? Usually such models presuppose a world that is made up of discrete entities (such as individuals), which enter in relations with each other.

[10] See for example discussion on knowers K1-K4 on page 7 of [Mavridis2005b],

[11] We will return to this idea in section 3.2.3, when discussing alignment of GSM's across different agents.

In the case of possible worlds semantics, there is also a second species of relations - "modal" accesibility relations, which connect possible states of the world with other possible states. For example, look at figure 2-2.

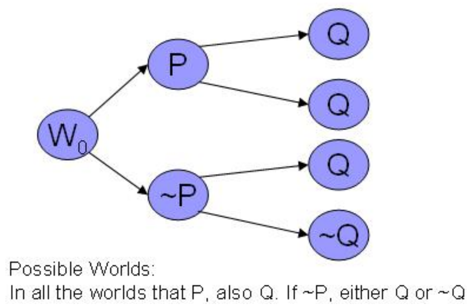


Figure 2-2: A fragment from a possible worlds model

In the example depicted in the figure, we do not know whether P holds or not-P holds. However, we do know in the case that P holds, Q should hold too; while in the case that not-P holds, Q might hold or not-Q might hold.

What are the difference of such models as compared to a classical physics model for example? Numerous and important. First, in a usual possible-worlds model, the world is already discretized into entities (individuals/objects) and relations, which are furthermore usually defined using natural language terms ("Tom", "apple") etc. In contrast, in the classical physics view, possesing mass distributions does not provide an unambiguously defined way to "break down" the world into objects. Second, in a possible-worlds model, continuous values (which figure prominently in classical physics world descriptions) are not used; everything is discretized and turned into crisp logical sentences. Third, while classical physics assumes determinism and complete knowledge, a possible-worlds world description can easily handle partial knowledge and stochasticity. Fourth, while the spatial frame has a primary position within the classical mechanics worldview, it does not within a possible-world semantics formalism. To recap: inherent objectification vs. a continuum, crisp logical sentences vs. continuous values, partial knowledge vs. complete knowledge, stochasticity vs. determinism, and finally no special treatment of space vs. primary importance of the spatial frame, are some important differences between possible-world semantics and classical physics worldviews.

The folk ontology

Yet another direction that one can take is the following: forgetting about the models we might construct in physics or semantics, what about the folk worldview that is implied by / embedded in the way that natural languages carve out the world? How does human language carve out the world? This is the direction that we will take in this thesis, when crafting an internal representation of situations for Grounded Situation Models (see foreword and section 4.2.1). But before we attempt to look at some answers to this question, which we will give in section 4.2.1, we are missing one main prerequisite in order to do so: the notion of life and purposeful agency, which we will briefly visit in the next section, before later presenting models of minds and notions related to such models: sensing, acting etc.

2.3 Introducing life in models of the world

2.3.1 "Process" vs. "Purpose"-type descriptions

Biological, behavioral and sociological models are vastly different from purely physical models, as they depend on the introduction of "higher level" notions, that slowly elevate from descriptions of processes or interactions among parts, to the level of function and most importantly, purpose. What is the difference between "process" and "purpose" in the sense used here? As an example: on the one hand, a Newtonian equation describes the infinitesimal changes in the state of the world that occurs when time changes by an infinitesimal amount. This is what I call a "process" law - it is concerned with changes across neighboring time instances. On the other hand, a "teleological" (purpose-oriented) law looks is concerned with the longer-range picture - dealing with time stretches that might ideally stretch to infinity (if the system is left unperturbed). For example, the "purposeful" view of gravity on earth, is concerned with the steady-state: gravity "wants" all objects to end up on the ground. Here, we are not concerned with how this will happen (the exact trajectory, rate of change of position etc.) - only with the "telic" state. Most "functional" explanations are similar in nature: the purpose of the heart is to sustain the circulation of blood etc.

Thus, to recap, the first comment made so far was that one can distinguish two "gross" categories of laws: "process" laws, mechanistically describing the temporal evolution of things, versus "teleological-functional" laws, dealing with purposes of phenomena - "steady-state" results that occur in case the system was left unperturbed. But how does all this relate to the introduction of life in models of the world?

The "goal hierarchy of life"

My position here is that in order to introduce life in models of the world, one unavoidably will end up moving from "process"-type explanations to "teleological". Any definition of life will have to call upon some "goal hierarchy" for living organisms. What do I mean by this term? A short explanation follows. Let us first focus on a single living being and its bloodline, then to the species where it belongs, and then to life as a whole, viewed as a unitary being.

A "living thing" is often defined[12] as a unity-preserving structure, where the ultimate "purpose" is the preservation and multiplication of its "structural footprint" across time. This ultimate goal of preservation of the single living being and its bloodline breaks down to two subgoals: preservation of the individual and creation of offspring - which

[12]Many alternative definitions exist, others more specific to already known forms of carbon-based life, others more abstract and informational: for example, entropy-based arguments. For the purpose of this exposition, we have chosen to focus on a definition based on "purpose".

are often contradictory in a limited-resource environment, and thus death of the individual at a certain age might be beneficial to the (being, bloodline) set as a whole. Preservation of the individual can in turn be broken down to multiple subgoals, more-or-less shared across organisms: supplying yourself with nutrients, getting rid of waste, moving away from dangerous environments etc. Now, let us leave the (being, bloodline) level, and move up to viewing all life as the all-encompassing unitary living being. Thus, we also move to the top level of the goal hierarchy: preservation of the (being, bloodline) is subserved to a higher goal itself; to preservation of life as a whole, where some leafs of the tree of life might be sacrificed for the sake of the tree - thus, justifying altruistic actions even outside expectations of personal future return of benefits, or return to the bloodline or the species. I.e. the "preservation of the individual / replication" level of the goal hierarchy should be augmented with "helping others", in order to preserve life as a whole.

This was thus a short description of what I mean by the "shared goal hierarchy of life"[13]. Recapping, in a top-down view, and also introducing a species-level:

- *Top level*: Preservation of life as a whole,
- *Species level*: Preservation of the species and helping other species,
- *Bloodline level*: Preservation of the individual/replication and helping other individuals,
- *Personal level*: Breaking down to subgoals like finding food, getting rid of waste, avoiding danger etc.

From goals of life to goals of artifacts

Now let's move on from living beings to human-made artifacts. In this case, the "goals" of the artifact are given by the human designer - or at least the intended goals. Again, notice that ultimately, even this "designer-chosen" purpose of the artifact, should ideally be finally subservient to the goals of life - the artifact is built through a set of actions of a living organism (the human designer), that being either misguided or not, should ideally have attempted to serve the ultimate purpose of life - as any other of his actions[14].

2.3.2 The object/purpose "dual" of the semiotic triangle

But how is all this relevant to our original discussion - to the semiotic triangle, to situated language and so on? My position here is that: once one introduces living beings (or agency) in the models of the world used in any computational theory of semiotics,

[13] Notice that the proposed hierarchy has survival of all life (thought of as a unitary being) at the top level, survival of the individual, replication and helping others at the second, and having enough nutrients, being away from danger, getting rid of waste etc. at the third. In this respect, the third level is reminiscent of the lowest "physiological" level of Maslow's hierarchy of needs [Maslow1943]. However, there are many differences: for example, Maslow's hierarchy is centered on the single human; while the goal hierarchy of life proposed here is centered on the totality of all living beings, viewed as a unitary being, etc.

[14] Here we are not claiming that all human actions, either consciously selected or not, indeed serve the ultimate purpose of life, just that ideally they should be subservient to this ultimate goal, which they often might not be.

the introduction of the notions of "goal", "purpose", "intention" and the such becomes almost unavoidable. This actually creates an interesting (although asymmetric) duality - behind the "original" semiotic triangle, one can imagine a dual triangle where the units are not anymore "objects" (physical objects, mental counterparts, words) but "parts of goals".

Referential vs. functional meaning

Let me be a little more specific on this. Consider language: it is often said that utterances can be seen as having both a *"referential"* as well as a *"functional"* meaning. For example, the *referential* meaning of the utterance "it is hot in here" is a pointer to a specific aspect of the current physical situation - more precisely, to the thermal state of the current situation, that ultimately might refer to the average kinetic speed of the molecules in a physical model of the world. On the other hand, if this utterance is said in a room with an air conditioner that is turned off and in the presence of somebody else that can open the air conditioner, the *functional* (goal-oriented) view of the utterance might be the following: by saying the utterance, the speaker acts in this way (moves his lips and vocal chords) in order to achieve the GOAL of the lowering of the temperature of the room. This is not achieved *directly*; the movement of air caused by his utterance will probably not alter the room temperature significantly - however, it is achieved *indirectly*, through the mediation of a listener: the listener, upon listening the utterance, will most probably turn on the air conditioner[15]. Thus, while the "referential" meaning of the utterance deals with a "state" of the situation, the "functional" meaning can be viewed as a "step" towards a "goal". Of course "goals" are again described in terms of situations - they are classifiers of situations (classifying them in favorable and unfavorable ones) - and thus the situation/goal asymmetric duality arises.

A practical application of this goal-oriented viewpoint is exposed in the work of Gorniak [Gorniak2005]. Gorniak studied situated language used by human players cooperating with each other while playing computer games. The situation in which the human players were embedded was predefined, and there was track-keeping of the state of the situation while they were playing. Furthermore, toy-goals were given to the human (similar to "open the blue chest"), and thus given the simple environment, the set of all possible plans that led to the satisfaction of the goal could be explored. Thus, in such a framework, words could be connected to both the physical situation as well as the goals and plans of the speakers.

[15]The capacities for action upon them that physical objects have are also called "affordances" [Gibson1977] - for example, some of the affordances of an apple include: eatable, throwable, smellable etc. Extending this concept from passive objects to humans, we have "social affordances" - for example, in the above example, we are exploiting the social affordance of the listener which deals with his capacity to turn on machines. Thus the analogy follows: while physical action exploits the affordances of passive objects towards goals, speech acts exploit social affordances of humans.

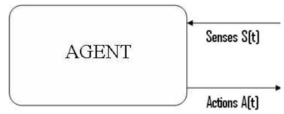


Figure 2-3: Agent as a black-box model

2.4 Models of Minds

2.4.1 What is a mind?

Having briefly touched upon some ideas related to the first vertex of the semiotic triangle (models of the world / external reality, including living beings), it is now time to move on to the second vertex: minds. The first question to ask is: what is a mind?

"Brains" and "Behavior"

We know from our experience that many organisms have brains - which are material organs, made of flesh and blood, that we know can guide their behavior. Of course, externally observable behavior is not only affected by the brain itself: reflexive properties of the neural system, or even physical properties of the human body can also create observables that might be thought of as "behaviors": for example, a bubble appears on my hand after it has gotten burnt, without any intervention of the brain. So it seems that we have touched upon two terms so far: "brain" and "behavior", and it seems that in higher organisms a big part of behavior is mediated by "brains".

The black-box idealization of minds

In order to introduce minds and models of minds, I will invoke an idealization. First of all, I will decompose "behavior" to a set of idealized, molecularised "actions". Taking a systems viewpoint, I will model the mind as a "black box" (figure 2-3), that produces such "actions" at specific time instants. I will assume that the only output of this black box are these actions. I will also assume that the only input to this box is a "sensory stream" - something that corellates somehow with the spatiotemporally local external reality, into which I will assume that the body carrying the mind being modeled is embedded.

First, notice here that this is a mathematical/informational-level idealization. In this idealization, I am only concerned with: (a) the *overall goal*, as well as (b) the *input-output model* of the mind, and maybe also[16] (c) some internal aspects of how the input-output computation takes place - but in (c) only from a *computational* point of view - NOT from the viewpoint of the *physical implementation*. By being concerned only with (a)-(c), and not with the physical implementation, I am targetting the first two levels of the three levels postulated by david marr [Marr1982], namely the computational and algorithmic levels, but NOT the implementation level. Thus, under the viewpoint taken

[16] If we move from a opaque black-box to a transparent box with more details of the information processing taking place.

here, the mind is an informational idealization of the functionality of parts of the physical body of the organism, and mainly of the brain of the organism in the case of the higher organisms (as in that case the brain determines a big part of "behavior" under normal operating conditions). How much "body" (apart from the "brain") will be included in the informational idealization that we call "mind", depends also on where we "cut out / carve out" the idealized "actions" and "sensory stream" that enter the mind[17]

2.4.2 The purpose of minds

Now, after introducing the "black-box" idealization of the "mind", let us continue our investigation, first by asking: what is the purpose of minds? As an informational idealization of embodied functionality, minds of living beings inherit directly from the "shared goal hierarchy of life": the ideal purpose of the mind becomes to serve the survival of the individual as well as replication, balanced towards the survival of the species, and ultimately, of life as a whole. I.e. the mind, upon accumulating on past genetic experience and sensory-stream derived lifetime experience, should[18] SELECT ACTIONS ACCORDINGLY, in order to serve the shared goal hierarchy of life. Viewed as a commandment, a mind should:

"SELECT WHAT ACTION TO DO, also with the help of your SENSES, in order to SURVIVE and REPLICATE and HELP others"

And always doing so with the ultimate purpose of the survival of all life as a whole.

Restricting the focus from all life to the individual or the bloodline

A more narrow-minded evolutionary view will only center on the individual, and forget about others or life as a while, and would prescribe:

Absolute Egoism
: "SELECT WHAT ACTION TO DO, also with the help of your SENSES, in order to SURVIVE. You might replicate or help others, but only in so far as you can expect a worthy return towards your survival".

This does not preclude replication as a possible action; but from this narrow viewpoint, that takes into account the benefit of the individual only (and not the bloodline or life as a whole), replication would only be justified if the cost of doing so was less than the expected benefits of the help that your children will give you - i.e. from this narrow viewpoint there is no space from real altruism, not even for your offspring - any apparent altruism will be disguised egoism on the basis of expected returns. But, if we center the ultimate goal outside the individual and onto the bloodline, then this changes to:

[17] For example, one degree of freedom might be: how much of the "milieu interieur" (the environment within the organism) should be included the boundary around which we will "cut out" in order to postulate sensory stream signals?

[18] Again, we are not claiming that the mind does select actions accordingly - only that it should.

Bloodline-altruism:

"SELECT WHAT ACTION TO DO, also with the help of your SENSES, in order to SURVIVE and REPLICATE and HELP your children[19].

Again, any help offered outside the bloodline can only be justified in terms of expected returns within the bloodline - i.e. another form of pseudo-altruism.

However, as we have said in the section introducing the "shared goal hierarchy of life" (section 2.3), one could adopt an even wider ultimate goal: preservation of life as a whole. Within that framework, even true altruism is justified, as long as it helps towards the preservation of the unitary all-encompassing living being. Thus, as we said before, the commandment becomes:

Alloflife-altruism:

"SELECT WHAT ACTION TO DO, also with the help of your SENSES, in order to SURVIVE and REPLICATE and HELP others"

This is the purpose of the mind that we would like to suppose for our models here. The mind tries to do the best that it can - although always sub-optimally, even if sometimes better and sometimes worse.

2.4.3 The purpose of artificial minds

Thus, in this viewpoint[20], purposeful action selection is the main thing a mind does. Now what about artificial minds? Once again the "mind" of such an artifact inherits the purpose of the artifact: to serve the purpose intended for it by its human creator, at least ideally. At this point it is also worth stressing, as we have mentioned before, the possibility of "misalignments" of purpose: the "intended purpose" of an artifact might deviate to its actual "steady-state" achieved goal. For example, a malfunctioning spaceship might accomplish the unintended "goal" of splashing into the Atlantic, while its intended purpose was reaching the moon, which it has not accomplished.

Thus, we have succinctly concentrated the purpose of minds as:

Purpose of minds of living beings:

"SELECT WHAT ACTION TO DO, also with the help of your SENSES, in order to SURVIVE and REPLICATE and HELP others"

Purpose of minds of artifacts:

"SELECT WHAT ACTION TO DO, also with the help of your SENSES, in order to SERVE YOUR DESIGNER'S PURPOSE"

[19] Notice that the inclusion of ancestral relatives and not only children in this definition might effectively extend it to the whole of humanity or all life, if we suppose that all humans or all life started from a single couple or a single organism".

[20] This viewpoint shares the fact that it positions a primary focus on action selection with stoicism (see for example [Epictetus1928]), but is also different from it in other aspects.

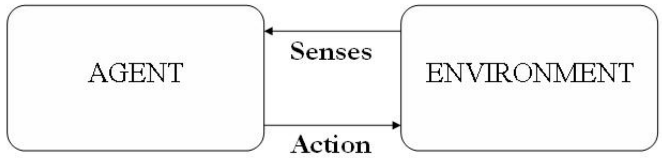


Figure 2-4: Agent and environment as black-box models

Some related questions that we will attempt to touch upon

Now let us briefly comment upon a number of related issues. Unpacking the black-box models of minds and their intended purpose, one can ask the following questions:

- How should actions be selected? What are different types of action-selection processes? [Q1]

- What is there within the black box? Should previous sensory data be forgotten? Where does "internal representation", "affective state" and "situation model" fit in this picture? [Q2]

- What is the position of the notion of "prediction" in the mind model? [Q3]

Let us try to briefly touch upon some of these questions [Q1-Q3].

2.4.4 Actions and Goals

In this section, we will first start visiting the question concerning "action selection" [Q1], which we will also revisit at a later section, after having also talked about "internal representation" [Q2] and "prediction" [Q3]

First, notice that outside the agent, another "black box" system can be postulated: the "environment", as viewed through the action/sensing channels (see figure 2-4)

The question now arises: what actions serve the agent's goals best?

Descriptions of goals

But even before that, at this stage, in order to proceed with some form of an answer to this question, one might try to explicitly describe the goal in some form. For example: I might have designed an artifact whose sole purpose is to fill up its visual field with pixels of red color. Then, I can specify the goal in terms of the sensory stream; when the part of the sensory stream corresponding to vision is filled with contents that correspond to "red color" then the goal is achieved. Thus, I have specified the goal as a partition on sensory stream states; some correspond to "goal achieved", and the others to "goal not achieved". Of course, many possible extensions / different settings exist; for example, the goal might not be specifiable as a condition on a single "moment" in time, but might be specifiable as a condition on a whole "sequence of moments" (i.e. the difference between: "I want the cup on the table" and "I want you to jump up and down three times").

Also, again note that here we have specified goals only in terms of desirable "sensory stream contents". Indeed, this can be all that an agent "directly see", so goals must eventually ground out to the sensory stream. However, as we will see later, the agent might postulate some model of "external reality", and try to explain its sensory stream contents in terms of this model, and thus use sensory stream information as evidence for specific states of external reality[21]. In that case, we might be tempted to specify goals viewed from within the agent as conditions on the state of external reality - but to be more precise, these can only be conditions on the "personal" internalized description of external reality, which however hopefully corellates to the "absolute" external reality through. Nevertheless, once again, these goals are primarily conditions on the sensory stream, and secondarily on the conjectured model of external reality that is supposed to create the observed sensations.

Reachability of goals / multiplicity of solutions

Anyway, no matter how the goal is specified, the following options exist:

a) the goal is never reachable

b) there exists a unique sequence of actions that will reach the goal

c) there exist multiple sequences of actions that will reach the goal

One further complication arises, in the cases where at the time of the start of the agent's actions, the future of the environment cannot be deterministically known, not only by the acting agent, but by anybody. If the future of the environment can be deterministically known by a "superstrong" knower, than indeed one of either a or b or c must hold. If the future of the environment can not, then it might be the case that even after some time has passed from the initiation of the acting agent, it might not be possible to know whether a, b or c holds at that specific time instant - only retrospectively (after the completion of time) we could judge, assuming that what had unfolded stochastically has become known after the completion of time.

Preference orderings among solutions / spreading the "reward" over time

In the case of c), a further question arises: is there a preference ordering among the various possible action sequences that will reach the goal? If there exists, we start nearing the notion of the existence of some "reward" function, that might de-absolutize, decrispify and spread the reward in time. I.e. the other question that arises quickly in this path is the following: how should the "reward" of the agent be distributed in time, as it reaches towards the goal? In the simplest case, one can assume a "crisp" and "momentary" reward function: you get all of the reward exactly at the moment that you reach the goal. Else, one can try to postulate subgoals and intermediate rewards, or other ways to spread the reward over time and distance from completion, so that even partial completion is better than nothing.

[21] For example: through a sequence of 2D views, I might postulate the existence of an external 3D space populated by an object. For such a model to be useful, it should not only explain the 2D views I have seen already, but also predict future views: i.e. both fit the past observations and predict the future observations.

Finite-horizon action selection

If one spreads the reward so that at each action-taking step there is some reward / punishment given, then one can reset the purpose of action generation from:

-> take the best possible sequence of actions that will reach the goal.

to:

-> at each step, take the best possible action, i.e. the action that will maximize the local reward for some steps ahead.

or even to:

-> at each step, take the best possible action, i.e. the action that will maximize the local reward for one step ahead.

Anyway, it should be obvious that in many cases, this single-step local reward approximation will produce sub-optimal results for sequences of multiple steps; nevertheless, this can serve as a viable first approximation.

Necessary and sufficient information for action selection

Having discussed descriptions of goals, reachability of goals, the introduction of "reward" and the spreading of it over time, and successive simplifications that lead to finite-horizon action selection, now we will examine another question, which is centrally important for this thesis, whose main subject is Grounded Situation Models - i.e. one special kind of internal representations. The question is:

What information can the agent use in order to select future actions?

The only input from the world that is available to the agent is its sensory stream, and thus, the information used for action selection has to originally come from there. Thus, the action selection procedure can at most be equivalent to a function that uses the whole sensory stream (past, present, and future!) as an argument:[22]

$A[t+1] = f(S[-\infty \ldots t-1], S[t], S[t+1 \ldots \infty])$ [C1]

Nevertheless, in all practical cases $S[t+1 \ldots \infty]$ is not available to the agent; only approximate predictions about it can be made, which will again need to be based on the past $S[-\infty \ldots t-1]$ and the present S[t]. Thus, at most, we can have:

$A[t+1] = f(S[-\infty \ldots t-1], S[t])$ [C2]

If we further assume an initial "turn-on time" t0 or "birth" for the agent, we are constrained to:

$A[t+1] = f(S[t0 \ldots t-1], S[t])$ [C3]

Notice that the above action selection function requires a "memory" for keeping past states of the sensory stream, or at least some transformation of them that contains all the "essential" for action selection information that can be distilled out of them. If this "internal state" memory has finite capacity of k time slots, then we might have a case where:

$A[t+1] = f(S[t-k \ldots t-1], S[t])$ [C4]

[22] Here, for simplicity, we will start with a discrete time index; the above can be rewritten with continuous-time actions and sensory streams. Also, look at [Mavridis2005b] for a relevant discussion of knowers and incomplete knowledge.

Finally, if no such "internal state" can be kept at all, than the agent is constrained to performing action selection only on the basis of the current sensory input, i.e.:

$A[t+1] = f(S[t])$ [C5]

Notice that the above "action selection" functions [C1-C5] are ordered by "optimality"; we can expect C5 to be potentially less-optimal than C4 and so on, as less information (where the missing information might be crucial!) is available to be taken into account in each case.

Recap

Now, let us try to recap, to see where we are in terms of the original question regarding action selection [Q1].

We have started by briefly touching upon possible descriptions of concrete goals, such as wanting-to-just-see-red (a goal specified as a condition on the sensory stream), wanting-the-cup-to-be-on-the-table and wanting-john-tobe-jumping (goals specified as conditions on external reality, as modelled from within the agent through sensory stream evidence). Then, we briefly talked upon the reachability of such goals, and the multiplicity of possible solutions to them, moving forward to the issue of preference orderings among solutions, and to the "spreading" of the "reward" of the reaching of the goal over time, so that the agent can get some satisfaction on the way to the desired destination. Thus, we also introduced finite-horizon action selection instead of full-length-to-goal planning, and spoke about the question of how much sensory information is necessary and sufficient for action planning of differing degrees of optimality. Thus, we went all the way from all-past and all-future knowing omnipotent infinite-step ahead planning agents, to simple one-step ahead action selectors, which select on the basis of only the current sensory stream contents.

When discussing necessary and sufficient information for action selection, we also briefly touched upon the subjects of [Q2] and [Q3]: First, regarding [Q2] (internal representation), it should have become clear that:

> For an agent to select his actions in the best realistically possible way, some form of keeping internal track of previous contents of the sensory stream is necessary, for most non-trivial cases of external "environments" in which the agent is embedded and "goals" that he is pursuing.

Then, regarding [Q3] (prediction), as knowledge of the future sensory stream contents would have been in many cases valuable if it had been possible, and as it is realistically not possible given the nature of time, the best we can do is substitute the missing knowledge of the future sensory stream contents with our predictions of them. Of course, the predictions can only be generated on the basis of previous sensory input again - and thus, although predictions might be used as intermediate byproducts in order to aid action selection, in the end again our future actions will only be effectively based on past sensory inputs.

2.4.5 Internal representation: what and why?

At this stage, we can attempt an answer towards two important questions relevant to [Q2]: first, why should internal representation take place? Second, what are some qualitatively different levels of internal representation? I.e. is it enough to momentarily copy sensory information? Is there a need for processing/transforming the sensory information in some way? Is there a need for keeping "traces" of past sensory information of some sort?

When is keeping internal state necessary?

First, from the previous sections the following important result should have became clear: *keeping internal state is necessary whenever, in order for an agent to optimally select actions to be taken in the future, he also needs information extracted from the past of the sensory stream - i.e. if using the present contents of the sensory stream only is not sufficient for good action selection.* The "internal state" that is kept by the agent can at most contain the whole of sensory experience so far; of course, this might not be stored intact, but "distilled" in some form, that only contains and exposes the necessary information for good action selection. What will be kept as internal state and how it will be transformed / distilled is an immense subject, which is also extremely relevant to the purpose of this thesis - one of the main positions of this thesis is that a "situation model" should form a major part of the "internal state" to be kept by any SCA. Anyway, in the previous subsection we saw how the need for keeping some form of internal state arises, in an argument starting from the viewpoint of optimal action selection towards goal satisfaction.

Various other arguments towards arriving at the necessity of keeping internal state

One can also arrive at the same result (that very often keeping internal state is needed) through different pathways of argumentation. Some examples follow.

In these examples, we will propose three qualitatively different levels of internal representation that can arise. At the first level, we are talking about intact internal representation (*"copying"*) of the current contents of the sensory stream:

$InternalState[t] = SensoryStream[t]$

At the second level, we are talking about processing of the current sensory stream contents in order to produce *"transformed"* versions of the data:

$InternalState[t] = f(SensoryStream[t])$

Finally, at the third level, we are talking about not only possibly transforming, but also keeping the transformed data for the purpose of using them for future action selection, and possibly recombining this internally held state with new sensory data to create the new *internal state*:

$InternalState[t] = f(SensoryStream[t], InternalState[t-1])$

Thus, even before talking about keeping internal state that persists across time, a first question that can arise is: what is the purpose of internal "copies" or representations (transformed or not), even if they are momentary, and not kept for later use? I.e.

once we have some effect of an external stimulus on a sensory organ, why should it be "transferred" and effectively copied/transformed to another part of the organism?

First level: The need for representations containing *"copied"* information

Let us first start with a very simple case, where no "representation" exists. Take, for example, the question of the information about the energy left in a living cell. The ATP molecules themselves are the energy that is left; there is no neural circuit or other transmission method that effectively transfers information about the energy left. On the contrary, consider the "fuel left" indicator needle in a car. In that case, the position of the indicator needle corellates with the amount of fuel left in the tank, forming what is a rudimentary, but nevertheless useful, representation. The purpose of this representation ultimately grounds out to the purpose of agency: taking the right action. Indeed, the car operator, by monitoring the needle, will be able to act in order to replenish the fuel of the car whenever required. Thus, the purpose that the designer of the car (artifact) designed the car for (i.e. moving people and things are places), will be able to continue to be satisfied[23]. But why is there neither "wire" nor "indicator needle" to transfer information about ATP across distances within molecules? Because in that case the process of replenishing is initiated locally, and without the intervention of a human operator.

Second level: The need for representations containing *"transformed"* information

So we saw a case where there was no "re-representation", and again saw that driving purposeful actions is again the reason behind any representation. Now let us briefly touch a second question: why should the sensory information often be "transformed" when internally represented? Why is loss of information sometimes necessary? Consider the visual system of a rabbit. One important set of stimuli for the rabbit to be able to notice are the stimuli originating from tigers - immediate action is required when sighting a tiger [fleeing, subserving the "move away from danger" subgoal of the goal hierarchy of life]. Thus, to start with, the visual sensors of the rabbit should be able to notice tigers. Also, as the action to be taken involves the leg actuators (muscles), there must be some sort of information transfer across a distance from eyes to legs, and thus some representation of part of the visual information across a transfer circuit is required - i.e. there is a need for representations containing "copied" information. The question now becomes: should the visual information be transformed in any way? Should information be lost? Here, I need to clarify what I mean by *"loss"*. By *"loss"* I mean the concatenation of equivalence classes in the original information - i.e. the inability to discriminate in the output (transformed information) between two or more original states in the input.

Let us return to the tiger and rabbit example. Any two sightings of a tiger will most probably not produce exactly the same visual stimulus, due to pose and lighting differences, elastic deformations, aging of the tiger etc. This is even more the case across

[23]Obviously, on the other hand, if there was no replenishing, the non-availability of fuel would inhibit movement. Also, notice that replenishing energy is a classic subgoal of the goal hierarchy of life (i.e. enabling nutrients to come in).

tigers; any two tigers will not look exactly the same. However, we expect the "fleeing" action to be initiated by any of these different stimulus. But notice however, that during the evolutionary or developmental lifetime of the rabbit, not all of these "variants" of visual signatures of tigers were shown to it. Thus, even if successful associations between some of the visual signatures of tigers and the fleeing behavior were possibly made, these will not cover other variants of the visual signature of tigers. Thus, some form of *generalization* across tiger stimuli is necessary for appropriate actions to be taken. But essentially, generalizing across tiger stimuli means ignoring some of the differences across the stimuli as being unnecessary to be noticed - as we want to initiate the same action across potentially different stimuli. But this is exactly concatenation of equivalence classes - we collapse two tiger stimuli that would otherwise be noticed as being different to the same class. Thus, we need to "transform" the visual signature of the tiger to some other representation, in which two different tigers or poses would have the same "value" in the representation[24].

Third level: The need for ***keeping "internal state"***

Hence, so far we have seen cases where: there is no need for representation (ATP example), where there is a need for a representation containing "copies" of remote information (fuel indicator example), and a case where there is a need for a representation containing "transformed/lossy versions" of remote information (rabbit example where different tiger stimuli were collapsed into one). In the above cases where representation existed, the ultimate purpose was (as always) driving purposeful actions, that would serve subgoals of the goal hierarchy of life (or the designer's intended goal). Now we will ask the question: there seem to be cases where indeed "copies" or "transformed versions" of current stimuli need to be represented. But these are "immediate" copies; they need not persist over time. Is there sometimes a need for "persistent" traces of them, i.e. of keeping some "internal state" in an organism?

In both of the above cases (fuel indicator / tiger-rabbit), we saw the need for "immediate" correlation between stimulus and representational contents; apparently, no need for "memory" or keeping internal state existed. However, a closer inspection of the second shows that keeping "traces" of the past are necessary; the category of the set of stimuli that count as "tigers" should be represented somehow - storing of some form of exemplars, and/or prototypes, some sensory distance metric, and/or parameters of some parametric classifier of whatever sort. This representation of the category of tiger stimuli will have to be shaped by the evolutionary or the developmental experience of rabbits - and thus, essentially, these representations are some form of "internal state" caused and made up by transformed information arising out of sensory experience of tigers. This close chicken-egg relationship between category formation and noticability of stimuli is examined beautifully in [Hayek1952]. In essence, Hayek argues that in the idealized limiting case, originally the sensory effect of the first "tigers" to be seen would be for them not be noticed as tigers, but to initiate the formation of the category of tigers. Later, as the category forms and is further shaped, the new incoming "tiger"

[24] On the question of how to find the optimal categorical boundaries for intermediate representations on the basis of utility for action selection, see for example: [Horvitz and Klein1993].

stimuli are perceived as tigers. I.e. if the category of "tigers" is not frozen at some point of the lifetime of the rabbit, each incoming tiger stimulus will have two effects: one direct, i.e. being reported as the sighting of a tiger, and one indirect, i.e. helping form the category of tigers, and thus shaping also how future tigers will be perceived. In a "simulated annealing"-type "gradual freezing/stabilisation" of categories, we would expect the second effect to decrease with time.

Detachment of stimuli as a driving force for internal representation, and the possibility of self-modifying representations

One should notice that in the above discussion we touched upon two interesting and highly relevant ideas, which I will try to restate more succinctly here. Before that, let us briefly revisit and rebuild upon some of the above discussion. We saw that whenever there needs to be a transformation of incoming information in order to enable generalization through the formation of categories, incoming stimuli initiate two resulting effects: they are not only represented as members of a category (first), but might also shape the category itself (second), and thus affect how future stimuli will be perceived. Thus, the stimuli do leave some persistent "traces", which indeed could count as "internal state" for the agent's model[25]. But also, this means that even when a stimulus is "absent", it still exists (similar stimuli have left traces) internally in the agent through the representation of its category, which is essentially a "trace" of the absent stimulus. Also, this internal representation of the category, is also what enables later recognition of the stimulus as belonging to the category, and thus also forms an "expectation" of the re-appearance of a stimulus belonging to the category sometime in the future. Thus, the two interesting ideas that have arisen are the following:

First, that during the lifetime of a complicated organism, its internal representations might not be "static" in terms of the way through which they correspond to external stimuli, but might well be "dynamic" and evolving; a simple example comes through the continuous updating of categorical boundaries through experience; what I call a "cat" might not have included a himalayan cat ten years ago, but now it might[26]. The situation-model structures proposed here, are quasi-dynamic: categorical boundaries might change during training of the robot, however the overall structure of the situation model remains the same[27]. The second idea that has arisen is similar in vein to

[25] Of course, one could also posit equivalent structural deformations of the internal circuitry of the agent; but if one chooses to parametrize the agent's model, these structural deformations amount to changes of internal state.

[26] Similar phenomena of adjustment of categorical boundaries arise in linguistic development of children - such as "overextension" and "underextension", usually along the initial stages of meaning trajectories of words. An example of overextension: the word "hat" might mean just about anything that one can wear on his head. An example of underextension: the word "cat" might mean only the specific cat that the family owns, and no other cat.

[27] Ideally, one would like to have self-evolving and self-adapting representations, but such ideas are not pursued extensively here. In such cases, one might expect two stages of development: first, slow and continuous augmentations / tunings of the representations, similar to theory augmentations/extensions during the development of scientific theories of physics, and second, radical re-arrangements of the representational apparatus, similar to scientific revolutions or "paradigm shifts" in Kuhn's terminology [Kuhn1962].

BC Smith's "sun-tracking flower" example [Smith1996]: it is the periodic absence or inaccessability of the external stimulus that causes the need for an internal "copy" of it. The fact that all tigers are not continuously accessible to the eye, but disappear and will re-appear is what drives the need for an internal "trace" of a tiger within the rabbit's mind[28]

Recap

Thus, in this subsection, we have seen several other arguments regarding the need of internal representation, apart from the argument given in the previous subsection. We have qualitatively distinguished three levels of complexity of internal representations: those just consisting of "copied" instantantaneous sensory information, those consisting of "transformed" instantaneous information, and those of consisting of "internal state" that has arisen through some function of current and past sensory information. Then, we considered yet another viewpoint towards the necessity of internal representations: the argument of detachment of stimuli as a driving force for internalizing them. Finally, we also briefly hinted towards the possibility of flexible self-modifying representations.

Now, having occupied ourselves with [Q1] (action selection) and [Q2] (internal representations) in this and the previous subsections, we will move on and discuss the utility of prediction [Q3].

2.4.6 Prediction

Let us now consider [Q3]. Apart from the need for keeping some form of internal state, we also briefly touched upon the need for prediction. When discussing necessary and sufficient information for action selection, it became clear that pre-knowledge of anticipated future sensory inputs would have been ideal. Of course, any non-prophesizing agent cannot directly have access to this pre-knowledge; thus, prediction (on the basis of past/present sensory information) is the best that we can do. Of course, the predictions can only be generated on the basis of previous sensory input again - and thus, although predictions might be used as intermediate byproducts in order to aid action selection, effectively again our future actions will only be based on past sensory inputs. But even as intermediate byproducts, they can be very valuable.

So, the question follows: how is prediction normally implicated in the action selection process? For example, one possibility is the following: Envision an agent that has some process generating a preference ordering for future sensory inputs, that at each moment can evaluate how "good" a forthcoming sensory input might appear to be for the agent[29]. Thus, if the agent is equipped with predictive machinery, that can produce

[28]For BC Smith, this condition of intermittent accessibility is possible due to what he calls the "flex and slop" of the universe, i.e. the state of intermediate coupling among entities (neither total inaccessibility nor complete accessibility), that is supported for example by the spatio-temporal inverse-distance attenuation laws of physics.

[29]For example, this utility function might be current goal-relative, as discussed in the previous sections on spreading the reward over time, and as will be discussed in the next subsection too.

predictions of the expected next sensory input of the agent as a function of the internal state and an action to be taken, i.e.:

$Prediction(S[t+1]) = f(S[t0 \dots t], A[t])$[30]

Then the agent can cycle through all possible actions to be taken, try to predict the sensory inputs that they will produce, and choose the action that will produce the most preferred predicted sensory outcome.

This is a classic route through which prediction can be practically utilized and motivated. And, as mentioned before, the predicted future sensory states, are nothing but the intermediate byproducts of past sensory states that are used as the means for the ultimate end of agents: purposeful action selection[31]

2.4.7 A hierarchy of action selection methods

Now, let us finally revisit Q1. What are the possibilities for specific methods of action selection in an agent? Here, we will try to present a hierarchy of increasingly complex methods. This section does not claim that the presented methods are novel - they have all existed in the literature for quite a while, and also in much more elaborate and advanced forms - a glance through any AI or reinforcement learning textbook (such as [Russell and Norvig2003], [Sutton and Barto1998]) will indeed confirm this statement. The aim of this section, is to introduce a hierarchy of five qualitatively different levels of action selection, which will also be useful later in this thesis.

Level 0: Fixed stimulus-response agents

At the lowest end (which we call L0 - Level 0), one might find very simple stimulus-response tables with fixed entries: the agent's actions are a constant function of only the current snapshot of the sensory stream. For example:

($Agent0$): Imagine an oversimplified agent that has only two possible actions at its disposal: move towards a stimulus (A1), and move away from it (A2), plus the trivial action of remaining stationary (not moving) (A3). Imagine also that it can only discriminate among three possible momentary sensory states: smell of food (S1), smell of predator (S2), nothing (S3). The stimulus-response table for such an agent might be:

S1 - A1 [IF $sense = smell\_food$ THEN $action = move\_towards\_smell$]
S2 - A2 [IF $sense = smell\_predator$ THEN $action = move\_away\_from\_smell$]
S3 - A3 [IF $sense = smell\_nothing$ THEN $action = do\_nothing$]

But how has this table arisen? In the case of an artifact, it was pre-wired by the designer. In the case of a natural organism, evolution has effectively selected for the table entries[32].

[30] Possibly not only single-step ahead, but also multi-step ahead i.e. $Prediction(S[t+k])$

[31] Notice that this is also again the purpose of another species of "intermediate byproducts": inferred logical sentences. When an agent uses first-order-logical representation internally in order to represent knowledge, then any further sentences that are inferred are ultimately just means towards the end of purposeful action selection. Inferences that have no chance of being used have no reason to be made.

[32] From a systems viewpoint, this agent acts as a memory-less system: $A[t] = f(S[t])$. This is the case of

Level 1: Internal-state-keeping agents

After visiting level 0, we now will move on to L1 - Level 1, namely what we will call internal-state-keeping agents. In this case, let us assume that we have an L0 agent, in which some internal state is also kept, which internal state is of finite length. For simplicity, let us revisit $Agent0$, the example agent used when discussing level 0, and let us augment $Agent0$ with internal state, in order to build $Agent1$. Let us assume that in $Agent1$ we have one internal variable (I1) that can have two possible states: hungry (I1), and not-hungry (I2). Now, we will have two tables: a (stimulus, state)-to-action table, and a (stimulus, state)-to-state table[33]. Let us also assume the existence of a new stimulus, namely S4 = $food\_in\_mouth$, and a new action, namely A4 = eat_food.

In this new example ($Agent1$), the (stimulus, state)-to-action table might include (among others) the following entries:

(S1,I1) - A1 [IF $sense = smell\_food$ AND $state = hungry$
THEN $action = move\_towards\_smell$]
(S1,I2) - A3 [IF $sense = smell\_food$ AND $state = not\_hungry$
THEN $action = do\_nothing$]
(S4,I1) - A4 [IF $sense = food\_in\_mouth$ AND $state = hungry$
THEN $action = eat\_food$]
(S4,I2) - A4 [IF $sense = food\_in\_mouth$ AND $state = not\_hungry$
THEN $action = do\_nothing$] etc.

While the (stimulus,state)-to-state table might include:

(S4,I1) - I2 [IF $sense = food\_in\_mouth$ AND $state = hungry$
THEN $state = not\_hungry$]
For every x:
(Sx,I2) - I1 [IF $sense = whatever$ AND $state = not\_hungry$
THEN $state = hungry$] etc.

L1-L0 differences

What is qualitatively different in such an agent ($Agent1$) that keeps internal state as compared to the simple stimulus-response agent ($Agent0$)? First of all, notice that there are effectively two "moods" in this agent: there is $Agent1a$ (hungry) and $Agent1b$ (non-hungry). These two moods switch: whenever food is digested, we move from $Agent1a$ to $Agent1b$, and after one time slot, we move back to $Agent1a$. In general, every state keeping agent, is effectively "switching" among moods, which during the period within which they remain stable, are each equivalent to a single L0 agent[34],[35].

---

using only "transformed" information that was examined in the previous subsection.

[33] In dynamical systems notation $A[t] = f(State[t], S[t])$ and $State[t+1] = g(State[t], S[t])$. This is the case of agent that "keeps internal state" in the three-level terminology of the previous subsection.

[34] "Mr.HungryNick and Mr.Non-HungryNick might be effectively almost two different people!

[35] In system-theoretic terms, this translates to the following: a discrete-state system at a specific moment, and when viewed only for that moment, is equivalent to a specific memory-less system. Across

Level 2: Conditionable stimulus-response agents

Now let us move on to L2 - Level 2, the level of what we will call conditionable agents. The main difference here, is that there exists at least one special stimulis: either a reward (S+) stimulus, or a punishment (S-) stimulus, or both. These stimuli can be used to effectively reprogram the agent, in order to alter the entries of its stimulus-response table during his lifetime. For example, consider a special type of level 2 agent, that comes prewired with a table such as: (for some actions and stimuli)

S1 - A1
S2 - A2

Now assume that S1 occurs, and the agent acts A1. Then, at a short temporal distance, S- is applied to the agent. In that case, the agent effectively "decreases" the strength of the connection between S1 and A1, and the next time S1 will occur, it might act stochastically out of its repertoire of possible actions. Let us assume that S1 re-occurs later, and that the agent acts A2. If after a short temporal distance S+ arrives, then the agent will "increase" the strength of the connection between S1 and A2, and might after a while only act A2 after seeing S1. Thus, the new table after the "reprogramming" of the agent through the right sequence of stimuli which included reward and punishment stimuli now is:

S1 - A2
S2 - A2

L2-L1 differences

Thus, we can now ask: what is qualitatively different between an L1 agent and an L2 agent? An answer is that, before and after special "training" sessions which contain the special reward and/or punsihment stimuli, an L2 agent is effectively an L1 agent. But during the "training" sessions, the agent effectively changes "personality", and learns how to respond differently to stimuli[36]. After the new personality has "stabilized" through repetitive and consistent reward / punishment sessions, we have a period of "fixed personality", which is effectively equivalent to an L1 agent. This lasts until and if the reward / punishment pairings with his actions change again. Notice that during the "training" phase, which in our example is entered whenever a current stimulus-action pairing is punished, the agent might also deterministically cycle through the possible actions to be paired with the stimulus, instead of stochastically acting for the stimulus-action pair which has recently been punished and under goes reprogramming. Notice that previously, a similar situation arose, while discussing L1-L0 differences. Again, there were effectively steady "moods" inbetween internal-state change periods. In general, one can rewrite any non-stochastic or near-deterministic L2 agent as an L1 agent, with a special internal-state variable that switches between "training phase" and "steady phase", and

moments, we effectively have various different memory-less systems being switched around.

[36] For the interested reader, an advanced proposal for a model explicating the role of affect in the traditionally purely cognitive areas of learning and decision making can be found in [Ahn and Picard2006].

with parts of the (stimulus, state)-to-action and (stimulus, state)-to-state tables fixed and unalterable. Thus, one also think of L2 agents as a special case of L1 agents that contain "re-programming" code in their genetically-derived initial tables[37].

Level 3: Generally fixed short-term planning

Now let us make a big leap, and move on from what is normally termed "reflexive" agents to "planning" agents. Here, instead of having (stimulus, state)-to-action and (stimulus, state)-to-state tables, we will assume that there exists an action list (possibly parametric actions), as well a utility-based or goal-oriented action selection procedure, which at every time step chooses a single action or a number of future actions towards the maximization of expected utility or the achievement of a goal[38].

The point to notice here is that at this level, the planners are supposed not to be flexible: they specify a plan completely and not partially, and once they have specified a plan, they will try to carry it out irrespective of what happens after plan execution has been initiated. Regarding possible plan failures, in such agents there might be at most a simple "plan failure" condition detector, that terminates the execution of the current plan, and ab-initio and ex-nihilo initiates replanning.

Level 4: Flexible longer-term partial planning agents

Here we assume that the agent possesses what Bratman [Bratman1987] calls a planner that can handle "partial hierarchical plans". The partiality refers to the fact that our future plans do not need to have all of their details filled in at a certain moment; in Bratman's words from an everyday example that he gives: [Bratman1987]:

> "First, my plans are typically *partial*. When I decide today to go to Monterey tomorrow, I do not settle all at once on a complete plan for tomorrow. Rather, I decide to go to Monterey, and I leave till later deliberation about how to get there in ways consistent with my other plans. Second, my plans typically have a *hierarchical structure*. Plans concerning ends embed plans concerning means and preliminary steps; and more general intentions embed more specific ones. As a result, I may deliberate about parts of my plan while holding other parts fixed. I may hold fixed certain intended ends, while deliberating about means or preliminary steps"

Such a flexible planner, can easily gradually refine the details of his plans during execution time, and negotiate changes or replan.

L4-L3 differences, "social planners"

As compared to Level 3 agents (fixed planners), flexible planners are able to plan partially (leaving numerous options and details open, to be specified at a later time), and

[37]Also notice, that similarly to the existence of a "universal" turing machine, one could have an agent whose internal state effectively keeps his "program".

[38]Such as a rational maximum expected utility decision mechanism, or a STRIPS-type planner etc.

are also able to replan often and partially and recover gracefully from plan failures, without having to re-initiate replanning ex-nihilo. Such planners might also be able to coordinate plans in real-time in conjunction with others, effectively becoming "social planners" (almost a new level on its own). Consider, for example, the kitchen cooking assistant task discussed in chapter 1. Say that my partner and I started to make greek salad; but, while cutting the tomatoes, I found out that some of them were rotten. Thus, I was able to change my original plan and use more cucumber instead; and thus, I informed my partner, and the overall sequence of actions towards our goal was dynamically coordinated and replanned.

Recap

Thus, in this section, we described a hierarchy of four levels of agents of increasing complexity in their action selection methods. We started from simple stimulus-response agents, moved on to simple state-keeping agents with fixed action tables, then introduced conditionable agents, and finally considered inflexible fixed planners and flexible partial planners. Notice that the lower-level agents have a limited temporal horizon; they live in the here-and-now, maybe keeping some "distilled" internal state. Also, they are quite inflexible in adapting their actions to changing environments. However, the "planning" agents at the higher end, usually have much wider temporal horizons and even explicit situation models and episodic memories, and also can flexibly plan, replan, and adapt. This overall five-level hierarchy is also followed in our detailed description of the proposed GSM-based agent model in appendix B.

Now, having dealt with all three questions posited before, namely [Q1] (action selection), [Q2] (internal representation] and [Q3] (prediction), let us move on and conclude this chapter by presenting a sampling of existing mind models, and discuss their similarities and differences, as well as virtues and intended areas of application.

2.5 Some Existing Mind Models

There exist quite a variety of models of minds in the literature[39]. In total agreement with Whitman Richard's important comment in [Richards1990]: although when it comes to the physical worldview, we have converged to a handful of models (Newtonian, quantum-dynamic, relativistic), with pretty stable conceptual entities comprising them and with the possibility of moving across them by imposing limiting assumptions, the situation is vastly different when it comes to models of minds. Here, many more models exist, most of which are partial and focus only on specific aspects of minds, and which very often contain conceptual entities that are unalignable and untranslatable across them.

Here, I will very briefly go through a selection of some such models. Notice that they differ immensely in purpose and form: some serve theoretical modeling purposes,

[39] Often referred to as "Cognitive Architectures" too. For a review of some classic architectures as well as some interesting comments regarding aspects that have not gotten the deserved attention yet, see [Langley and Laird2002].

others are mainly suited towards computational modeling of cognitive-science experiments, others have been proposed for or used in building real-world AI systems. Also, notice that due to the differences in focus, there are quite differing levels of detail in the descriptions of minds in these models, and that very rarely if ever have equivalences / translations among the models been attempted.

Simplistic ad-hoc models

At the simpler end of the spectrum, one can postulate simpilistic *stimulus-response table* machines (at level-0 in terms of the action selection methods hierarchy described before - minds such as those used in simple Alife simulations), *finite-state automata*, or simple *propositional-logic type descriptions of minds* consisting of unstructured lists of beliefs and desires of the form:

believe(Agt, Prop), want(Agt, Prop)

These models might also be coupled with some form of simple planner (such as those postulated by theoretical papers targetting some specific aspect of cognitive function, for example [Cohen and Perrault1979]). Such models often lack in flexibility, and cannot cope with the problems arising from real-world continuous-valued sensory data, or with the requirements for real-world natural-language interactions and theory of mind, and most frequently do not explictly organize around a notion of a situation-type or an episodic memory. However, for the purpose of illustrating restricted domain-specific aspects of minds, they have often proven adequate.

SOAR and ACT

SOAR [Laird et al.1987] is a well-established cognitive architecture that has been in development for more than two decades. It is essentially a large *production system*, effectively consisting of symbolic rules of the form:

IF (condition) THEN (action)

Where the actions taken might either effect the internal state of the system or be primitive external actions. The production rules are in turn organized in terms of operators connected to specific types of problems. Tasks in SOAR are in essence attempts to achieve goals; and operators are applied and grouped towards that effect.

Yet another production rule system family is ACT, with ACT-R (Anderson and Lebiere 98) being a more recent standard member of the family. ACT-R is tailored towards modelling human behavior, and contains more specialized architectural subsystems than SOAR, including two long-term memories (declarative and procedural), and provisions for keeping quantifications of utility, benefit and cost when choosing among different productions towards a goal.

The above systems have been widely used in the past. Regarding shortcomings, one can mention some of those that arise from the strongly discrete, logical sentence-like nature of production systems: no uniform and obvious way to connect to continuous-valued and noisy sensory data, to support categorical classification, to include affective variables[40].

[40]Affect, largely ignored during the early years of AI, has proved to be increasingly important both in

Anigrafs

Yet another modelling direction is taken in Whitman Richard's Anigraph models [Richards2005]. These are multi-agent models of minds, that place primary emphasis on action selection through a voting process (Condorcet Tally) among the constituent agents, and on the topology that arises from the relative positioning of the agents within this voting process. Here, the main modelling goal is to abstract away from the details of sensing and internal representation, and to create an isomorphism between agents and graphs, that captures the essence of the decision making mechanisms of the agents. By moving from the complexity of agents to the concreteness of graphs, one can then find precise conditions for specific aspects of the agents by postulating equivalent statements in the well-defined world of graphs. For example, a "theory of mind"-type embedded model of another agent's mind (agentB) within an agent (agentA), is equivalent to the existence of a subgraph within the graph of agentA corresponding to the graph of agentB.

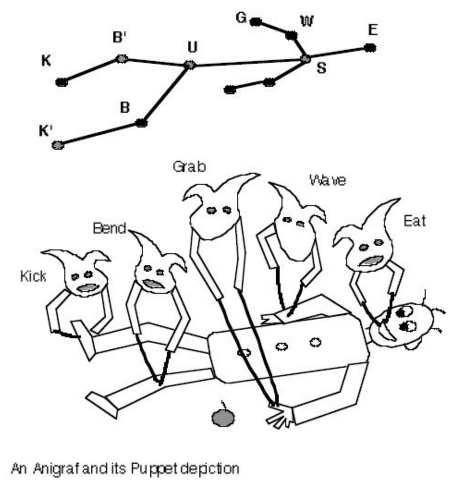


Figure 2-5: Anigraf as a society of agents: in this example, each agent controls one physical actuation point, and patterns of movement are created through sequencing

GSM-based Agent Models

In the next chapter (chapter 3), the proposed GSM-based agent model will be introduced in detail. In comparison to the above examples of mind models, some of its advantages include the explicit attention given to connecting with continuous-valued and possibly noisy sensory data as well as with natural language utterances, as well as its clear consideration of the situation datatype and episodic memory. Regarding its shortcomings, it places less emphasis on defining special structures for planning, goals and affect (which however can still be implemented within the system, as we shall see in chapter 9), and does not explicitly model logical inference or cognitive constraints yet. Furthermore,

modelling human behavior as well as in AI. See for example [Picard1997], [Ahn and Picard2006] and [Minsky2006].

there is no presciption of how habitual (conditionable situation-response) and more deliberative (goal-based) action selection systems can co-exist[41]. Again notice that action selection was not targeted during the design of the model, but anyway extensions can be easily foreseen, also in the form of simple augmentations.

Given the stated advantages of the GSM-based agent model and despite the above shortcomings, it has proved to be quite effective for the purpose of creating Situated Conversational Assistants, as well as for theoretical reflection, as we shall see in more detail in chapters 6 and 7.

Section recap

Thus, in this section, we have discussed the plurality, intranslatability, and application specificity of existing mind models (in contrast to the more unified picture of established world models in physics), and we have provided a sampling of some existing mind models: simplistic, SOAR/ACT, Anigrafs and GSM-based agent models.

2.6 Recap

In this chapter, we took a big-picture viewpoint, and positioned our discussion within a version of the "Semiotic Triangle". There, we discussed models of signs, minds and worlds. We started by introducing the semiotic triangle, and by re-examining the previously introduced notions of Symbol Grounding, Situated Language and Situation Models, within this new framework (section 2.1). Later, we entered a detailed discussion: initially talking about models of the world (section 2.2), then introducing teleology and life (section 2.3) in such models, and then, moving from the "physical" to the "mental", we will talk about "models of minds" (section 2.4). Within our discussion of mind models, we also specifically talked about internal representation, prediction, and action selection, and this lengthy section was closed with a quick presentation of some existing models of minds - coming from a variety of directions and addressing differing needs. Overall, this chapter hopefully provided a proper prelude for the introduction of the overall agent architecture framework within which the proposed GSM-based agent model resides, in chapter 3 which follows directly.

[41]There exist even more elaborate and interesting proposals that separate processing from perception to action into more strands, such as Sloman's CogAff Architecture **??**. In H-CogAff, the three strands that are proposed are: Reactive, Deliberative, and Reflective.

Chapter 3

The proposed GSM-based Agent Model

Here, having discussed some basic notions about descriptions of reality, models of minds, action selection, internal representation and prediction in the previous chapter, I will introduce the overall agent architecture framework within which the proposed GSM-based Agent Model will be positioned[1].

As a connecting bridge, let us remember some of the reasoning justifying the arisal of internal representations (including "situation models"), given in the previous chapter. In section 2.4.5, I assumed that within the agent model, internal representations of external reality might arise, and that they often do arise for two good reasons. First, reality is not always accessible to the senses, but predictions about future encounters with it must be made (for example, by keeping track of objects encountered in the past, and expecting to see them again when their location is revisited). Second, in order to be able to generalize about effective reactions to future stimuli given past stimuli, one needs to categorize stimuli, i.e. "throw away" unnecessary differences and keep their "essence" (for example, having seen a view of a tiger and having had bad experiences with this encounter should cause you to take precautions when you see a different view of the same tiger, or even a different tiger). But one cannot do the above unless he possesses an internal representation of the stimuli to process, so again this need justifies the existence of internal representations. But having tied the purpose of internal representation to prediction and generalization, one might ask again what purpose is served by successful predictions and generalization. Once again, the purpose grounds out to the more-or-less shared goal hierarchy of beings (or the designer's goals for artifacts): Predictions and generalizations ultimately enable meaningful action selection towards the purposes of the goal hierarchy of the agent, and this is why they are ultimately useful.

Now let us return to this chapter. I will start by a brief presentation of the overall agent model and its constituent representations and processes in section 3.1 (more details can be found in appendix B, and then proceed to the very important question of *alignment* across two different situation models, which is discussed in 3.2. This question in layman's terms amounts to: how can we reconcile multiple different viewpoints of the same situations? How can two beings that might look at the world through different sensors, use different ways to "break down" the world, and have different vocabular-

[1]Essentially, this is a proposal for a GSM-based mind model / GSM-based cognitive architecture for artificial agents.

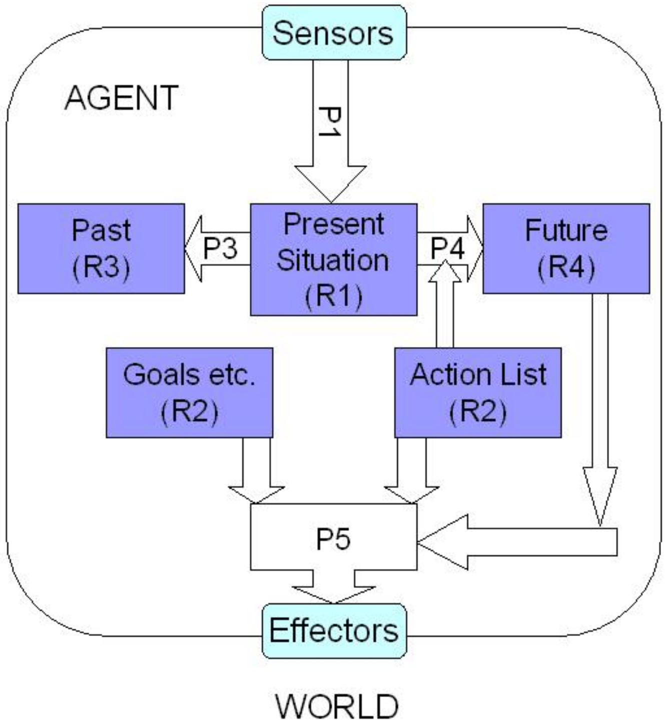


Figure 3-1: Block diagram of the proposed agent model

ies / perceptual categories, be able to negotiate meaning among them, and effectively communicate?

3.1 The proposed GSM-based agent model framework

Agent models, as introduced in the previous section, are representational / algorithmic descriptions of the "minds" of beings, which are assumed to be connected through sensing and acting to reality. According to our previous arguments, internal representations of external reality do arise inside agent models. These representations of external reality we will term "situation models". These representations also come along with processes operating on them. Inside agent models, apart from situation models, further representations might encode a description of the internal state of the agent (goals, affect). Furthermore, two other types of representations arise in agent models. First, representations storing and organizing past states and situations, and second, representations holding predictions for the future / imagined situations. A complete enumeration of the representations and processes that comprise the proposed agent model follows below:

Representations:
R1) Present Situation model (not necessarily exclusively physical model)
R2) Other internal state (current goals, affective variables)
R3) Past states/situations (includes events)
R4) Future predictions / planned or imagined situations

R5) Action list

Processes:
P1) Sensor-to-SituationModel process
P2) SituationModel-to-InternalState proc.
P3) Storage/Maintenance of Past process (includes event recognition)
P4) Future Prediction Generation process
P5) Action Selection Process (SituationModel-to-action table or planner)

Required Models external to the agent:
M1) ExternalReality-to-Sensor model
M2) Action-to-ExternalReality model

In order not to interrupt the flow of this text, a closer look at the above components is given in appendix **??**.

3.2 Aligning Situation Models

Here, the questions that we will try to investigate, are the following:

- Given two situation-model-based agents of the form described in the previous paragraphs, which are both embedded in the same situation, in what ways can their situation models differ? (Q3.2a)
- Up to what extent can one translate from the "worldview" of the first agent to the "worldview" of the second? (Q3.2b)

A related very important theoretical but also practical question is also the following:

- How can two beings that might look at the world through different sensors, use different ways to "break down" the world, and have different vocabularies / perceptual categories, be able to negotiate meaning among them, and effectively communicate? (Q3.2c)

Apart from the theoretical interest, these three questions are also quite important in a practical sense: if we want a machine to effectively communicate with a human, there should be some extent of similarity / overlap among the situational "worldview" of the machine and that of the human. It is not enough for the machine to share a common vocabulary with the human; in order for communication to be possible, the linguistic symbols should be grounded in a similar way, and thus the situation models of the two agents should be related[2]. Thus, language learning for robots is definitely not only word learning; the robots should derive a situation model similar to the human, feed it

[2]As a simple example: what I call "green" should have sufficient overlap for what you call "green", what I call "the red ball that is in front me" should exist as a token in your situation model and should have sensory expectations that anchor to the same object in external reality etc.

through external reality in a similar way, and use similar linguistic labels - in our terms, the robot first needs to somewhat "align" its situation model with the human. This is why the above questions are so important. We will now try to briefly address these questions in this section.

3.2.1 The "situation-viewpoint" of an agent

The processes and representations of figure 3-1, can be further broken down (for the definitions of the representations R1-5 and the processes P1-5 refer to Appendix B). The situation model of the "present situation" rep (R1), is fed through process P1, the so-called "Sensor-to-SituationModel" process. This process, can be further broken down to three qualitatively different parts, which when viewed as a cascade and augmented with the ExternalReality-to-Sensor model (M1), specify the relationship of the present situation model (subjective internal reality) to the external situation (objective external reality), i.e. in layman's terms, the relation between the "world inside" and the "world outside".

Parts of the Reality-to-Situation Model pipeline

The three relevant parts of P1 augmented with M1 are the following (see figure 3-2):[3]

- *Reality-to-SensoryStream (M1):*
  Spatiotemporally local external reality is projected into a "sensory stream"
- *Sensorystream-to-IndividuatedStream (P1a):*
  Sensory streams are segmented to "objects" - i.e. some information is marked as pertaining to an object, which object is also given a unique ID
- *IndividStream-to-ObjectPropertyList (P1b):*
  All of the segments of the sensory streams that resolve to a common object are collected, and object properties are calculated on their basis - color, size etc.
- *ObjectPropList-to-ObjPropCategList (P1c):*
  Continuous-valued or fine-granularity property values are quantized to meaningful categories - for example, color is quantized to "red", "green", "yellow" etc., size to "small", "large" etc.

The situation-viewpoint of an agent

Thus, in the proposed framework,in order for somebody to completely specify the transformation that the "objective external world" undergoes in order to become the "subjective internal world" in a specific agent, he needs to know the following information: The models of M1, P1a, P1b, P1c, in whatever form they might be (for example: parametric)

This information (the above quartet of models) I will from now on call the "situation-viewpoint" of an agent. Notice that the above models contain parameters that might be time-varying: for example, for a physically mobile agent, we expect that the position of the sensors of the agent will change with time.

[3]As introduced before in section 3.1, and in detail in appendix **??**.

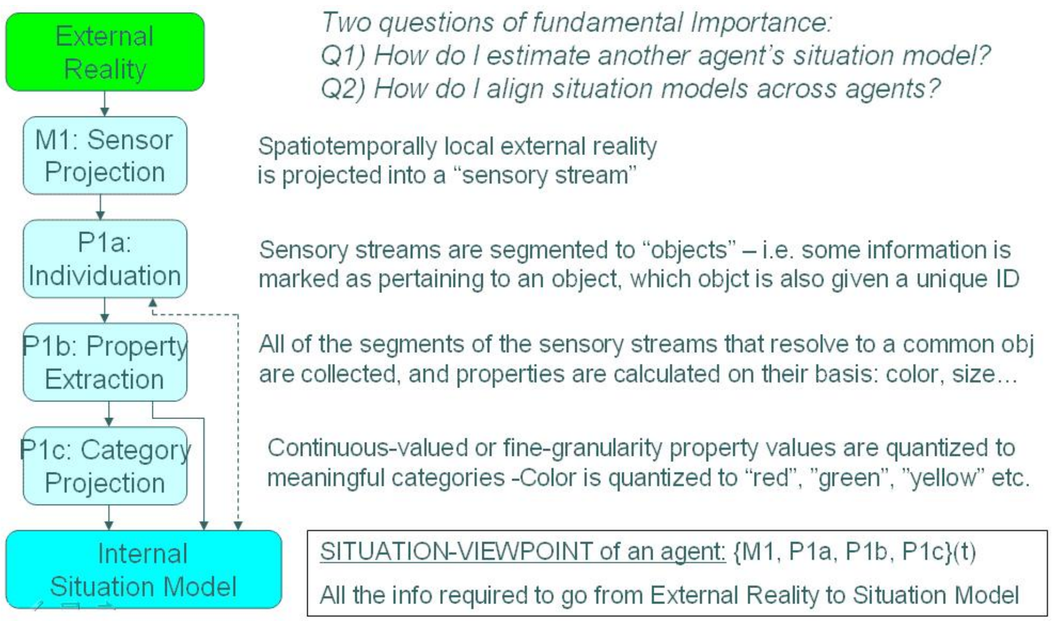


Figure 3-2: The four stages of the Reality-to-SituationModel pipeline

3.2.2 Species of differences across situation-viewpoints

Given the above four parts of the "situation-viewpoint" of an agent, we can now consider and enumerate some qualitatively different types of misalignment that might exist across two agents, i.e. some species of ways that their "worldviews" on external reality might differ (towards the question Q3.2a). Imagine that we start with two identical agents, both situated (unrealistically) the same position in spacetime and thus being fed with the same data, and also having exactly similar processing chains, i.e. having the same "situation-viewpoints". Then, both the contents and structures of their situation models will be the same. Let us now start to introduce difference, and see some ways through which we can make them different (towards an answer to question Q3.2a):

Some species of differences across situation-viewpoints

D1) Position / Pose misalignment (stage M1):

We can keep the same sensory organs, but change their position/pose. (change of parameters of reality-to-sensory stream model). Depending on the radius of coverage of the organs, and also the pose- and scene-dependent occlusions that might arise, their might or might not be partial or total overlap between the contents of the GSM's of the two organisms (supposing they were previously empty). Also, any observer-position / pose dependent contents might be different: for example, relative positional coordinates, apparent luminosities etc.

D2) Sensor modality / tuning misalignment (stage M1):

We can change the modality and / or the spectral tuning of the sensory organs (change of reality-to-sensory stream projection functions). Then, if there existed totally disconnected (uncorellated) "parts" of reality, we might not have had any common information among the two sensory streams. But such total disconnection does not exist:[4] Say that I see in infrared and you see in the visible spectrum: our sensory streams are usually correlated. Say that you only see and I only hear: again, our sensory streams are somewhat correlated. Similarily for the case of dolphin (sonar) and human[5]. Of course, whether there exists adequate correlation for any practical purpose depends.

D3) Objectification misalignment (stage P1.1):

Differences in objectification / individuation (change of mapping from sensory streams to object-specific sensory streams). When first listening to a new language, I cannot separate phonetic units the way that a fluent speaker can; thus, my units might misalign with his (audio segmentation). Or else, what you might consider as being two separate visual entities (for example, a door and its knob) I might consider as being one (this time there is a form of partial alignment, as the union of what you consider as *two* objects corresponds exactly to the *one* object I see, but anyway there is different granularity - what I construe as a single entity, is divided in two entities for you). Also, of course, even if momentarily we might have the same contents as the output of the second stage; but our processes might generally have different results - more on general vs. momentary alignment follows soon, in section 3.2.3.

D4) Property dimension misalignment (stage P1.2):

We might have a different number of properties, and different mappings between the sensory data belonging to an object and its properties (change of mapping from object-specific sensory stream to property vector). Then, the information that I regard as "shape" might include part of what you regard as "size" and "shape" etc. Also, we might end up having a different number of property dimensions: for example I might have all of the dimensions you have, also derived through equivalent processes, but might have an extra one, say color, which you lack.

D5) Property category misalignment (stage P1.3):

We might have the same property values for an object (i.e. see the same color), but classify it to different categories ("pink" for you while it is "red" for me) (i.e different mapping from continuous property values to property categories).

Some ways of alignment - towards Q3.3c

Having seen some species of differences across situation models, we can now try to hint at possible ways of alignment (towards the question Q3.3c). However, before doing that, I will try to answer an obvious objection to the above list: one might ask - what about affect and goals? Differences in affective state or different goals can make agents perceive

[4]Look at the argument given in the "reality" section of [Mavridis2005b].

[5]The possibility of human-dolphin intention recognition is examined in a similar framework in [Mavridis2005a].

the world quite differently - giving attention to different parts, or furthermore even making some parts totally unnoticed. Indeed this is the case - but as a first approximation, we will here make two simplifying assumptions. First, that the situation model does not have a distribution of attentional salience among its parts. Second, that perception is a bottom-up continuous pipeline, feeding the situation model through the senses - and that no top-down goal-driven active perception mechanism[6]. Extension towards these two directions (attentional salience within the GSM and active perception) is discussed in chapter 9. Now, after having clarified what our first approximation covers, let us, we can now try to hint at possible ways of alignment across different situation models, under the simplifying assumption that only one of the above species exists at a time:

D1) Here, essentially we have to negotiate a common spatial coordinate frame across agents - origin and unit measures, as well as synchronise their clocks. Similar objects/events in the situation model can be identified, and serve as an "anchor", if there is some spatiotemporal overlap throughout the experience of the agents.

D2) This is a difficult case. Here, we have to take into account the extent and the form of that correlation that exists across the sensory modalities / tunings. This problem is similar in essence to the problem of aligning multimodal traces within a single agent - for example, aligning tactile with visual information.

D3) Similar to "aligning segmentations" in the simple vision-only case, or to learning phonemic segmentation, or speaker attribution, for example.

D4) We have to align two multi-dimensional sensory spaces (each property consists of a number of these dimensions), that are being produced from the same object-specific sensory stream. Each dimension is a function of the object-specific sensory stream. We can try to parametrise the functions in some form, and try to "tune" the parameters in order to match the output of the other agent - for example, learn how to derive the property value "size" from the set of pixels of a region.

D5) At first sight, the simplest of the above: specific examples can be presented (different colors) to the two agents who are attempting alignment. Then, one agent describes them, and the other uses the (color, label) pairs as training points for categorical learning, in order to align his categorical boundaries with those of the describing agent.

Notice that in the above hints of possible ways of alignment, we have made the very often unrealistic assumption that we can observe the intermediate stage results of the pipeline across the two agents - i.e., that if one of the agents is a human, I can directly for example measure the absolute value of the "size" property of a seen object that is encoded neurally somewhere in his brain. Thus, as the above is obviously not the case, for the human-machine alignment case, special psychophysical / behavioral experiments or existing data have to be used, that rely on verbal reports or other externally observable corellates. We have also only dealt with the special simplified case where all four pipeline stages except one are aligned - usually, we would expect misalignment across more stages, which would further complicate the alignment process.

[6]However, in the GSM proposal, apart from the bottom-up pipeline, there is also a top-down expectation-driven path (not goal-driven), which is explicated in 5.2.2 (matching process).

3.2.3 Levels of agreement between two GSMs

General compatibility versus compatibility in given situations

Now, we are returning to revisit the question of general vs. restricted alignment that arose a while ago: If we are to say that two "worldviews" are aligned, up to what level of generality are they? How many levels of agreement can we have among two such systems? In order to deal with this question, we now need to introduce the notions of "informational containment" and "informational equivalence". A layman's definition of informational equivalence is simple to state: two descriptions are informationally equivalent if knowledge of the contents of one allows me full knowledge of the contents of the other and vice-versa. Let us know try to introduce "informational containment".

The partial order of descriptions of reality

As we have discussed in section 2.2, many possible representations of reality exist. If we try to order the possible representations of reality according to informational content, a partial ordering will arise. Thus, three different cases can arise when comparing the informational content of any two given representations of reality, where the comparison is concerned with whether the first contains all the information of the second or vice-versa.

Case 1: Equivalence
Both representations include equivalent information

Case 2: Containment
One representation includes all of the information contained in the other but also more (in the form of finer-grained distinctions)

Case 3: Non-comparibility
The two representations contain different information, which is not directly comparable (and exactly because of the existence of this third possible case we have a partial ordering and not a total ordering relation).

For example, consider the partial representations of reality implied by the following three verbal descriptions:

D1) There exists a blue triangle and a red sphere
D2) There exists a blue object and a red object
D3) There exists a triangle and a sphere
D4) There exist two different objects.

Both D2 and D3 contain less information that D1 (and thus would be among the descendants of D1 in the partial ordering lattice). However they are not informationally comparable to each other (D2 to D3). Of course, D2 and D3 contain a common "core",

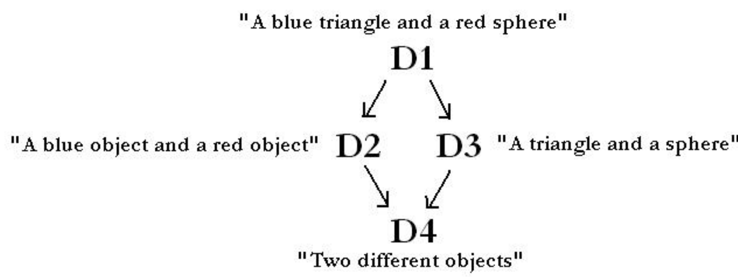


Figure 3-3: The position of the descriptions D1-D4 within an informational lattice.

which has the verbal description D4. The positioning of the above descriptions in the informational lattice can be seen in figure 3-3.

Let us now try to formalize the above with a set of well-defined relations among descriptions: First, let us introduce "informational containment":

A description A is *informationally contained* in another description B, if knowledge of the contents of B allows me full knowledge of the contents of A, but not vice-versa

Thus, in the above example, D2 is contained in D1, D3 is contained in D1, but neither D2 is contained in D3 or vice-versa. Now, we can move on and also introduce "informational intersection" and "informational union". Almost mirroring set theory in our definitions, we define:

The *informational intersection* I of two descriptions A and B is a description that has the property of being informationally contained in both A and B, and which furthermore has the property that no other description I' exists which contains I and which is contained in both A and B.

In the above examples, D4 is the maximal intersection of D2 and D3. Finally, let us define informational union:

The informational union U of two descriptions A and B is a description that has the property that both A and B are contained in U and that no other description U' exists in which both A and B are contained and which is contained by U.

Again, in the above examples, D1 is the union of D2 and D3.[7]

Levels of agreement between two situation models

Now, we can finally attempt an answer to the question: How many levels of agreement can we have among two situation models, regarding the generality of their agreement? Let's try enumerating some:

[7] The above definitions of informational equivalence, containment, intersection and union can also be defined in an alternative formulation in terms of parametrisations of mathematical descriptions of reality and existence of functions, for both deterministic and stochastic viewpoints. For more details, see p.18 of [Mavridis2005b].

L1) We can have absolute agreement given any possible input: i.e. All the functions of the four stages are identical, at most allowing permutation of indices j,k at the output.

L2) We can have informational equivalence given any possible input: Given whatever input, the outputs of the two systems are informationally equivalent.

L3) We can have informational equivalence given a set of specific inputs

L4) We can get informational containment given a set of specific inputs

L5) We can get at least some informational overlap (i.e. non-empty informational intersection) given a set of specific inputs

Up to what extent can we translate?

Now, we have equipped ourselves with enough tools in order to revisit Q3.2b. Up to what extent can one translate from the "worldview" of the first agent to the "worldview" of the second? That is, given that an external observer knows the "situation-viewpoint" of two agents, and the contents of the situation model of one of the two, how much of the situation model of the other can he recover? This question is of immense importance for practical applications, not only for an external observer, but also for either one of the two agents - the answer to the above question dictates the extent of the possible estimates of the other agent's situation model. This extent, in turn, affects the possible communication across the agents. Let's have a look at some cases:

In the first case (L1), no translation is necessary - complete mutual sharing is possible. In the second (L2), a translation is guaranteed to exist, that will enable perfect knowledge transfer, given any input. However, the translation still has to be negotiated. The situation is similar in L3, but only for some inputs - some external situations, not all. In the case of L3, for those inputs where equivalence cannot be achieved, we expect L4 to hold. Of course, L4 can also hold on its own. In that case, the informationally "superior" agent can try to translate all of the knowledge of the "inferior" to a subset of his own; however, all of knowledge will not be accessible to the inferior, and thus the converse will not be successful. Finally, in most practical cases, we expect L5 to hold. There, only a subset of the information possessed by each agent can ever by translated in terms of the other's knowledge. Notice that all of the above are upper bounds on what could be done if we knew the translation function; in practice, this function has to be empirically derived (even if partially) through negotiation.

Now, let's see a practical application of the above:

A usage of the notion of "informational intersection"

For example, consider that we have two agents (A and B) communicating, where one of them (A) has just uttered a referential description of an object - something like "the small red ball". What would be the minimum level of agreement that could guarantee successful referent resolution, if we assume that the translation function has been successfully negotiated? This question is of particular importance for communicating agents - when they are planning their utterances, in order to plan a description of a referent, the speaker must take into account his estimate of the hearer's situation model, and plan his description accordingly. Let us now attempt an answer within our framework: If the current situation is within the set of specific inputs mentioned in L3-L5, then for the general case, the answer is L3. Of course, if (as is often true) the number of commonly visible objects is small, and a partial description can be adequate to single out a referent, then even agreement across a single property dimension might be enough - all the way down to L5.

A usage of the notion of informational "union"

Thus, above in the case of L5, we saw an application of the notion of informational "intersection". However, we have not yet dealt with "union". Imagine a third, external agent C, who wants to estimate the situation model contents of two other agents A and B. How "detailed" must his view of external reality be, in order to be able to include the "subjective views" of both A and B in it? This question is again of particular importance, this time for the case of an agent that tries to "learn" a language by observing interactions of others. Within our framework, the answer is, that his model of external reality should at least be informationally equivalent to the informational union of the situation models of A and B. Thus, these two important applications (description planning and theory of mind for learning by observation), demonstrate the descriptive precision and power of the proposed framework of informational relations: containment, union, and intersection.

3.2.4 Recap of section

Thus, in this section, we have tried to address three questions regarding the alignment of situation models across agents. First, we asked in what ways can their situation models differ? (Q3.2a) Then we asked: how can two beings that have different Reality-to SituationModel pipelines, be able to negotiate meaning among them, and effectively communicate? (Q3.2c). And finally, we asked: Up to what extent can one translate from the "worldview" of the first agent to the "worldview" of the second? (Q3.2b). On the way, we introduced various notions that proved helpful towards this endeavor: the "situation-viewpoint" of an agent, the partial ordering of descriptions, and the notions of informational equivalence / containment / intersection and union.

Topics of similar flavor exist in the literature - albeit not implicating situation models, and not at this level of analysis. For example, Goldstone in [Goldstone and Rogosky2002] presents a model of conceptual alignment across beings; however, the simplifying modeling assumptions taken, and the narrowness of the scope, deem this to be an unrealistic over-simplified toy-model with little explanatory power or possibility of use in real-world systems.

Finally, as we have noted in the beginning, it is worth restressing the importance of the three questions that were asked - an importance which towards the vision of SCAs arises from the fact that they are crucial towards effective human-machine communication and natural language learning for machines. We will also refer back to this section later in this thesis, while discussing embedded situation models and mind reading[8].

[8] which are discussed in section 5.1.1 and also while discussing future extensions which are leveraged

3.3 Recap

In this chapter, we introduced the overall agent architecture framework within which the proposed GSM-based Agent Model will be positioned. First, in section **??**, a brief narrative recapitulating in a few paragraphs the essence of the previous chapter was presented, which opens the way for the explicit and detailed presentation of the agent architecture and its constituent representations and processes in section 3.1. Finally, the very important question of alignment across two different situation models is discussed in 3.2. Now, having seen the overall agent framework, let us introduce Grounded Situation Models, motivating them through multiple pathways, and driving their design through two explicit desiderata.

significantly by the current system in chapter 9, among other places.

Chapter 4

Introducing Grounded Situation Models

In this chapter, initially multiple motivational pathways towards the introduction of GSMs will be given (section 4.1). The largest portion of the chapter (section 4.2) will deal with the derivation of two main characteristics of the overall GSM design (hierarchical GSM structure with embedded GSMs, triple-layered property descriptions) through two explicit desiderata for GSM's (NL-like parsing of situations, bidirectionality between language and the senses). Finally, section 4.3 will address the question of whether GSMs can be considered as "imagistic" or "propositional" representations.

4.1 Motivating GSMs

There exist many ways to motivate the introduction of GSMs: some start from an engineering perspective, others from a more scientific outlook, yet others come from everyday "folk psychology". Here we will briefly visit some. First, let us start from the engineering viewpoint and discuss the practical roles of GSMs:

- *How do we interconnect language with perception/action subsystems in an SCA?*
  Such a role and motivation behind GSMs was already exposed in section 1.4: acting as a hub interconnecting language with perception and action subsystems - see figure 4-1.

- *How can we find a systematic way to build an SCA starting from a behavioral specification?*
  This role of GSMs was already mentioned before[1]: acting as a starting point in the proposed SCA design method. The behavioral specification is translated to a GSM representation with associated processes, and the rest of the SCA is built around the GSM. The specifics of this approach will be given in appendix D.

Now, leaving aside engineering considerations, one can try to see what the GSM resembles in established AI or cognitive science terms. For example:

[1] in section 1.2.6.

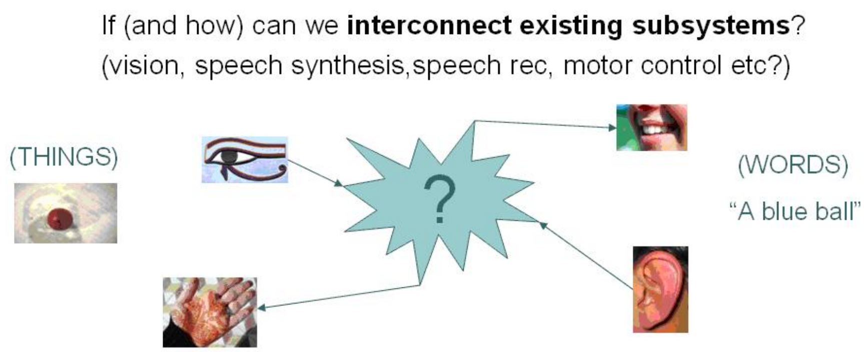


Figure 4-1: GSM as a hub interconnect existing subsystems

- *GSMs as "structured blackboards"*
  In more traditional AI-terminology, the representational part of the GSM is a sensory-updated "structured blackboard". Of course, the GSM is also a type of "internal model", in Marvin Minsky's terms (see section 8.5.9).

- *GSMs as cognitive science-like "situation models"*
  In terms of its contents, the GSM shares many of the characteristics of the "situation models" described in the cognitive science literature (see section 1.3.3).

Finally, one can use less scientifically precise "layman terms", and use the "theatre of mind" metaphor[2]:

- *GSM as a "theatrical stage"*
  One can view the GSM as a manipulable "theatrical stage" in the agent's "mind", which may be filled either with the contents of the agent's present here-and-now physical situation, or a past situation that is being recalled, or an imaginary situation that is being described or planned.

Having hopefully provided adequate motivation for the introduction of GSM's, now we will move on to trying to drive the overall design of our proposal through two explicit desiderata.

4.2 Desiderata driving the design of the proposed architecture

In the given proposal of the representations and processes comprising the overall architecture of GSMs, two main desiderata have driven our choices: First, that the GSMs should "parse" physical situations into the *ontological types that reflect human natural language semantics*. Second, that they should enable fluid *bidirectional translation from sensation-to-language and language-to-action*. Let us consider each in turn:

[2]For the repulsion of the classic homunculus attack that normally accompanies this term in the case of GSMs, look at section 8.5.7. Also, for the classic mental imagery debate and how it relates to our proposal, look at section 4.3.

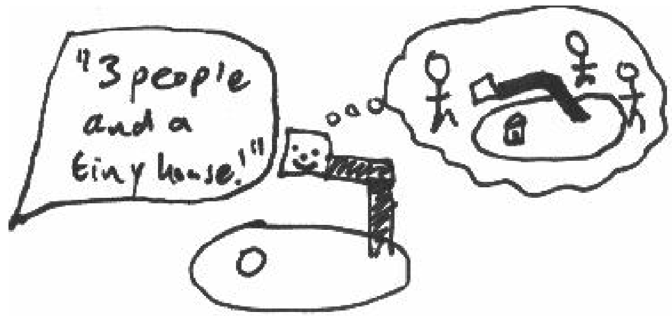


Figure 4-2: The "theatrical stage" of the robot's mind - the bubble that contains "depictions of thoughts" in comics

4.2.1 First desideratum: NL-like parsing of situations

At first sight, this seems like a vague statement. How do natural languages parse out situations? We will try to investigate this question. But first, let us ask a different questions: Are there other ways to parse situations, too, that would be different than the way natural languages do?[3]

There indeed are numerous such ways. For example, consider the way that a *3D scanner* would parse a given situation, say me sitting on my desk where my laptop is: the situation would become a list of triangular patches, with the positions of their vertices and some apparent color accompanying them. Then, my laptop and I are not considered as separate entities: we are both represented by undifferentiated triangles. Also, apart from no distinguishing of "objects" (apart from their constituent primitive triangles), there is no explicit noticing of properties: there is no direct noticing of size or weight or anything like that, while in language there clearly is. Also, there are no mental entities such as my thoughts or emotions in the parse of the situation that the 3D scanner provides - while in language there clearly are.

Now consider yet another different way: the way that a *reductionist chemist* would describe the above situation: a huge list of atoms, bound into molecules, occupying different positions.

In contrast to the 3D scanner description and the reductionist chemist description, let us now focus on a step-by-step examination of how *natural languages* usually decide to parse situations to constituent parts. Consider an utterance. On the surface, utterances can be classified into speech acts, and can be broken down into syntactical constituents, ultimately decomposable to syntactical units such as nouns, adjectives, verbs. Getting deeper, we have semantic units such as agent, patient, action etc. But what is the underlying situation model implied by these units? The assumption made here is that, across natural languages, there is a pretty much similar situation model structure which is implied by them. And this situation model that underlies natural language, along the

[3]A relevant discussion regarding the possibility of many different species of descriptions of *external reality* was given in section 2.2. There, we briefly considered some examples of such descriptions, such as those used in classical physics, possible-world semantics etc. Descriptions of *situations* are descriptions of external reality, but only *partial*: they are constrained only to the aspects of it that are relevant, and usually only cover a limited spatiotemporal extent.

lines of this proposal, should be modeled as consisting of *objects* and *events*, each having *parts*, carrying *properties*, and entering into *relations*[4].

It is worth noting here that the objects and events that appear in natural language, do not only belong to the *physical* realm, but also to the *mental*. Whether philosophically one realm might be reducible to the other is not relevant to our discussion here; what matters is that in language both my "thought" as well as my "limb" exist as objects, and both "I realized" as well as "The ball moved" exist as events. Of course, after deciding to represent both realms, one has to create processes that update the contents of these representations, which ultimately must be derived through the senses (at least for the minds of others). I.e. I have no other choice but perceiving "limbs" and "thoughts of others" ultimately through my eyes and ears.

Descriptions of the state of the physical (physical objects, parts & properties)

Let us start with the physical, by considering the way that objects are represented in our proposal: these might be either animate (agents) or passive objects. (Example: "man" versus "table"). Passive composite objects consist of parts (simple objects), which have properties. (Example: legs of table (parts) and size/color of legs (properties)) Animate agents are two-partite: they have both a directly observable physical (body) as well as a hypothesized mental part (mind). The body is described with similar terms to passive objects, with the exception of an expectancy for more complicated self-kinesis[5].

Descriptions of the mental and "embedded" GSMs

There is no standard way, though, of describing the mind[6]: depending on the complexity of the task, one might conjecture percepts, beliefs, intentions, goals and so on, as well as many different configurations and types of processes operating on them. Different areas of psychology implicitly propose quite different models, and furthermore other such models are implied in the folk psychology which is embedded in language. However, knowledge of a model of one's own mind, can provide a starting point for conjecturing a model of the mind of others. This is the approach taken here: we will suppose that as a starting point, the robot will model the minds of others on the basis of its own agent model (SCA model). And furthermore, we will start by assuming that others have a GSM with the structure of his own (but different contents), and that the senses of others affect the change of their GSM contents through processes similar to the robots (but driven through different sensory data, given the different locus/orientation of the sensory organs of others)[7].

---

[4]The specifics of the representation chosen for describing situations in the GSM proposal follow later in section 5.1, and their customisations that were implemented for Ripley the Robot are described in section 6.3.

[5]The specifics of the representation chosen for describing the state of the physical in the GSM proposal follow later in section 5.1.2, and their customisations that were implemented for Ripley the Robot are described in section 6.3.

[6]See also the similar important comment made by Whitman Richards in [Richards1990].

[7]The specifics of the representation chosen for describing the mental in the GSM proposal follow later in section 5.1.1, and their customisations that were implemented for Ripley the Robot are described in

For example, I might "carve" my beliefs about the current situation in a specific way: I might only notice color and size properties for all simple physical objects. Thus, when I try to model the mental contents of another human, I might start by conjecturing that he "carves" his beliefs about the current situation in a similar way that I do, but that even though the structure of his beliefs is the same, the actual contents might differ. Then, within my GSM, I will include an "embedded" GSM that corresponds to my conjectured structure and contents of the mental part of the other human. Thus, the GSM of the robot will also have to include "embedded GSMs" with conjectured structures and estimated contents, which will represent the situation models and contents ascribed to other agents (for example, human partners). Differences among the robot's GSM and the embedded GSMs ascribed to the human play a crucial role: for example, if the robot knows that human has not seen that there is a red ball on his left but that the human needs it, it might choose to inform the human by uttering: "there is a red ball on your left!", and thus aligning this partial content of the human's GSM with his. Thus, "mental" objects and events (as found in natural language) will ground out to entities within either the self-model of one's mind or within the embedded models of other's minds. Of course, these entities are not directly observable, and have to be conjectured on the basis of sensory data (at least for the minds of others) - and thus effectively, as expected, from the observer's point of view, "beliefs" and "intentions" are just convenient hidden variables used in order to explain and predict long-term patterns of observable physical behavior[8].

Descriptions of temporal sequences: moments and events

However, so far we have mainly dealt with "static" entities, such as objects. Once time is introduced, the notion of events appears, quantizing the temporal in a way similar to the way objects quantize the spatial. Also, in a similar manner to which objects can be referred to in language ("the small red one"), events can be ("when I moved my hand"). Events are being represented in the current GSM architecture as being caused by appropriate sequences of "moments", i.e. momentary snapshots of GSM state. Thus, an appropriate sequence might create a "movement started"-type event for object No. 34 etc.[9]

Recapitulation of first desideratum

In order to satisfy the first desideratum (carving out situations in a way similar to natural languages), the GSM contains representations of the self, the agent's communication partner, and salient passive objects that form the physical common ground. The representations of the self and communication partner contain descriptions of both the physical (body parts and properties) as well as the mental (beliefs and desires) aspects of them. These representations are hierarchical - for example, the physical starts at

section 6.3.

[8]Also see the discussion on the "intentional stance" and GSMs in section 8.5.6.

[9]Specific designs for event representations were proposed and implemented, and will be presented later in this thesis, in particular in sections 5.1.4 (general form) and 6.3 (customized for Ripley the Robot).

the level of composite objects, which break down to simple objects and their relations, which in turn break down to a set of lowest-level nodes representing properties (color, shape etc.).

4.2.2 Second Desideratum: Bi-directionality between senses and language

Now, let us consider the second desideratum: how can we achieve bi-directionality between the senses and language? First of all, why do we need bi-directionality? Let us suppose the robot is asked a question, such as: "what color is the ball on your left?" while the ball is visible. Then, the robot needs to translate visual sensory data to linguistic. On the other hand, if the human informed the robot that "there is a black ball on your left" while the robot is looking elsewhere, then the robot would need to be able to instantiate an appropriate token in its situation model, which could provide sensory expectations to be used when later looking towards the left. Furthermore, such a token should be capable of later being augmented, and ideally interchanged with a sensory-derived token. I.e., when the robot actually looks towards that direction, it should be able to match what it sees with what it expects, and in the case of success to augment its knowledge of the black ball with the sensory-derived information. For example, after seeing it it might be able to know its exact size, while before it only had some rough prior expectations. Thus, it is clear that we need bi-directionality, if we are to handle both *requests for information* speech acts (question answering) as well as *inform-type* speech acts. The *triple-layering* of the proposed GSM architecture is what enables it to achieve the required bi-directionality. The first layer readily interfaces to stochastic sensory information, the second to categorical information which connects to language, and the third to single-valued continuous parameters required for driving action routines.[10]

First example of the purpose and form of the three layers

An example will help clarify the purpose and form of the three layers. As was said before, the "lowest level" nodes of the hierarchy of physical objects are PROPERTY representations: boxes holding the estimated color, size, shape, weight of objects etc.

Let us consider the case of the color of an object, for example a red sphere. Here we are motivating the introduction of three different types of representations due to the suitability of each for different tasks - and we are doing so pretty much along the lines of the discussion of Minsky in chapter 8 of [Minsky2006]: How should a brain proceed to select which representation to use? As we have emphasized several times, each particular kind of description has virtues and deficiencies. Therefore it makes more sense to ask, "Which methods might work well for the problem I'm facing-and which representations are likely to work well with those methods?" How should we represent this property (color) of an object? Should we use a vector? Should we use a probability distribution over vectors? Should we use a set of discrete categories? The answer is: it all

[10]Specific designs for the triple-layered representations were proposed and implemented, and will be presented later in this thesis, in particular in sections 5.1.3 (general form) and 6.3 (customized for Ripley the Robot).

depends on what one wants to do with the representation. Let us thus here discuss three purposes of the representation, and propose suitable forms for each purpose.

First Purpose: SENSING - Accumulating sensory-derived information & providing sensory expectations
Proposed Representation: The Stochastic Layer

First, having chosen a color space for the description (RGB for simplicity) suppose that we see the sphere, and accumulate the following measurements over time: (0.9, 0.1, 0.05), (0.91, 0.09, 0.04), (0.89, 0.11, 0.06), ... It is clear that there exists variation in the measurements, which might be caused either by noise or by a variation of illumination conditions or (rarely in the case of color) by a variation of the "true" color of the object. What the cause of the variation (or uncertainty) is, is not our prime concern here. What is needed, is to find a way to *represent some knowledge of the variation*: and a natural way to do so is through a probability distribution, for example: Mean value = (0.9, 0.1, 0.05), diagonal variance = (0.01, 0.01, 0.009) Of course, we could have chosen a parametric form or a histogram approximation or any other form: what matters here is not the specific representation chosen, but the fact the we can represent multiple alternative values and assign probabilities to them.

Now consider the case of *sensory expectations*: I have seen a red object on the table a while ago, but in the mean time I have moved away and the red object has not been in view for a while. Now, I return to the table - I expect to see the object again, having pretty much the properties it had the last time I saw it - i.e. the same color (or similar, due to lighting variations - or natural change, such as rottening of an apple), the same position (unless it has rolled or someone or the wind has moved it), the same or similar size etc. I indeed look towards where I expect the red object to be - and I get some sensory information. Say that indeed I see an object, and I measure its properties - its color, position, etc. Is it the same object, or another? In order to answer this question, I need to match my sensory expectation with the incoming sensory information - if the match is adequate, then I can consider this to be the same object - else, I have to suppose the old object is not there anymore (or has ceased to exist), and a new object has been introduced[11]. Again, the probability distribution described above can serve as an appropriate way to represent my sensory expectations, and effectively encode any uncertainty due to the possible drift of the properties with time.

Thus, to summarize, this is the purpose and form of the Stochastic Layer:

Purpose: accumulating sensory-derived information and providing sensory expectations for comparison with incoming data

Form: representing multiple probabilistically weighted alternative vector values through a probability distribution.

Second Purpose: ACTION - Deriving action parameters or depicting situations
Proposed Representation: The Continuous Layer

[11] more on this in the discussion of the "objecter" in section 6.4.

Now, suppose that we want to *depict* the seen object, by drawing it. We cannot afford to keep multiple alternatives anymore; a single value is needed. We might for example choose the value[12]: (0.9, 0.1, 0.05) Furthermore, when driving other *action parameters*, very often a single value is needed. For example, no matter whether we have uncertainty about the position of an object that we are about to grasp, a single initial target position has to be fed to the motor control subsystem, even if it is to be refined later. Thus, to summarize, this is the purpose and form of the Continuous Layer:

Purpose: deriving action parameters or depicting situations

Form: representing a single "most likely" property value, through a continuous vector.

Third Purpose: LANGUAGE - Describing situations or imagining through descriptions

Proposed Representation: The Categorical Layer

Finally, consider not having seen the object, but having heard that "the object behind you is red". Then, a categorical description is required, corresponding to a choice among a finite number of verbal categories. And indeed that is what the Categorical Layer provides: a probability distribution upon a set of discrete verbal categories (such as "red", "green", "blue" etc.). Thus:

Purpose: describing situations or imagining situations through descriptions

Form: a probability distribution upon a set of discrete verbal categories ("red" etc).

Now, having briefly introduced the three purposes that the representation needs to serve and the three layers of the representation that serve the different purposes, we can also try to re-examine the purposes through a different viewpoint: through the classic dichotomies of continuous vector / discrete symbol (or signals/symbols, or derived from vision/language) and single-valued / probability-weighted multi-valued (or deterministic / stochastic). Notice how these correspond with the three proposed layers:[13]

Stochastic Layer: Probability-weighted continuous vectors

Continuous Layer: Single-valued continuous vector

Categorical Layer: Probability-weighted discrete symbols

Thus, the three layers are able both to serve the three purposes (sensory data-related, action parameter-related, and language-related), but also effectively cover the continuous vector vs. discrete symbol and single-valued vs. probability-weighted multi-valued dichotomies. In this way, as we shall see, they are able to bridge two well-established traditions: The *signals / connectionist* tradition (a mathematical / parametric description of situations), with the *symbols / logical sentences* tradition (a logical/sentential description of situations).

[12]Some sensible choices include choosing the mode or the mean of the stochastic layer to feed the continuous.

[13]The fourth possible combination, Single-valued discrete symbol, is discussed in the section on bridging with FOPC in section 8.4.1.

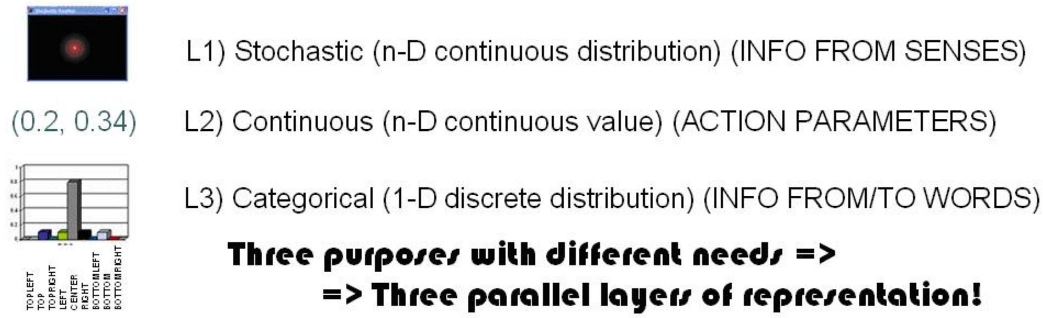


Figure 4-3: Three purposes & three parallel layers of representation(each one serving a purpose)

Second example of the purpose and form of the three layers

As a further example, consider a ball that partakes in a situation, and consider another of its properties - this time not its color as in the previous example, but its position. Imagine that you:

C1) See a ball and estimate its position with some uncertainty

C2) Need to grasp the ball

C3) Hear that "the ball is at the left"

The question then again becomes: What is a natural rep of position for each of the cases C1-C3?

Once again, the first case is best served by a "stochastic layer" representation, the second by a "continuous", and the third by a "categorical", as further illustrated in (figure 4-3)

Again, notice the continuous vector / discrete value and the single-valued / probability-weighted multi-valued dichotomies, and how they are reflected in the three layers (figure 4-4):

L1 (Stochastic Layer) is n-D continuous vector / multi-valued probabilistically weighted

L2 (Continuous Layer) is n-D continuous vector / single-valued

L3 (Categorical Layer) is 1-D discrete value / multi-valued probabilistically weighted

The fourth combination is 1-D discrete value / single-valued, which essentially is FOPC-like logical sentences (see section 8.4.1).

Notice the information loss across the arrows in figure 4-4: for example, when we move from stochastic layer to continuous layer, we need to choose a single value - thus, many different stochastic layer content might have been projected to the same continuous layer value, and thus total inversion is not possible given just the continuous layer.

Processes accompanying the GSM

A number of specific processes accomplishing the following tasks form the crux of the GSM: translating among the three layers, feeding either verbal or sensory information to

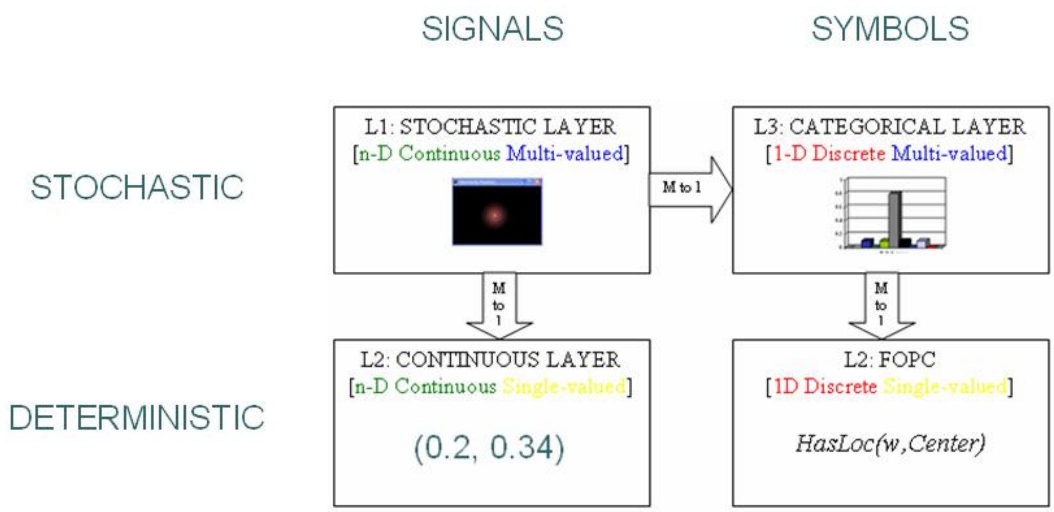


Figure 4-4: The four combinations across the two dichotomies: n-D continuous / 1-D discrete and single-valued / multi-valued. The three first combinations are the three layers, the fourth is FOPC-like)

them, matching sensory expectations to them, or deriving words and action parameters from them.[14]

Recapitulation of second desideratum

Towards the second desideratum (bidirectionality between sensory-and-verbally derived information), the contents of each lowest-level node "ground out" to three layers of representation. The first layer ("stochastic layer") contains a multi-dimensional continuous probability distribution over property values, encoding information in a form suitable for interfacing with sensory subsystems. The second layer ("categorical layer") contains a uni-dimensional discrete distribution over verbal categories, encoding information in a form suitable for interfacing with language subsystems. The third layer ("continuous layer") contains a single multi-dimensional continuous value, which can provide single-valued action parameters for motor control subsystems. These three layers enable the desired bidirectional translation: verbally described "imagined" objects and events can be augmented / verified through sensory information and vice-versa. Apart from specifying the representation, a set of basic processes is proposed. These perform elementary GSM operations, such as updating the contents with recent sensory or linguistic information etc.

4.3 "Amodal" GSMs and the imagery debate

Finally, at this stage it is worth addressing the well-known but often quite superficial and ill-posed "mental imagery debate" [Kosslyn1985]. In its simplest form, and using quite

[14]These processes are described in detail in section 5.2, and their implementation on Ripley the Robot is described in section 6.4.

ambiguous layman's terms, the debate goes as follows:

The question is: How do we represent "ideas" in our mind - using "mental images" or "language-like sentences"?[15]

The two sides representing the two possible answers are:

(The imagists) We use "mental images" - when I think of an apple, something that resembles in some way the picture of an apple "lights up" in my head.

(The propositionalists) We use "language-like sentences" made up from symbols - when I think of an apple, the node corresponding to the symbol "apple" "lights up" in my head, together with other nodes related to it.

Here we take neither side - it is not just images that exist in the head, nor just unstructured language-like propositions, nor just a mix of both. We propose that a grounded situation model (GSM) serves as a mediating amodal representation that lies at the center of the interconnection of sensory-derived percepts with linguistic structures as well as actions. The GSM is amodal; it consists neither of "images" or "sentence-like strings of symbols". It is not "imagistic" in the sense of being neither a viewer dependent 2D image, nor an absolute 3D spatial model. For example, aside from a geometric description of physical parts, it includes invisibles such as the beliefs of other agents. Nor is it just an unstructured list of language-like propositions, for example describing beliefs or wants or previous discourse. It is a representation accumulating information coming in from multiple modalities (vision, touch, proprioception, language), which has a structure analogous to the situation the robot is embedded in or is imagining.[16]

4.4 Recap

In this chapter, we started by discussing multiple motivational pathways towards the introduction of GSM's (section 4.1). The largest portion of the chapter (section 4.2) was devoted to the derivation of two main characteristics of the overall GSM design that is proposed (hierarchical GSM structure with embedded GSM's, triple-layered property descriptions) through two explicit desiderata for GSM's (NL-like parsing of situations, bidirectionality between language and the senses). Finally, section 4.3 addressed the question of whether GSM's can be considered as "imagistic" or "propositional" representations.

Now, having derived the two main characteristics of the overall GSM design through the two explicit desiderata, we will move in and discuss the specifics of the proposed solution: the details of the hierarchical representational structures that comprise the GSM, together with their associated processes, and a modular implementation architecture for building real-world systems.

[15] Mutual exclusivity presupposed in the naïve version of the debate.

[16] As a reminder, for a a list of defining characteristics of GSMs, the reader should refer back to the foreword.

Chapter 5

Specifics of Grounded Situation Models

In the previous two chapters, the overall agent architecture was discussed (within which GSM-based systems reside), GSMs were motivated and introduced, and an overall architecture was derived from two explicit desiderata. The breakdown of situations into agents, composite objects, simple objects, properties, as well as the three-layered property descriptions (stochastic, categorical, continuous) should by now be familiar to the reader. Here, we will discuss the specifics of the GSM proposal: the exact contents of the GSM representation, the details of the associated processes that operate on it and move data in and out of it, as well as a multi-module architecture for implementing GSMs through intercommunicating processes in real-world systems.[1]

5.1 GSM representation

A specific proposal for a GSM representation, which can be customized to specific demands, and which supports embedded GSM's and bidirectionality between senses and language has been made, and representations and accompanying processes have been coded[2]. An overview of the proposed hierarchical representations can be seen in figure 5-1.

As mentioned before, the overall GSM design was driven by two desiderata. Here, we will provide more details on the design than those that were given in the previous chapter - enough details in order for the interested reader to code general-purpose classes that implement GSMs in his language of choice. These general-purpose classes can be later customized for a specific implementation: starting for example from a set of explicit behavioral goals - as will be done when discussing the customized GSM implemented for "Ripley the Robot" (in chapter 6), or starting from a target psychological test to be passed - as will be done when discussing the quasi-minimal GSM design method, that can produce an SCA that passes the "Token Test" (in appendix D).

As discussed before (chapter 4), in order to fulfill the first desideratum, the GSM should reflect the natural hierarchy of agents, bodies, body parts, and properties that

[1] The interested reader can also refer to [Roy et al.2004] and [Mavridis and Roy2006a].

[2] For pseudo-code containing the data structures and the accompanying processes, consult appendix **??**.

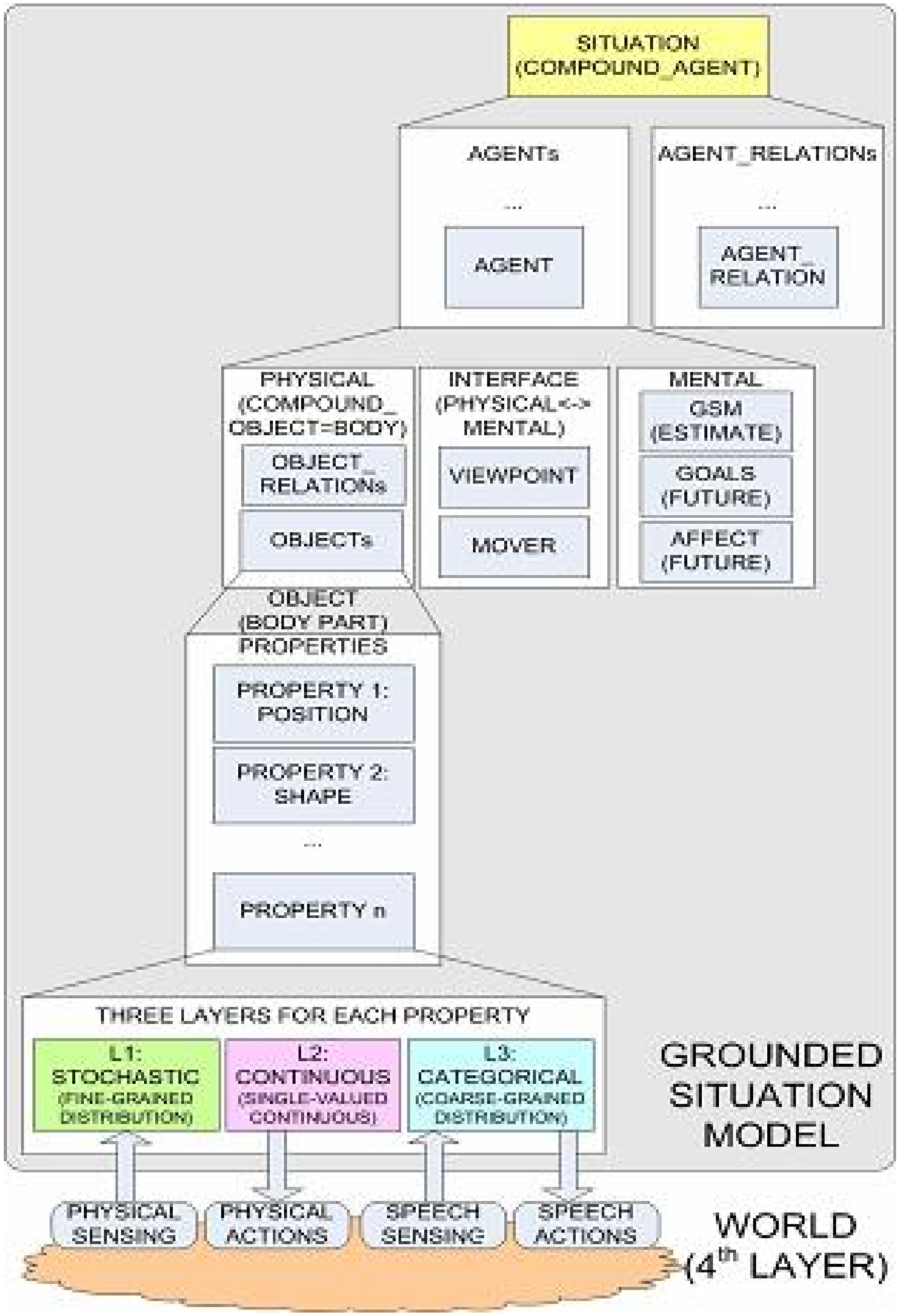


Figure 5-1: Hierarchy of Representations comprising the GSM

```
Present situation:
Time=137
    o  Agent #1 ("I")
        •  Physical:
           Compound Object with 5 objects (body parts)
           Object #1: ("head")
                Property #1: ("position")
                     Layer #1 (Stochastic): GaussMix: G1{m=(1.5,1.89,2), s=2}, G2{...},...
                     Layer #2 (Continuous): (1.6, 1.85, 2.05)
                     Layer #3 (Categorical): #7 ("At the center") with P=.95, ("left") with P=.02, ...
                Property #2: ("pose")...
                Property #3: ("color")...
                ...
           Object #2: ("neck")
           ...
        •  Interface:
           Viewpoint: R eye center (1.4, 1.89, 2.1), viewdir (-14deg, 5deg, 0deg)
           Mover: Head Motor: attachment point (1.5, 1.89, 1.8), rot axis: (...)
           ...
        •  Mental:
           Goals:
           Estimated GSM: (what Robot believes Robot's GSM contents are)
           ...

    o  Agent #2 ("You")
        •  Physical:...
        •  Interface:...
        •  Mental:
    ...    Estimated GSM: (what Robot believes Human's GSM contents are)

    o  Agent #3 (Object on table)
        •  Physical:...
        •  Interface: NONE
        •  Mental: NONE
Past moments: stored situations for t=136, t=135...t=1
Event history:
Event #5: type=start_moving, starttime=127, endtime=127, participants: agent#1object#2
Event #4: type=appeared, starttime=102, endtime=102, participants: agent#3object#1 ...
```

Figure 5-2: Example contents of a situation model during a particular time instant

is implied in human languages. This is accomplished through the *hierarchical representation* shown in figure 5-1. Also, in order to fulfill the second desideratum, namely bidirectionality between language and sensing, the *three layered property descriptions* have been devised (stochastic, continuous and categorical layers, shown at the bottom of figure 5-1). Before considering the hierarchy in detail, let us see an example of the contents of a situation model at a particular instant (figure 5-2):

Let us briefly consider the example of figure 5-2 in more detail. The current situation (for time=137) contains descriptions of the robot (agent No.1), the human partner (agent No.2), and the physical objects (agent No.3). Each agent is broken down to three parts: physical, interface, and mental. The physical part, describing the physical body of the agent, is broken down to a set of objects, which are broken down to properties. Each property is described with three parallel layers of respresentation. Past situations ("moments") are also stored, as well as an event list. Compare this example of contents with the structure of the GSM representation depicted in figure 5-1.

Now, let us move on to a detailed examination of the different structures that partake in the hierarchy. We will start at the highest level ($COMPOUND\_AGENT$), and go all the way down to $PROPERTY$ structures and the three layer-description that they consist of. Then, we will consider sequences of situations occupying more than a single time instant, and special event structures that mark the timeline. The reader can also consult figures 5-1 and 5-2 while reading through the detailed description of the types that comprise the hierarchy. We will break our description of the hierarchy to three parts:

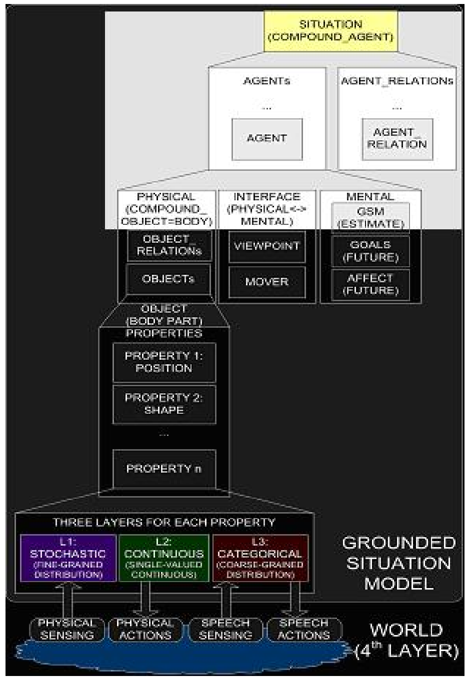


Figure 5-3: The top-level representations

Top-level, mid-level, and bottom-level.

5.1.1 Top-level representations

At the top-level of the representational hierarchy, we find $COMPOUND\_AGENT$ structures, that have $AGENT$ and $AGENT\_RELATION$ structures as their children. In turn, as we shall see $AGENTs$ are divided to three parts: $AGENT$_BODY, $AGENT$_MIND, and $AGENT$_INTERFACE (see Figure 5-3). In more detail:

The $COMPOUND\_AGENT$ type

The first thing to be noticed, viewing the hierarchy top-down, is that the "instantaneous situation" type is termed $COMPOUND\_AGENT$. A $COMPOUND\_AGENT$[3] is a type consisting of all the agents ($AGENT$) that participate in the situation, as well as the relations among them ($AGENT\_RELATION$). Thus, at a specific time instant, all the information relevant to a situation is encompassed in an instance of a $COMPOUND\_AGENT$ type, which will include one $AGENT$ token for the self ("me"), another $AGENT$ token for each human participating in the modeled situation ("you"), and one more for each inanimate object ("it"), as well as $AGENT\_RELATION$ tokens for possible relations among them (for an example, look at figure 5-4(a) below).

[3] The choice of "compound agent" as a term over the more natural "situation" having been inspired by notions such as Plato's "world-being" [Plato1929].

(a) (b)

Figure 5-4: An example: the bodies of three $AGENTs$ participating in a situation-type ($COMPOUND\_AGENT$) - the three $AGENTs$ being the robot (self), the human (user) and a spherical object, construed as potentially agentive. Left: Depiction arising from continuous layer of GSM (A), Right: Depiction of corresponding external reality (B)

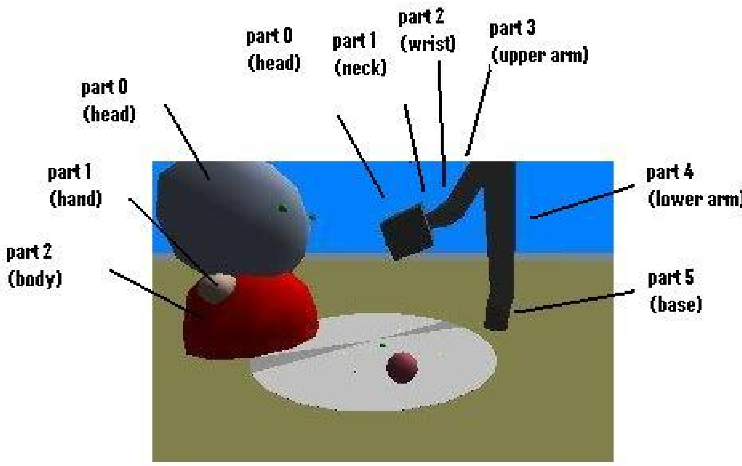


Figure 5-5: Some AGENT BODY parts for Ripley (robot and human user)

The $AGENT$ type

Every agent structure breaks down to a three-part representation, consisting of the body (physical realm), the mind (mental realm), and the interface (between physical and mental). All inanimate objects are modeled as being potentially agentive - however, as long as this is not the case, two of the three parts of the agentive representation (the "mind" and the "interface") remain unpopulated.

$AGENT$ BODY: Physical realm of agent description

The body is modeled as a "compound object", consisting of simple objects (body parts) and spatial object relations. For example, the human body might be modeled as a set of capped cylinders and ellipsoids, each having specific properties. In figure 5-5, you can see the decomposition of the robot's and the human's body into parts for the case of the GSM implementation for Ripley the robot:

$AGENT$ **MIND: Mental realm of description of agent**

The mental realm is represented by a recursively embedded GSM associated with each body, enabling the SCA to model the contents of other minds[4]. For example, imagine an SCA (let's call him S) with two agents residing in his situation model: himself (S) plus a human that is around (H). Let us suppose that each one of them possesses a GSM: $GSM(S)$ and $GSM(H)$. Within the GSM of S, there will be two embedded GSMs (at the first level of embedding): one will be S's estimate of the contents of the GSM of H, i.e. $GSM_S(H)$, and the other will S's estimate of the contents of his own GSM, namely $GSM_S(S)$. In layman's terms, the beliefs[5] of S, contain some beliefs that are "S's beliefs about the beliefs of H", i.e. S's estimates of the beliefs of H, and some other beliefs that are "S's beliefs about the beliefs of S", i.e. S's estimates of his own beliefs.

Two approximations to the contents of embedded GSMs

Here, at a first approximation, one can make either or both of the following simplifying assumptions: First (A1), if we suppose that there is total "conscious accessibility" of S's GSM to himself (i.e. "S believes that he believes everything that he indeed believes"), then it follows that: $GSM_S(S) = GSM(S)$, i.e. the self-model of the GSM coincides with the actual GSM held by the agent. Second (A2), as is the case with young children, at early stages of development of a "theory of mind" [Premack and Woodruff1978], we might also make the "universally shared beliefs" assumption - that is, assume that "S believes that H believes everything that S believes", i.e. that $GSM_S(H) = GSM(S)$. Of course, if we make both assumptions A1 and A2, then the need for "embedded GSMs" is trivialized: essentially, we are taking the viewpoint of universality of belief, or "absolute belief": i.e., that every agent has the same beliefs as we do.

The utility of not making the above approximations

However, all interesting phenomena and all real utility of embedded GSMs, starts when we do not make the simplifying assumptions. For example, one can suppose that S believes that some of his beliefs are not shared with H: then, he could issue an INFORM statement[6] to update H's beliefs ("hey, Jack, there's a tomato over here!"). Or, S, by virtue of knowing the differences between his beliefs and H's beliefs, can comprehend his actions of reference correctly: for example, suppose S believes there are three balls on a table, two of which red and one green. Furthermore, suppose that H has not seen the second red ball, and thus S believes that H believes that there only two balls on the table: one red and one green. Thus, when S hears H saying: "Please, give me the red ball" then S can service this request because S knows, that the description "the red ball" that H has just uttered is not ambiguous, because it is relative to H's beliefs and not S's: thus, it must refer to the single red ball that is contained in $GSM_S(H)$[7]. Thus, we have seen at least two very important uses of embedded GSMs which might have different contents

[4]Optionally other structures estimating other parts (not only the GSM) of the agent model of the other SCA's (such as goal structures etc.) can be included.

[5]Layman's terms have been used here only for simplicity - we should bear in mind that "the set of beliefs" and "the GSM" are not equivalent.

[6]More precisely, an INFORM speech act, as discussed in section 1.2.5.

[7]Again, this is a first approximation. At a second approximation, one must take into account that when H planned his reference, he planned it in accordance not to H's beliefs, but in accordance to H's beliefs about S's beliefs etc.

than the embedding GSM: first, they can provide motivation for INFORM speech acts, and second, they can allow the speaker-and-hearer relativity of referential descriptions to be taken into account.

Second-level embedded GSMs

Now notice that further embedding can take place recursively: for example, again in layman's terms, we might have things such as: "S believes that H believes that H believes", i.e. S's estimate of H's self-model of his beliefs and so on. Similarily, we can have S's estimate of H's estimate of S's GSM and so on. Going down to the second level can prove fruitful in solving certain problems, for example in a better solution of the speaker / hearer-relative reference resolution problem described above, and so on. However, going into greater depths, is seldom required, if ever[8].

More details on usages of embedded GSMs (such as mind reading) can also be found in the next paragraphs on the $AGENT$ INTERFACE, and further discussion of them can be found in sections 4.2.1, 8.1.2, 8.5.6, and 9.1.3.

$AGENT$ INTERFACE: Third part of description of agent

The interface consists of the specifications of contact points between mental and the physical, i.e., "sensory entry" and "action output" parameters. It should ideally contain all the information that would enable: first, prediction of the mental realm given knowledge of the physical reality around the agent ("sensory entry" parameters - such as position / pose / characteristics of vision sensors), and second, prediction of the physical reality around the agent given knowledge of his action decisions. In the terms introduced previously in chapter 3, the "sensory entry" parameters are the essential parameters of the assumed model M1 (ExternalReality-to-Sensor model), while the "action output" parameters are the essential parameters of the assumed model M2 (Action-to-ExternalReality model). Let me elaborate a little more on this.

First, notice that in the previous section where embedded GSMs of others were introduced, we barely mentioned anything about how they are estimated. Notice that in order to *estimate the GSM contents of others*, several assumptions must be made regarding the answers to the following questions:

Q1) What is the assumed structure of the GSM's of others? Does it use the same property dimensions, categories etc.?

Q2) Apart from the structure, is there equivalence in its causal grounding, i.e. is it fed through a similar mechanism from external reality (so that its internal signs signify the same externals?), i.e. in the terms of our proposal, do they have the same Sensor-to-SituationModel process (P1 as introduced in chapter 3) with the estimating agent?

The obvious first-approximation simplifying assumptions to these questions are:[9]

A1) The other agent has the same GSM structure that our GSM has.

A2) The other agent's GSM is fed in the same way through reality, i.e. his P1 is the effectively similar to ours, maybe with the exception of the location and direction of the

[8] Furthermore, under certain conditions, such as not having to deal with complicated fraud detection, it can be shown that going beyond two levels of depth is usually not required.

[9] In case one does not take the above approximations, the discussion made in section 3.2 is highly relevant.

sensory organs. If we assume A1 and A2, then for the example of a simplified agent whose GSM is only fed through vision and hearing speech, it follows that (in layman's terms again): "If I know what he has seen and what he has heard, then I can estimate what he believes"

Now, let's make a third assumption:

A3) My beliefs of the external reality that is local to the agent whose GSM is am estimating are complete and correct

Thus, now the previous statement becomes: "I know what he has seen and heard, because my beliefs cover that. Thus I can estimate what he believes" This is exactly the basis of the simplified mind-reading procedure proposed here. Thus, mind reading can be accomplished through:

- Possesing or acquiring beliefs relevant to the local environment of the other agent.

- Estimating the sensory input of the other agent, on the basis of the basis of the "sensory entry" parameters (essential parameters of the ExternalReality-to-Sensor Model), which are contained in the Interface Part of the description of the other agent.

- Estimating the contents of his GSM, on the basis of by using our model of his Sensor-to-SituationModel process (P1)

For a simple example[10], consider the following:

Let's assume the "sensory entry parameters" of the other agent consist of an "eye position" and an "eye pose", and that we have a way to estimate these (through an assumption or through eye tracking, for example). Then, we can visualize the robot's GSM, through a virtual camera corresponding to the other agent's "eye", and feed the other agent's estimated GSM with the output of the virtual camera, and at the same time feed the other agent's estimated GSM with our audio stream. Then, effectively, we are "reading his mind" - we estimate what he believes on the basis of our estimates of what he saw and what he heard (figure 5-6).

Now, let's move on from the task of *estimating beliefs* to that of *predicting changes to physical reality, which are caused by actions of another agent.* In that case, first we need to do action prediction, and second, we need to estimate the effects of the actions to reality. For the second task, we need to possess the "action output" parameters, which are the essential parameters of the conjectured Action-to-ExternalReality Model (M2)[11]. For the first, we need estimates of more parts of the agent model of the other agent, including his action lists and action selection processes (R5 and P5 in the terms of chapter 3), and so on. Then, after having estimated the other agent's GSM contents through the simple procedure described above, and knowing the rest of his agent model including R5 and P5, we can in principle predict his future actions[12].

Thus, to recap, we have seen what the contents of the "Interface part" of the agent description are, and what their purpose is. First, the *"sensory entry" parameters*: these are the essential parameters of the ExternalReality-to-Senses model (M1), i.e. those that enable us to estimate beliefs of others if we furthermore know about their P1 and R1. And second, the *"action output" parameters*: these are the essential parameters of the

[10]That has also been demonstrated in the past using an implementation on Ripley.

[11]For more details on M1, M2, P1 etc. please refer to chapter 3 and appendix **??**.

[12]The interested reader might also look at the relevant comments made in [Mavridis2005a].

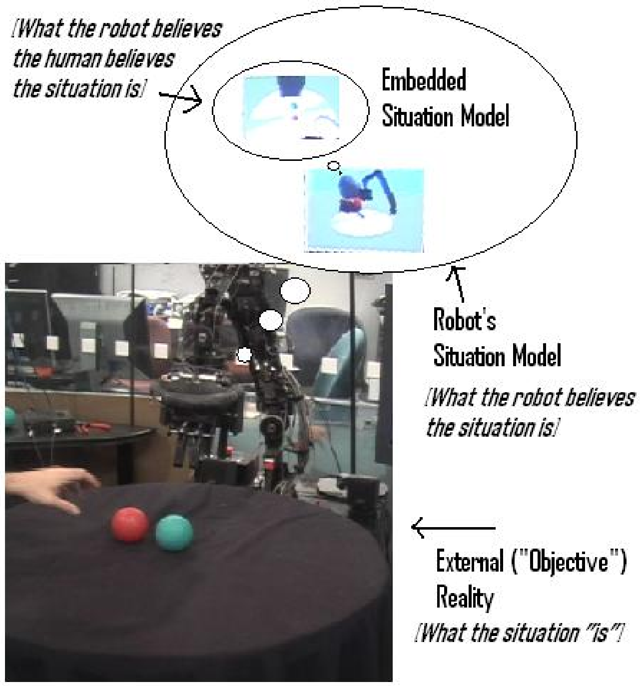


Figure 5-6: Simple Mind Reading through a synthetic camera

Action-to-ExternalReality model (M2), i.e. those that enable us to estimate physical results of actions taken by other agents, that when coupled with prediction of the actions of others can help us estimate the future states of physical reality. In short, one can say that the "Interface" contains all the parameters that situate the model of the mind of the other agent in external physical reality.

Now, having described the three parts of the $AGENT$ structure (mental / physical / interface), let us move on in our exposition.

The $AGENT\_RELATION$ **type**

Contains relations that hold between agents, as well as relative terms. Examples of possible relations holding between two agents: contact, attachment, support or inclusion, or even non-physical relations. Examples of relative terms: relative position, size, or again even non-physical. For a discussion and examples see section 8.3.3.

Now, having seen $COMPOUND\_AGENT$ structures, and how they break down to tri-partite $AGENT$ structures and $AGENT\_RELATION$ structures, let us move to the mid-level: $COMPOUND\_OBJECT$

5.1.2 Mid-level representations

The $COMPOUND\_OBJECT$ **type**

Consists of a list of $SIMPLE\_OBJECTS$, as well as a list of associated $OBJECT_RELATION$s among them.

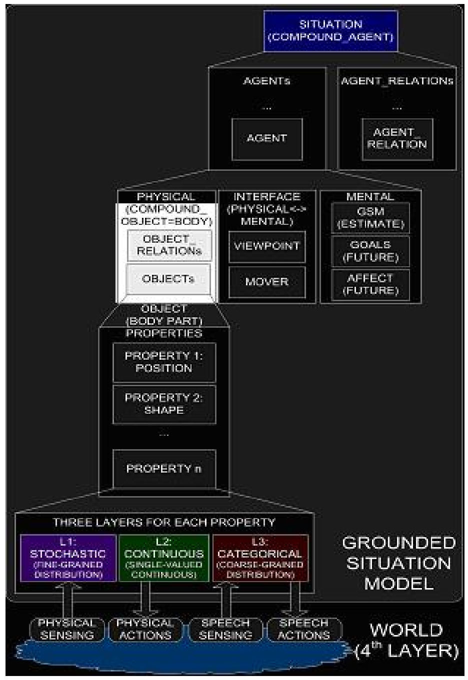


Figure 5-7: The mid-level representations

The $OBJECT\_RELATION$ type

Contains relative terms between body parts, i.e. simple objects, for the case of relations of body parts that are parts of the same $AGENT$. Here, in a similar fashion to $AGENT\_RELATION$, one can here encode relative terms between body parts etc. The difference with $AGENT\_RELATION$ is that in the case of $OBJECT_RELATION$s we are encoding relations among objects (body parts) of the same $AGENT$ (body), while in the case of $AGENT_RELATION$s we are encoding relations across different $AGENT$s, either across specific body parts of them, or by viewing the whole agent as a single entity. For more details see section 8.3.3.

The $SIMPLE\_OBJECT$ type

Objects in the physical realm bottom out in clusters of properties. For example, the representation of a ball might bottom out in a set of properties that model the look, feel, and location of the ball.[13]

5.1.3 Bottom-level representations

The $PROPERTY$ type

In order to fulfill the second desideratum, i.e. allowing bidirectional translation between sensory-derived data/expectations and linguistic descriptions, each property is

[13]For example, a list of the properties encoded for the case of the current implementation on Ripley the Robot can be found in section 6.3.

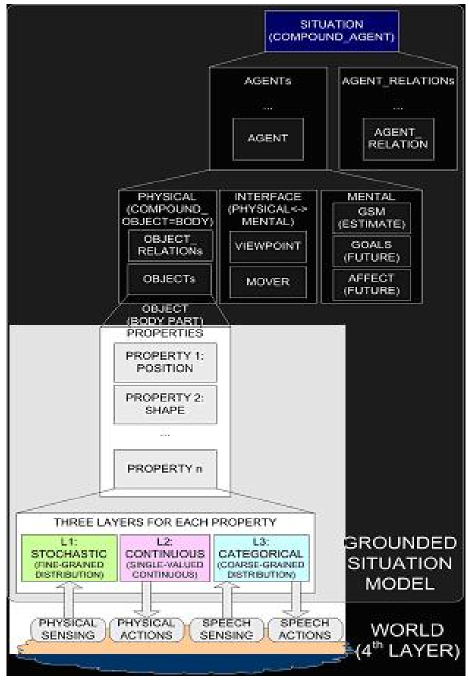


Figure 5-8: The bottom-level representations

encoded by a set of three layers of linked representations (Triad), as motivated and introduced in the previous chapter, in section 4.2.2. As a reminder, by property here we mean a property of a $SIMPLE\_OBJECT$, such as its color, position, pose etc. Thus, as shown in figure 5-8, having described the $COMPOUND\_AGENT$ structure starting from the top and moving down to $AGENT$s, $AGENT\_RELATIONs$, parts of $AGENT$s etc., we are now focusing on the bottom part of the figure, and specifically on the part that is under the spotlight (i.e. the part that is not dark).

A quick summary of the form and function of the three layers to which each $PROPERTY$ breaks down can be found below[14].

Layer 1 (L1 - Stochastic Layer)

Maintains stochastic representations of properties, suited for sensory measurements. Let us assume that we have acquired multiple noisy measurements of the position property of a particular object by computing the centroid of a tracked visual region over time. We would like to encode our knowledge of the position in a summary form, which should give little weight to non-persistent outliers, which should not cause any significant loss of measurement resolution, and which should still retain an ability to remember the spread of sensed values and our confidence in them. We should also be able to lower our confidence when measurements become outdated, and furthermore actively drive the acquisition of more current sensory information whenever required. To satisfy the above requirements, it would be reasonable to represent position property as a stochastic variable, through a probability distribution (e,g., a continuous parametric form, or as we have implemented it, a discretized histogram).

[14]For more details and examples, the reader can refer back to section 4.2.2.

Layer 2 (L2 - Continuous Layer)

Maintains continuous single-valued encodings of properties, suited for use as action control parameters. Consider a scenario where we want to execute an action which requires the position of the object as a parameter. For example, we might want to use the object as a target for a lift motor routine. The stochastic distribution must be sampled in order to guide action. In our current implementation, the continuous layer may be generated by selecting the maximum density point from L1. A second motivation for maintaining L2 is to support simulation based reasoning. To simulate interaction of objects over time, a single value for properties such as size, orientation, and position leads to computationally tractable physical simulation, whereas stochastic representations would be far more complex and time-consuming to manipulate.

Layer 3 (L3 - Categorical Layer)

Maintains discrete, categorical encodings of properties, suited for interfaces with natural language. Consider the scenario of asking the robot where an object is. To respond, a verbal spatial category must be produced and communicated by the robot (e.g., "at the left"). We need to be able to provide a single discrete value corresponding to the verbal category chosen, or better yet, provide a probability distribution over multiple spatial categories. This is what the categorical layer, L3, accomplishes. It represents a property as a distribution over a number of verbal categories (while in L1 we had a fine-grained distribution over sensory-derived measurements). For example, we might have "left", "right", "center" in the case of position, or "red", "blue" in the case of color etc. We have suggested that the categorical layer is motivated by the need for production of verbal descriptions. It is equally motivated by the converse need, translating from verbal descriptions to property representations: the robot might be told that "there is an object at the center". If there is total confidence in the linguistic source, the robot can represent the information as a discrete distribution over the categories, with P(location = center) = 1 and all other probabilities zero.

To summarize: In short, the GSM represents a situation as a hierarchy of objects in the physical realm linked, optionally, to a mental realm. The realms bottom out in a linked three-layered representation comprising stochastic (L1), continuous (L2), and categorical (L3) levels.

5.1.4 Moments and Events

So far, we have only talked about instantaneous "snapshots" of situations. As mentioned in section 4.2.1, the GSM should also be able to deal with sequences of situations. Thus, the types of "moments" and "events" and "histories" are introduced. The "compound agent", i.e. a particular configuration of the GSM, represents a "moment" in time - a snapshot of the state of the situation. Sequences of "moments" create "histories". An "event" is a structure providing landmarks on the sequence of moments. It consists of an event type ID, start/end time indices, and a list of participants (agents or bodyparts)[15]. For example, look at figure 5-9.

[15] One can also note here the analogy between the quantization of space through objects, and the corresponding quantization of time through events.

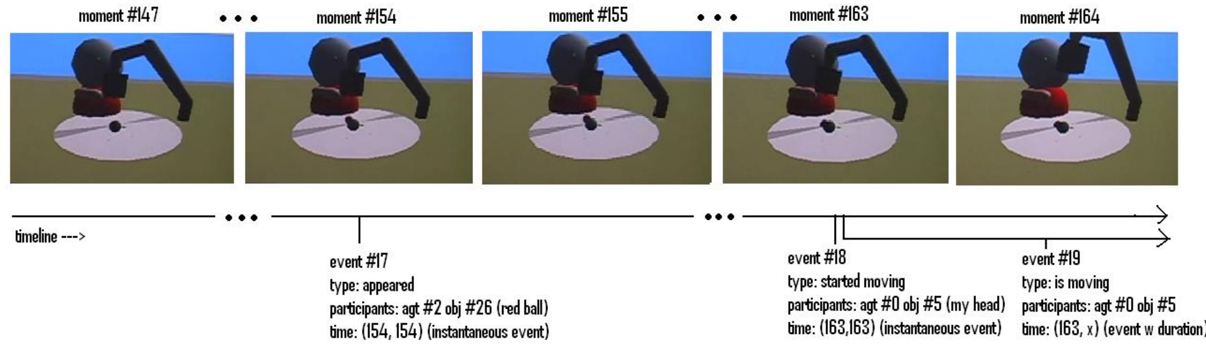


Figure 5-9: Example of Moments and Events

In figure 5-9, an example of a history fragment containing moments and events is depicted. This history fragment starts at moment 147 and finishes at moment 164. It contains three events: event no. 17, event no. 18, and event no. 19. Event no. 17 is an "object appearance" - type event, and its sole participant is the red ball (object no. 26). It is an *instantaneous event*, and takes place during moment 154. In contrast, event no. 19 is an *event with duration*, (of type "ismoving"), with the participant being the head of the robot, and has started at moment 163 but has not finished yet within this history fragment.

Now, having discussed the contents of the basic GSM representational hierarchy, starting from $COMPOUND\_AGENT$ structures and going down to $PROPERTY$ structures with three layers, and having also introduced the moment and event types, we can now proceed and discuss the basic GSM processes that operate upon the described representational structures.

5.2 GSM processes

One can view the GSM as being involved in two basic purposes: belief maintenance (update beliefs about the situation on the basis of linguistic or sensory evidence), and action control (getting action parameters - words to be said, locations of objects to be reached etc). The constituents of the GSM hierarchy, and each object's layered property representation is created and maintained using update procedures described in this section[16]. In this conceptualization, perception tries to "mirror" parts of the fourth layer (external reality) into the GSM. Action, on the other hand, can be seen as trying to "mirror" parts of an imaginary desired GSM into external reality]. A sensory evidence-to-GSM matching processes is also a member of the set of the basic processes of the GSM.

5.2.1 The basic GSM operational cycle

The basic GSM operational cycle is shown below: (figure 5-10)

[16]Note that conceptually, we can treat the robot's external physical world as a fourth property layer ("L0") that interacts with the other layers via sensorymotor processes (as seen in Figure 5-1].

```
//initialisation
InitL1L2L3

//main loop
DoForever {
//Sub-cycle 1: Update GSM through senses or words
        //Sub-cycle 1.1: Senses-to-GSM
        If (incoming sensory data)
                Senses-to-L1
                L1-to-L2
                L1-to-L3
        //Sub-cycle 1.2: Words-to-GSM
        Else If (incoming speech data)
                Words-to-L3
                L3-to-L1
                L1-to-L2
        //Sub-cycle 1.3: DiffuseGSM (If no incoming sensory data or words)
        Else
                Diffuse-L1
                L1-to-L2
                L1-to-L3
        Endif

//Sub-cycle 2: Service requests for words or action parameters
        //Sub-cycle 2.1: GSM-to-words
        If (incoming request for words)
                L3-to-words
        //Sub-cycle 2.2: GSM-to-actionparams
        Else If (incoming request for action parameter)
                L2-to-actionparams
        Endif
}
```

Figure 5-10: Pseudocode of basic operational cycle of GSM - two first level sub-cycles, and a total of five second-level sub-cycles

Notice that the main cycle is first broken down to two first-level sub-cycles: updating the GSM and servicing requests (i.e. Input to and Output from the GSM), which are then broken down to a total of five sub-cycles, depending on the type of data and/or request (words, senses, actionparams).

In the following figure, a pictorial overview of some of the basic sub-cycles can be seen, for the example case of sensing the position property of an object through vision:

In the first column (left-to-right), one can see the contents of the three layers of the position property in the case of arrival of sensory information - the robot is currently seeing the object. In the second and third columns, the robot is looking away and the object is not in the field of view anymore - and thus uncertainty about the position of the object increases. Finally, in the fourth and fifth columns, an object is instantiated not on the basis of visual, but on the basis of linguistic information: "Imagine a red ball at the center" etc.

Now, to familiarize the reader, we will first briefly look through subcycles 1.1-1.3 (situation model updating when sensory, linguistic, or no information arrives), and then consider all the processes in more detail.

Subcycle 1.1: Incoming Sensory Data

When a sensory object reaches vision, first a matching process (part of senses-to-L1) tries to match it to existing objects within the GSM. In this example, we assume that the GSM originally contained no passive objects, and thus upon first viewing the red ball an AGENT containing a single OBJECT within the GSM was instantiated (created). Again, within senses-to-L1, its initial property contents are then updated through the incoming

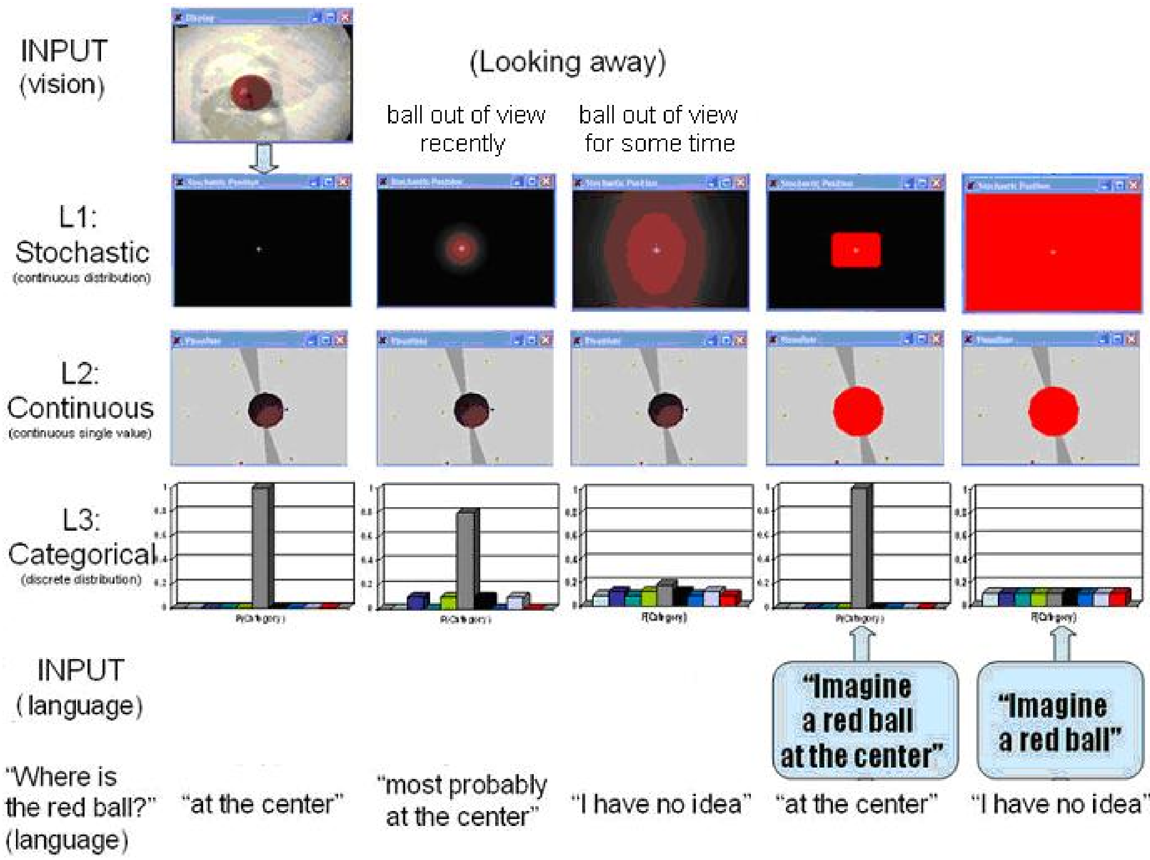


Figure 5-11: GSM Subcycle examples (left-to-right): Column 1 - subcycle 1.1, Columns 2&3 - subcycle 1.3, Columns 4&5 - subcycle 1.2

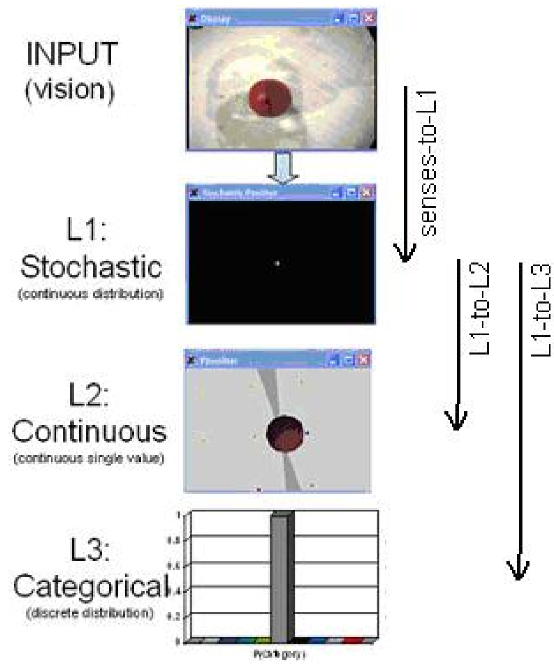


Figure 5-12: Subcycle 1.1 - Objects instantiated through vision (senses-to-L1, L1-to-L2, L1-to-L3)

sensory information. Then, L1 feeds L2 (obtaining a single value out of the distribution in L1), and finally, L1 feeds L3, obtaining a distribution on verbal categories, such as "center", "left" etc. for the case of position.

Subcycle 1.3: No relevant incoming information (sensory or speech)

When no new information about an object arrives (neither through vision nor through language), the uncertainty about the old information needs to increase, and thus a diffusion process diffuses the contents of L1. Then, as was also the case for subcycle 1.1, L1 feeds L2 (obtaining a single value out of the distribution in L1), and finally, L1 feeds L3, obtaining a distribution on verbal categories, such as "center", "left" etc. for the case of position.

Subcycle 1.2: Incoming speech-derived data

In this case, words are translated to a distribution over verbal categories in L3, and then L3 feeds L1 (transforming the verbal category distribution to a sensory expectation), and once again L1 feeds L3 (deriving a single value from the sensory distribution)

Many more details on these cycles and subcycles are given in the next sections, and the reader should also refer to figures 5-11, 5-12, 5-13 and 5-14 while reading through the paragraphs. Let us now move on from the general description to discuss the specifics of all the parts of the subcycles, i.e. the basic GSM processes (such as words-to-L2, L1-to-L3 etc), in more detail:

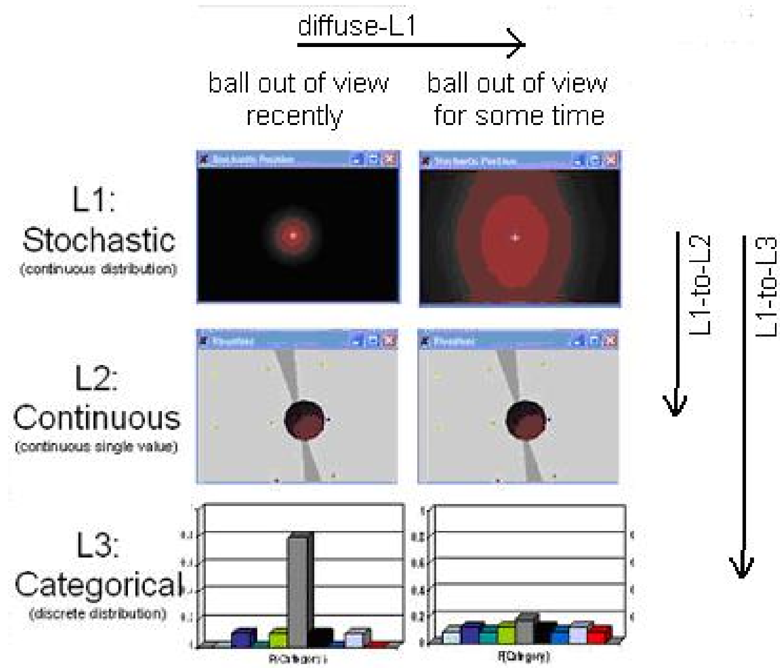


Figure 5-13: Subcycle 1.3 - Persistent objects (diffuse-L1, L1-to-L2, L1-to-L3))

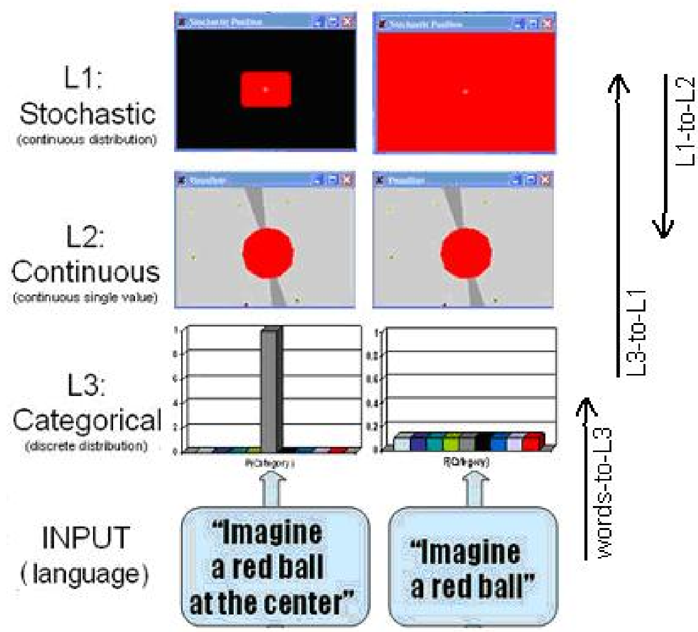


Figure 5-14: Subcycle 1.2 - Objects instantiated on the basis of speech (words-to-L3, L3-to-L1, L1-to-L2).

5.2.2 The basic GSM processes in detail

We will use the updating of an object's position property as an illustrative example (as seen in figure 5-11 from the previous section). We will adopt the notation Ci/Rj for the columns/rows of this figure. Pseudo-code is available at appendix **??**.

Initial State of stochastic layer (Init-L1):

Given no information (sensory or linguistic) about the position of an object, we are faced with a choice: what should be the initial probability distribution on positions? In our robot's particular case, objects are assumed to be on the table - thus the object's location must be bounded in space defined by the surface of the table. As a first approximation the a priori probabilities of unseen object positions are spread uniformly across the table.

Updating the stochastic layer when sensory information is available (Senses-to-L1):

Now let us suppose that an estimate of the position of an object is generated by the visual system. First, a matching process (termed "The sensory-object to GSM-object matching process" and described later in this section) attempts to assign matches between incoming sensory information and sensory expectations based on the previous state of the GSM, which reside in the stochastic or continuous layers. Then, objects are either created, destroyed, or updated depending on the outcome of the matching process. Here, we will first consider the method of updating the stochastic layer contents, in case a match between a sensory object and an existing object in the GSM is found. Later in this section, we will explicate the matching process. Let us now describe the update process. How should the probability distribution of the stochastic layer be updated, given incoming sensory information? We have chosen to calculate the new distribution as the weighted sum of the old distribution with a rectangular envelope centered at the new measurement. In the limiting case, this envelope consists of only one bin, namely the bin which contains the new measurement. The weight and envelope can be adjusted to fit the noise and temporal characteristics of the measurement.

Diffusing the stochastic layer when sensory information is not available (Diffuse-L1):

As a general rule, we assume that over time, knowledge becomes less reliable without information refreshing. For example, let us suppose that sensory information is not currently available about an object's position because the robot is not looking at it. Over time, the robot's confidence in knowing the position of the object should decrease (someone might move it while the robot is not moving, etc.). To model this confidence decay in L1, we use a diffusion process. The new value of each element of the position distribution in L1 is given by the weighted sum of its old value with that of its neighbors within a pre-specified neighborhood. The expected rates of change dictate the settings of the weights. Diffusion parameters are set separately for each property modality. Color and shape beliefs are diffused much more slowly since they are far less likely to shift over time (but color, will, for example, shift in perception as lighting conditions change). For example, in C1 an object has been visible for some period of time and is still visible. In R2C1, the resulting distribution has become very sharp after the object was stable and visible for some time - in fact it consists of a single bin (under the cross-hair). The robot knows where the object is with certainty. In contrast, in R2C2 and R2C3, the robot's head has looked away, and the object has not been visible for some time (C2), and even more time (C3). The diffusion process has taken over and spread out the distribution.

Speech-derived information updating the categorical layer (Worlds-to-L3):

The categorical layer consists of a distribution over a set of verbal positional categories ("right", "center" etc.) . If the robot receives information that the property value "left" was given through speech for the object under consideration, then the robot sets P("left") = 1 while the probability of other categories is set to zero. If such information is absent, it has two choices. Either the system can assume an empirical prior over the verbal categories, or it can use a non-informative uniform prior, and again we have chosen to implement the latter. In C4, the position is specified by the verbal information "...at the center". Thus, in R4C4 we have P("center")=1 while P(other category)=0. In contrast, when no spatial information is given through speech we get a uniform pdf (see R4C5).

The stochastic layer feeds the categorical layer (L1-to-L3):

Whenever information enters the GSM (either via L1 or L3) or when a change occurs due to diffusion, the three layers must be updated in order to ensure cross-layer consistency. If the change has occurred at the stochastic layer, then update information feeds the categorical and vice-versa. The continuous layer is always fed via the stochastic. The stochastic layer contains more specific information than the categorical, and thus the forward feeding process is many-to-one and straightforward. Each property has an associated classifier, which here we will call "categorical classifier". The classifier maps continuous sensory-derived values to categories. The classifier could in principle be implemented by any algorithm, such as SVM's, neural networks, etc. For simplicity, we have implemented nearest neighbor classification around predetermined centers[17]. Initially, all verbal categories are assigned zero probability. Then, each bin of the stochastic layer is considered. The probability of the verbal category associated with the center of the bin (according to the classifier) is increased by the amount of probability that corresponds to the bin of the stochastic layer that is under consideration. As a result, we obtain probabilities of verbal categories as the sum of the probabilities of their corresponding bins in the stochastic layer. The narrowly-spread stochastic distribution in C2R2 has created the narrowly-spread categorical in R4, and the widespread of C3R2 leads to the one in R4.

The categorical layer feeds the stochastic layer (L3-to-L1):

If we try to invert the previous transformation, a one-to-many mapping results. In order to achieve uniqueness, we enforced the constraint that the stochastic layer bins that correspond to the same verbal category should be equiprobable. Originally, the stochastic layer elements are all assigned zero probability. Each category is considered in turn. The elements that correspond to the category under consideration are marked, and the

[17] For more refined spatial models, see discussion in section 8.3.3.

probability of the category under consideration is spread equally among them. In C4, when R4 is fed to R2, the area corresponding to the bins whose centers would be classified as belonging to the "center" spatial category is filled with equal probability. In C5, each category corresponds to a rectangle such as the one shown in the C4R2 for "center", thus the whole of C5R2 is equiprobable.

Translation from the categorical layer to descriptive speech (L3-to-words):

Consider the case of R4C1. Unquestionably, as P("center") approaches unity, the robot can describe its position as "at the center". But things are less clear in C2 and C3. There, according to a decision tree created with preset thresholds on the probabilities of the three most highly probable categories and the entropy of the distribution, numerous different resulting verbalizations occur. For example, if P(most likely category) > 0.7 and < 0.9, then we get "most probably at the <spatial category>" (C2). As a further example, when the distribution is almost equiprobable as quantified by its entropy, then we get "I have no idea" (C3). The decision thresholds that were currently arbitrarily set, but could be empirically learned.

The stochastic layer feeds the continuous layer (L1-to-L2):

Here, we are seeking a single representative value for the distribution of the stochastic layer. Here we have chosen the statistical mean (and not mode), as no bimodal distributions arise. In our example, all the distributions shown in R2 share the same mean, i.e. the center of the table. Thus, if the robot were to look at an object, in both cases the same target fixation point would selected to guide the motor system.

The continuous layer drives action parameters (L2-to-acti):

Here, we simply select the required properties for the action to be taken (for example, the position and size of an object for a grasping action), and fetch it from the continuous layer, that has been fed previously through through the stochastic layer.

The sensory-object to GSM-object matching process (part of senses-to-L1):

Let us present an example - for the case of vision, and for the objects in the GSM that are within the current field of view, consider a case where:
- I have sensory *evidence* for two objects through *vision*:

- ObjV1: sensory property estimates are:
  position(1.2,0.4,0.1), radius(.2), color(0.8, 0.1, .05)
- ObjV2: sensory property estimates are:
  position(0.2,0.5,0.05), radius(.1), color(0.1, 0.1, .75)

- I have sensory *expectations* for three objects *within the GSM*:

- Obj14: sensory property estimates are:
  position(1.25,0.39,0.11), radius(.21), color(0.79, 0.11, .03)

- Obj22: sensory property estimates are:
  position(0.2,-0.5,0.05), radius(.1), color(0.1, 0.8, .05)
- Obj23: sensory property estimates are:
  position(1.2,0.4,0.1), radius(.2), color(0.8, 0.1, .05)

Now the question arises: how should we decide how (and if) to associate sensory objects to our sensory expectations? First, let us tackle the "if" part. Three cases might arise in the matching:

- A sensory object exists which cannot be matched well enough to a sensory expectation (case a)
- A sensory expectation exists which cannot be matched well enough to a sensory object (case b)
- A sensory object exists which can be matched sufficiently with a sensory expectation (case c)

Thus, we must decide what a suitable matching metric and algorithm is, and also what are appropriate actions for the cases a)-c).

Let us tackle the actions first.

- For the case of a), a new object must be created within the situation model (CREATE).
- For the case of b), an existing object in the situation model, has to be deleted, because its sensory expectation was not verified through sensing (DESTROY) [18].
- For the case of c), an existing object in the situation model must be updated on the basis of the properties of the sensory object it was matched with. We gave a more detailed description of this case in the previous paragraphs describing sensory updates.

Now let us return to the example case. In the above example, we notice that the measured properties of ObjV1 are pretty near the sensory expectation of the properties for Obj14 - thus, we would expect a successful match and a resulting UPDATE of the properties of Obj14 through the information of ObjV1. On the other hand, ObjV2 has properties that are quite different than those expected for Obj22 and Obj23 - thus, as a match cannot be found, a new object will be created within the situation model (with a new unique ID given to it - say Obj24) in order to account for the sensed object (CREATE). Finally, the expectations for both Obj22 and Obj23 were not fulfilled - thus, both of these objects will be destroyed (DESTROY).

Now let us revisit the matching metric and algorithm question. There can be two variants to the metric: first, a simplistic case (continuous-layer based), measuring the match between the sensory object properties and the categorical layer properties of the GSM object under consideration, and second, a stochastic-layer based case, where the match between the sensory object properties and the stochastic layer properties of the GSM object is measured. In the first case, we match a set of vectors to a set of vectors of

[18] An "open world" assumption underlies this choice.

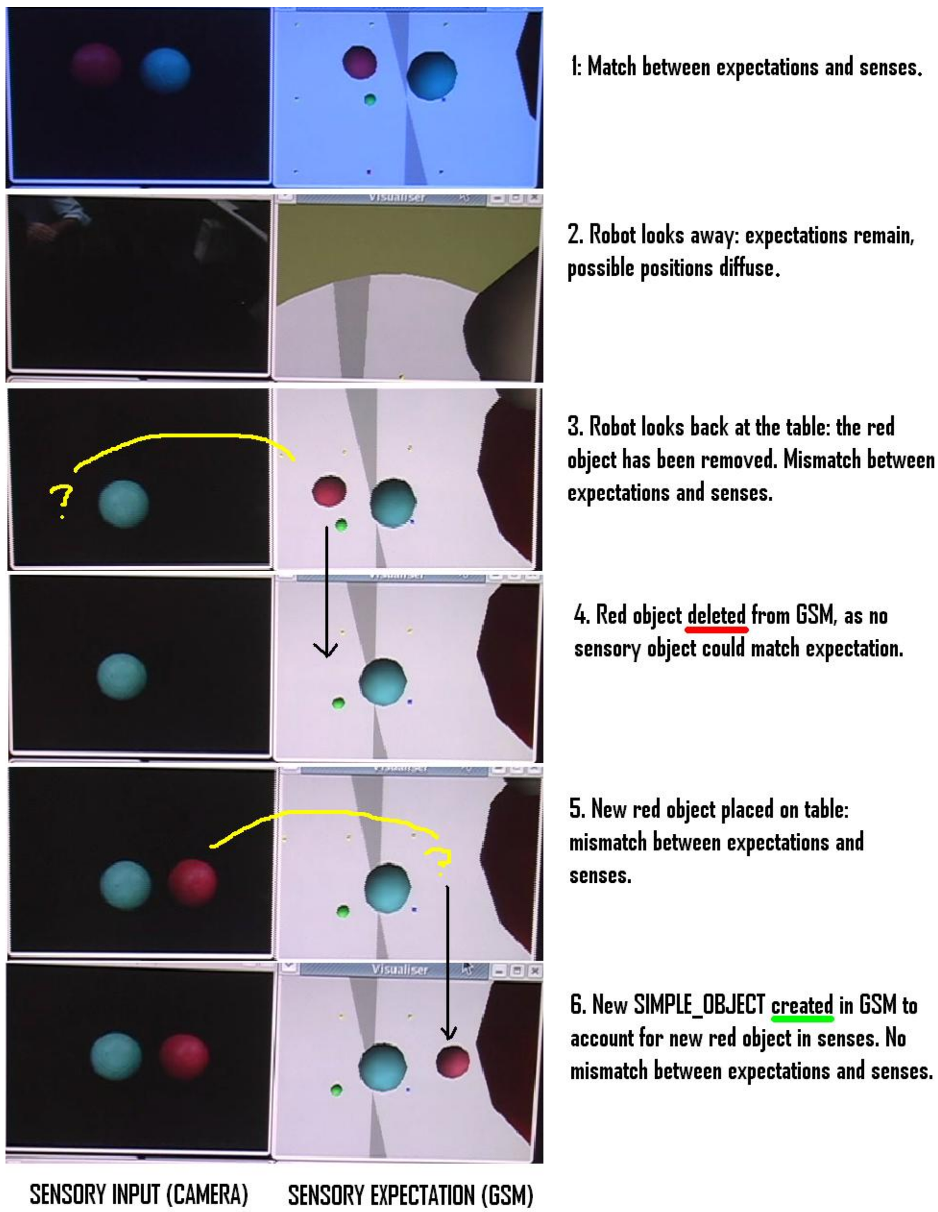


Figure 5-15: Matching sensory objects (sensory evidence) to GSM objects (sensory expectations)

equal number and dimensions - and thus, a weighted distance norm can be used, such as[19]:

$$d(ObjV_i, Obj_j) = w_1 d_{pos}(ObjV_i, Obj_j) + w_2 d_{col}(ObjV_i, Obj_j) + w_3 d_{rad}(ObjV_i, Obj_j)$$
(where pos=position, col=color, rad=radius)

In the second case, we match a set of vectors to a set of probability distributions of equal number and dimensions. Thus, we assign[20]:

$$spd(ObjV_i = Obj_j) =$$
$$pd(pos = sensedposition)pd(col = sensedcolor)pd(rad = sensedradius)$$
where pd= prob. density

In practice, this expression is calculated up to a scalar multiplier, by taking the probability of the bin of the position expectation histogram (stochastic layer) within which the measured position value falls, and similarily for color, radius etc. Of course, in both cases, an arbitrary threshold for successful matching needs to be chosen - maximum distance for match or minimum scaled probability density.

Having discussed the matching metric, let us now move on to the algorithm question. As the number of sensory objects within the viewport (and the corresponding expected objects in the GSM) is small, we employ an exhaustive search through all possible matches, as described in ([Roy et al.2004]).

Now, having seen the basic GSM processes in detail, let us move to the GSM processes that are related to temporal model construction - i.e. to deriving and accessing moments and events.

5.2.3 Processes for temporal model construction

As described before in section , the temporal model of the GSM consists of $MOMENTs$ and $EVENTs$.

$MOMENTs$ are created in a straightforward manner. The current GSM state is copied and time-indexed. In the current implementation, $MOMENTs$ are stored forever. For round-the-clock operation, some form of memory filter / consolidation must be added, but this has not been explored yet.

$EVENTs$ are created and continuously updated based on the current and previous moments, through "black boxes" processes that we term "event classifiers"[21]. These black boxes are continuously fed with the current $MOMENTs$, and paste $EVENTs$ to an event list. $EVENTs$ might be instantaneous or optionally encode duration. For example, when velocity (approximated by positional differences) rises above a preset threshold, it triggers the creation of the instantaneous "start moving" event. In contrast, an $EVENT$ having duration is first created, and then its end time is continuously updated as long as the event holds (e.g., the "is moving" event has duration equal to the period

[19] Also look at [Roy et al.2004].

[20] This is just pd(sensed property values|stochastic layer contents).

[21] See also section 8.3.1.

that an object is observed in motion). The "event classifiers" are again very simple in this first prototype. However, they are plug-in replaceable by more complicated trainable classifiers, utilizing hidden markov models, stochastic context free grammars etc.

Now, having discusses the basic GSM representational structures and processes, let us move on to the introduction of the proposed GSM modular implementation architecture.

5.3 GSM modular implementation architecture

A modular architecture running on multiple CPU's in which the GSM resides in a central module has been implemented 5-16).

The software implementation of the GSM and its associated algorithms is organized around a set of modules (As seen in figure 5-16):

Situation Model:

The module holding the current state of the GSM. This module broadcasts its contents to other modules over network connections, and processes requests for object creation / deletion / updates from the modality-specific modules in order to maintain the GSM object hierarchy.[22].

Visor, Proprioceptor, Imaginer (modality-specific modules):

Each of these modules propose changes to the current GSM state, which is broadcast from the Situation Model. Visor listens to the visual stream, while Proprioceptor connects to the robot's position and force encoders. Imaginer processes linguistic descriptions about real or imaginary situations. Via the imaginer, the situation model can now be fed not only through the senses but also through linguistic descriptions, and be later updated by either.

Inquirer:

Provides the capability of answering simple questions about the present, such as "What color are the objects at the left?", and also of acting on objects described through the present: "Touch the blue one". Carries out object referent resolution, and requests appropriate actions.

Rememberer:

Through this module, the past becomes accessible. It uses the "event" lists in order to resolve temporal referents such as "when the red one appeared" etc. Then, and after

[22]An important design question here is concerned with the possibility of conflicting information regarding an object entering through different modalities: for example, I heard that there is a ball, but I cannot see it; more on how we deal with this problem in the next chapter 6. More complicated treatments could be possible future extensions.

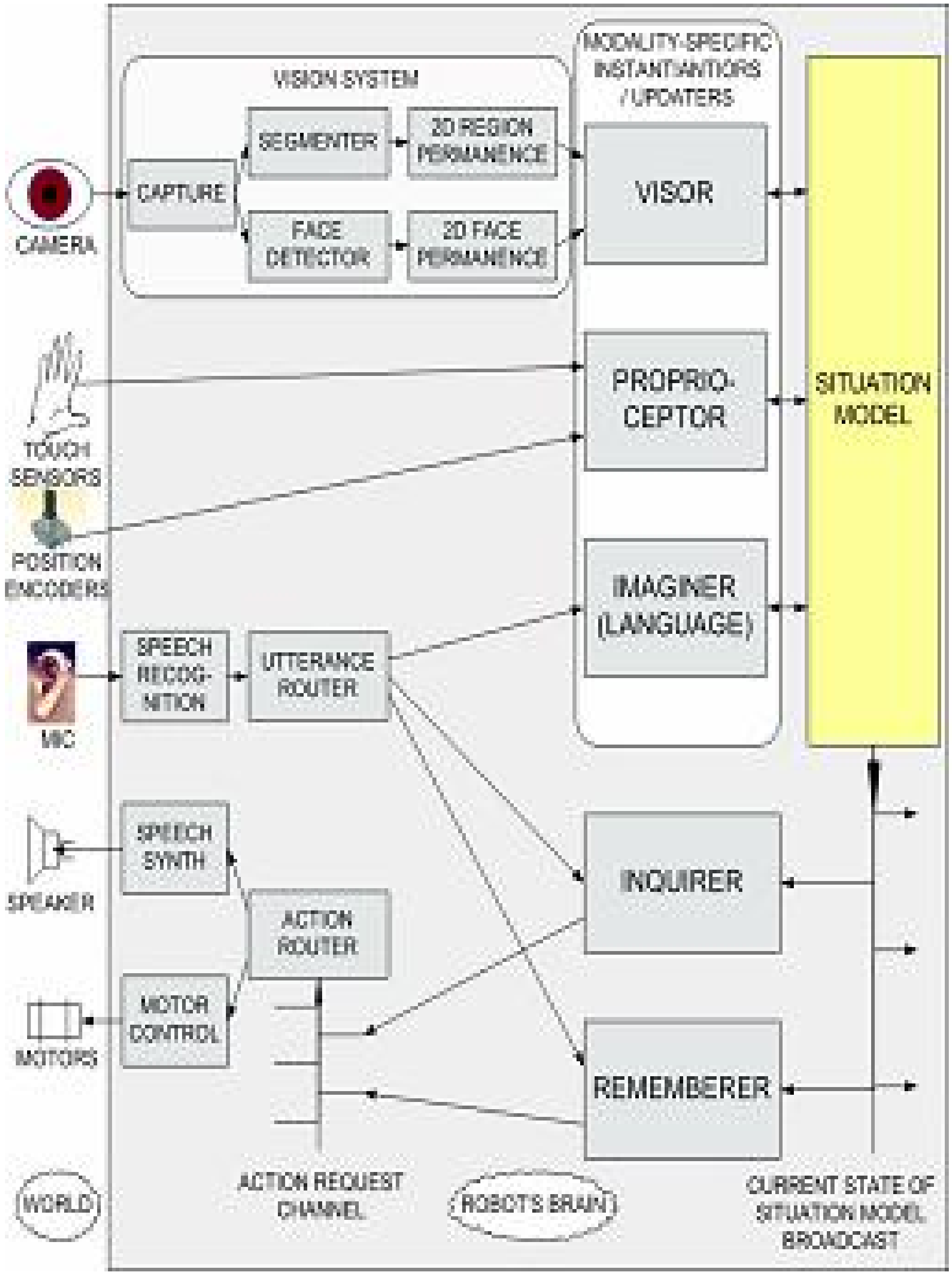


Figure 5-16: Proposed modular architecture

also having resolved the object referents at the right times, it feeds the appropriate "moments".

Utterance Router:

This module classifies incoming utterances, and forwards them to an appropriate module for servicing: Imaginer, Inquirer or Rememberer. INFORM speech acts ("there is a ...") are sent to the Imaginer module, REQUESTS for action or answers regarding the present are sent to the Inquirer, while REQUESTS involving the past are sent to the Rememberer.

Action Router:

This module routes requests for actions (generated by the Inquirer or Rememberer) to either the speech synthesizer (in the case of speech actions) or the motor controller (in the case of motor actions).

Notice here, that the primary data structure that is exchanged among modules is the present state of the situation (a $COMPOUND\_AGENT$ data structure, in the terminology introduced before in section 5.1). Changes to this are proposed by the various sensory-specific modules (visor, imaginer etc.), which then drive both language and motor actions (through the inquirer and the rememberer). $MOMENTs$ and $EVENTs$ are only held in the Rememberer.

Among the other modules, various types of messages are passed. Proprietary messages[23] supply information from vision, touch sensors, proprioception, while simple utterance events feed the utterance router, and action requests feed the action router. More specific details will be given at the next chapter (6), when we leave the level of the overall modular implementation architecture, and enter the specifics of the implementation in "Ripley the Robot".

5.4 Recap

Having discussed the overall agent architecture within which GSM-based systems reside in chapter 3, we then motivated and introduced GSMs, and derived elements of an overall GSM architecture from two explicit desiderata in chapter 4. There, we first introduced the breakdown of situations in agents, composite objects, simple objects, properties, as well as the three-layered descriptions (stochastic, categorical, continuous), without getting in further details.

In this chapter, we discussed the specifics of the GSM proposal: the proposed contents of the GSM representation (introducing data types such as $COMPOUND\_AGENT$, $EVENT$ etc.), the details of the associated processes that operate on it and move data in and out of it (introducing processes such as (senses-to-L1), (L1-to-L2) etc.), as well as a

[23] Left unspecified in this generic chapter, at the level of description entertained here.

multi-module architecture for implementing GSM's through intercommunicating processes in real-world systems (introducing modules such as "imaginer", "action router" etc.).

In the next chapter, we will see how all of the above were customized on the basis of a behavioral specification, and implemented in a real-world system: Ripley the Robot.

Chapter 6

The implemented real-world system: Ripley the Robot

An operational implementation of the GSM proposal, using the platform of the conversational robot Ripley has been built, and demonstrated numerous times in Media Lab open houses and to other visitors. This system was designed on the basis of a behavioral specification, which was later extended. Given the embodiment/hardware of the robot, the specifics of the GSM proposal described in the previous chapter, were further customized in order to meet the behavioral specification. The robot's language comprehension abilities that were achieved are comparable to those implied by a standard and widely used test of children's language comprehension (the Token Test for Children), and in some directions also surpass those abilities: for example, the robot is able to respond to references to past experiences and also to be informed about parts of the situation through language.

In this chapter, we will start in section 6.1 by describing the embodiment and hardware of Ripley, then move on in 6.2 to the behavioral specification on the basis of which its GSM was designed. In section 6.3 we will see the customized representations that were coded for Ripley, as well as the processes and modular implementation architecture in 6.4. These representations, processes as well as the architecture were based on the specifics of the GSM proposal that were described in chapter 5, and were customized in order to meet the behavioral spec. Finally, in 6.5 we will discuss how the implemented system compares with other existing conversational robots, and in 6.5.1 point out three novel achievements of Ripley as compared to all of the other systems.

6.1 Embodiment and hardware

He we start by describing the physical: the embodiment that we were given. The interested reader might ask: "How dependent is the approach and the implementation to the specific embodiment that you had? What must change if the embodiment changes?". An answer to this question is attempted in 8.5.1. Let us now proceed to the case of our implementation on the robot Ripley, which we will describe in this chapter.

"Ripley the Robot", is a manipulator robot under development in our lab. Apart from

Figure 6-1: The robot, the author, and the table top with the objects that are manipulated.

numerous experiments that have used it as a platform in previous years[1], we also anticipate it will continue serving as a basis for developing more sophisticated linguistic abilities in the future. The robot's environment consists of a table populated by objects, and a human interacting with the robot and the objects, who is standing near the edge of the table. (figure 6-1). The robot's originally intended purpose was to serve as a "conversational helping hand".

The manipulator robot is an arm with 7 degrees of freedom, equipped with: force feedback actuators (a), a gripper with force-sensitive touch sensors integrated into each finger tip (b), joint angle encoders (c), dual cameras mounted around the gripper (d), speakers and a microphone (e). Let us consider each of these components in turn:

(a) The *force feedback actuators* enable the robot to enter a special gravity-cancellation mode (zero-grav), under which it becomes virtually weightless, and thus very easy for a human to manipulate and move around. This mode is used for recording trajectories that have been used for action learning. The robot learns in a way similar to taking the "baby's hand" in your hands and moving it around, in order to demonstrate a move. This method of motor demonstration is advantageous over vision-based imitation motor learning, due to its simplicity - no complex human body / arm tracking through vision is required. Furthermore, the force encoders enable the robot to "weigh" objects, a feature that has been used in the implemented system.

(b) The *gripper* is able to grasp / handle / put down simple objects of roughly 3-6 inch diameter, and the touch sensors can successfully indicate grasp success / failure, but unfortunately have little useful resolution inbetween the two extrema, that might have been useful for hardness/softness measurements.

(c) The *angle encoders* provide proprioception for the robot (knowing in what configuration your body is). They require re-calibration through a standing reference position every time the system is reset.

(d) The *dual miniature ELMO cameras* are used for mono or stereo vision, and pro-

[1] [Hsiao et al.2003], [Roy et al.2003] [Roy et al.2004], [Mavridis and Roy2006a] etc.).

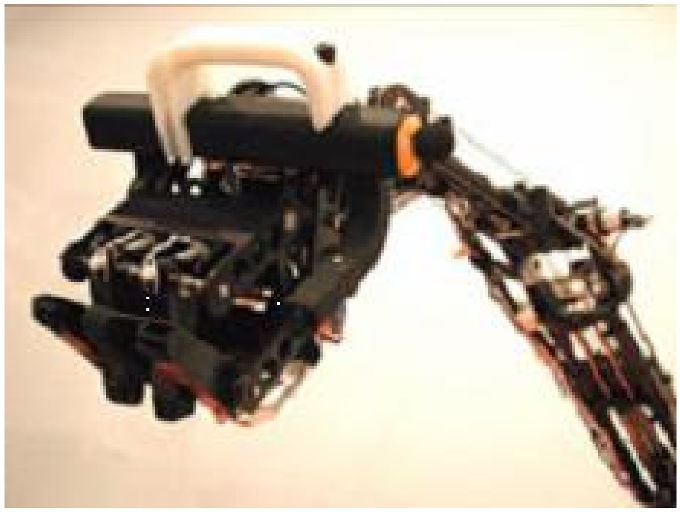


Figure 6-2: Ripley's head - manipulation handle, gripper, cameras mounted at the sides of gripper

vide analog signals which are later digitized. Their mechanical fixture unfortunately does not allow precise calibration or positioning.

(e) A *set of speakers* hidden behind the robot provide speech output, and a *wireless microphone* worn by the human user provides speech input.

In terms of reachability in the operating space of the robot, although a big part of the space is reachable in terms of positioning the head in the point under consideration, this is not the case for arbitrary pose/position configurations, as was reported in the investigation carried out in [Mavridis2004a]. As can be seen in the figures of this investigation, only a small part of the operating space is reachable for a given pose. This creates problems in taking multiple views around an object, in attempted multi-view geometry reconstruction experiments. More on multi-view reconstruction in section 9.1.1.

The robot's software, consisting of multiple modules along the overall modular architectures described in 5.3, is currently distributed on 8 PCs plus the robot's onboard controller, all of which are interconnected through gigabit Ethernet on a single 12-port router. More details on the software in section 6.4.

6.2 Behavioral Specification

There were two historical stages in the behavioral specification of our specification. There was an initial part first, which was implemented - i.e. the specifics of the GSM proposal that were described in the previous chapter, were customized in order to meet the behavioral spec and the embodiment. Later the specification was extended, and the system was consequently extended in order to meet the new requirements.

6.2.1 Initial Behavioral Spec

We set the following behavioral goals for the system to be implemented:

Figure 6-3: Human hand-related event detection

B1) *Answering questions about the present physical context:* For example: "What color is the one on the left?" (a question about objects on the table).

B2) *Quantifying and expressing confidence in beliefs:* Thus, when asked about the location of an object that it hasn't seen for a while, it might answer, "Probably at the left, but maybe not", expressing uncertainty since the object might have been moved while the robot was not looking.

B3) *Respond to spoken requests:* For example: "Look to the left" or "Hand me the small red one", for which the robot should respond with situationally appropriate motor actions.

B4) *Imagining situations described through language:* This covers the understanding of commands such as "Imagine an object at the left", or descriptions such as "There is a small object at the right". Such speech acts must be translated into representations that may later be related to sensory input.

B5) *Remembering and resolving temporal referents:* The robot should keep track of salient past events and be able to talk about them. The robot should be able the robot to answer questions such as "What color was the object that was at the center of the table when the red one appeared?".

Indeed, as we will see in the next sections, our implemented system was able to meet and surpass the above behavioral goals that were set.

6.2.2 Extended behavioral Spec

We later augmented the original behavioral spec given in the previous section with the following additional behavioral goals:

B6) *Resolving indexical human hand pointing:* so that the robot can understand indexical expressions such as "what color is this one" which consist of words plus finger pointing.

B7) *Recognizing human hand-related event detection:* so that the robot can answer questions such as: "Where was the blue one when my hand started touching the red one?"

Notice here that B7) is of utmost important towards action coordination with humans, when cooperating towards a common goal - for example, imagine a construction task. In that case, it is not enough for the robot to be able to perform actions on objects itself ("pick up", "put down" etc.), but it also should be able to detect the performance of these actions by humans (when the human has performed a "pick up", "put down" etc.), so that it can coordinate its actions. In general, for a robot to cooperate with humans, it is not enough to be able to manipulate objects itself - it should also be able to

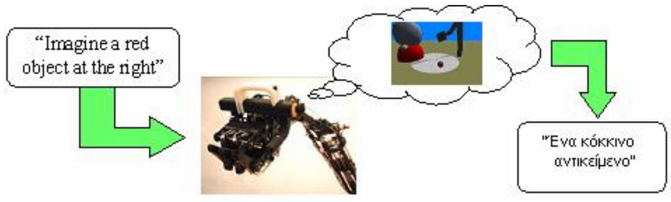


Figure 6-4: Rudimentary translation - using the situation model as an interlingua

recognize what humans are doing to objects. This capability, apart from being essential for turn taking in such tasks, is also very useful for learning of construction procedures by imitation - not at the level of minute moves, but at the level of primitive actions on objects ("pick up") etc.

B8) *Extending speech interface to multiple languages:* Here, the goal is for the robot to be able to achieve all the above functionality not only in English, but also in other languages such as Greek, and to also be able to switch between languages in real-time, depending on the language it is addressed with.

Notice here that due to B8) an extremely basic demo of another capability of GSM's can be demonstrated: situation models acting as an interlingua for translating between languages. For example, one can give an INFORM statement in English: "There is a red object at the right", and then ask about information about the object in Greek: "Ti xrwma einai to antikeimeno?" (i.e. "What color is the object") and get an answer in Greek: "Einai kokkino" ("It is red"). This is an extremely rudimentary example of translation, but with future extensions towards story understanding (inform statements describing sequences of events - see discussion in section 9.1.4), the capabilities of translation might become significant.

Again, as we will see, the above goals that were set were met[2]

6.3 Representations

The creation/updating of the above representations will be discussed in the next sections, i.e. processes and modular implementation architecture. Here some details of the data structures will be given.

6.3.1 Body Parts ($SIMPLE\_OBJECT$ data type)

Each $SIMPLE\_OBJECT$ has a uniquely generated ID. It also contains the following $PROPERTY$ instances that were encoded in triads (stochastic, continuous, categorical) within $SIMPLE\_OBJECT$:

P1 - Position (3D): Object center position
P2 - Pose (3D): Object pose in scaled quaternions, convertible to matrices / angles

[2] However, no speech recognition for Greek was setup - the multilingual capabilities are limited to console text input. Greek speech is also fed to an english synthesizer, creating a funny-sounding phonemisation. Anyway, these extensions would be straightforward in the future.

Figure 6-5: The instantaneous contents of a position property triplet for an object located near the center of the table: stochastic, continuous, categorical.

P3 - ShapeParams(3D): Object shape params such as radius for spheres etc.
P4 - Weight (1D): As acquired from the force sensitive sensors
P5 - Color (3D): Simple RGB-space color

An augmented set with some further experimental $PROPERTY$ instances (for work under progress) was also used for some demos:

P6 - 3D voxelised occupancy grid: for 3D model acquisition[3]
P7 - Linear Velocity (3D) - for experim. with dynamics/prediction through physics sims
P8 - Angular Velocity (3D) - as above
P9 - Force Accumulator (3D) - as above
P10 - Torque Accumulator (3D) - as above

6.3.2 Situation ($COMPOUND\_AGENT$ data type)

The situation was modeled as consisting of four agents (robot, human, table, objects):

- Robot Model ($AGENT$ No.1):
  Body ($COMPOUND\_OBJECT$): Capped cylinders leading to a rectangular box.
  Interface: Viewpoint and viewdirection of two eyes. Ears assumed omnipotent.
  Mental: Equal to GSM (assumption: robot believes it believes all that it believes, i.e. all beliefs are consciously penetrable for robot)
- Human Model ($AGENT$ No.2):
  Body ($COMPOUND\_OBJECT$): Capped cylinders for torso and hand, sphere for head
  Interface: Viewpoint and viewdirection, ears omnipotent.
  Mental: Either I) Equal to GSM (assumption: robot believes human believes what robot believes), or II) Fed through synthetic vision from Robot's GSM through human viewpoint (assumption: robot believes that the human believes what is/was accessible from human's viewpoint) - used in some experiments.
- Table model ($AGENT$ No.3):
  Body ($COMPOUND\_OBJECT$): cylinder with very small length
  Interface: N/A
  Mental: N/A

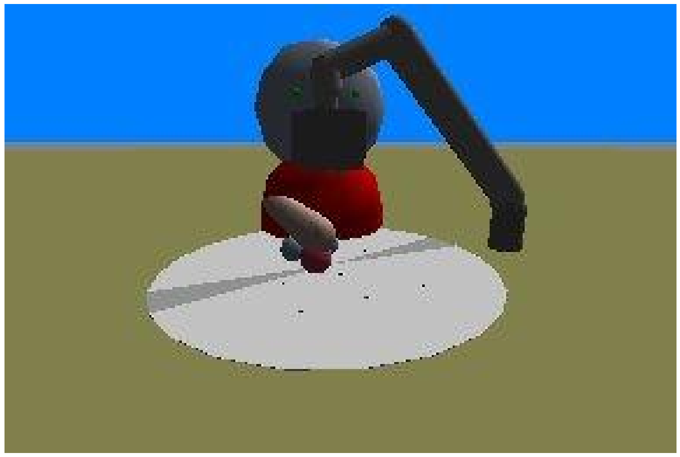


Figure 6-6: the most-likely-state visualization (continuous layer) of a situation - Robot model looking down on table model which is populated with two object models, while the human model is touching one of the object models with his hand

- Model of Objects on table ($AGENT$ No.4):
  Body ($COMPOUND\_OBJECT$): spheres
  Interface: N/A
  Mental: N/A

Relative properties among objects ($AGENT\_RELATION$):
Used for experiments with spatial relations among triads, as discussed in 8.3.3.

6.3.3 History ($HISTORY$ data type)

A history consists of $MOMENTs$ (time-indexed compound agents) and $EVENTs$. The following eventtypes are supported[4]:
Creation/existence/destruction,
Coming in view/remaining inview/getting out of view,
Movement start/movement continues/movement stops,
Move up starts/move up continues/move up stops (for pickup detection)
Move down starts/move down continues/move down stops (for putdown detection)
Start touch/is touching/stop touch
Start pickup/is picking up/stop pick up
Start putdown/is putting down/stop put down

In figure 6-7, you can see a real fragment from an event list of a history gathered during Ripley's operation.

6.3.4 Messages outside situation model

Proprietary - some described within next section.

[4]For the corresponding verbalisations, have a look at appendix A.

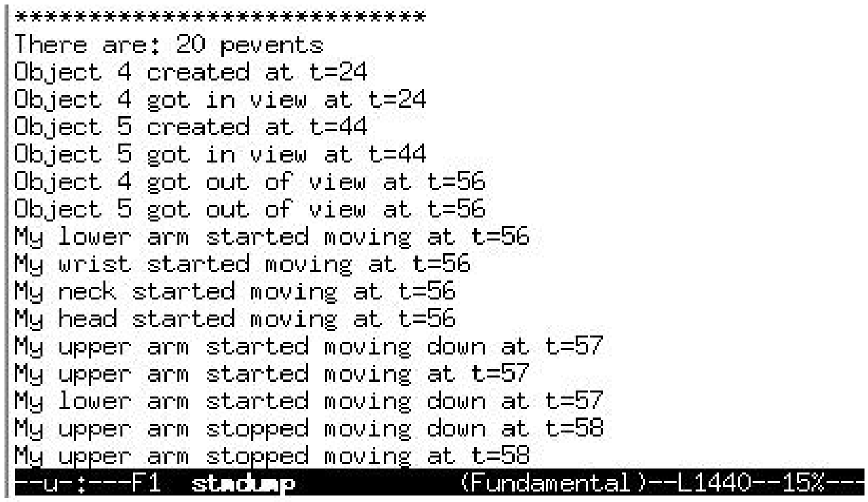


Figure 6-7: Description of a fragment from an $EVENT$ list of a $HISTORY$

6.4 Processes and Implementation Modules

Here, I will unite the previously separated sections on processes and implementation modules, for ease of presentation. All software is running under Linux, with most modules written in C++ (some Java too). Libraries that were used in the code include OpenCV (computer vision) and OpenGL (graphics). The modules intercommunicate with each other through the PVM protocol [PVM]. A wrapper on top of pvm was written in order to provide channel naming etc.

Zooming in the block boxes and exposing the details of the current system:

6.4.1 Vision front-end

A long push-forward architecture, with three parallel streams (objects, faces, hands) is used for vision, and it occupies one dual Xeon rack machine per camera. A short description of the vision system is: (look at figure 6-8)

Two camera capture modules (Left/Right), each feeding an object segmenter, a skin segmenter, and a face detector, each of which followed by an objecter 2D. The outputs of the skin segmenter and the object segmenter fed into two stereobjecters. All outputs as well as stereobjecter outputs fed to visor.

Capture, segmentation, facedetect:

The analog composite video signal is fed through conditioning boxes to firewire video digitizers, which drive the capture module. Hand and object detection is performed using trainable color probability-based foreground/background segmenters that decide on a 5x5 pixel block basis, and only report regions fitting area / perimeter / inclusion of center criteria. The output of these modules is two streams for each: one containing pixel-by-pixel regions, the other fitted ellipsoids with an average color attached to them.

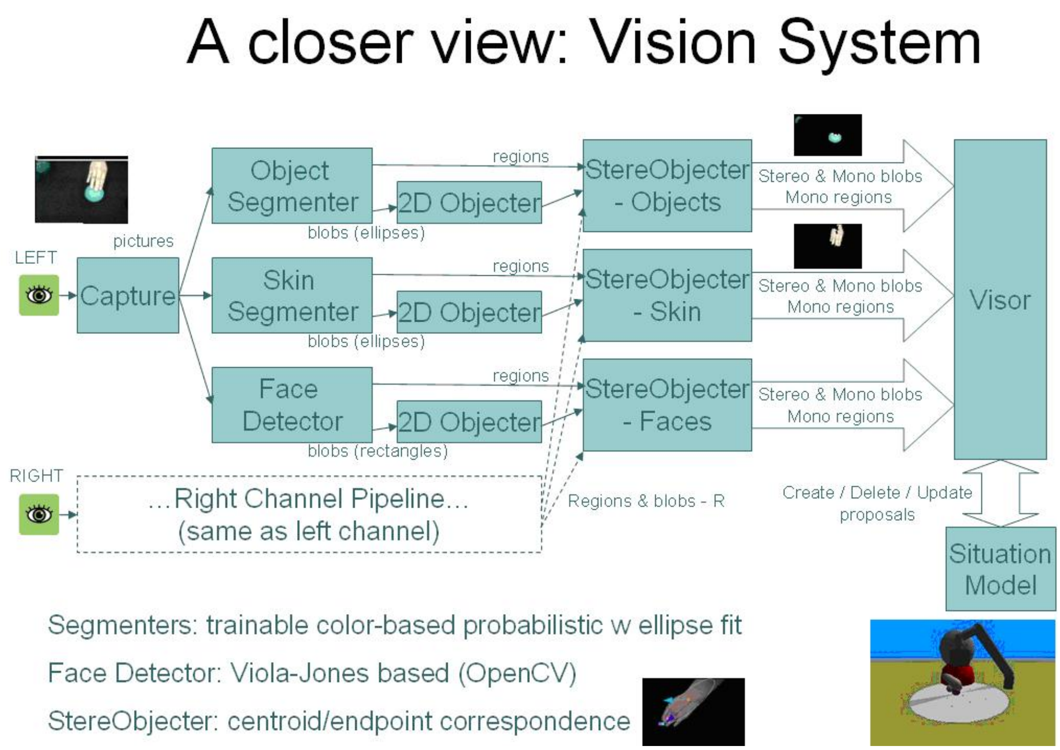


Figure 6-8: Block Diagram of vision system

Face detection is performed using OpenCV-derived code [OpenCV], which is based on a cascade of classifiers method.

2D objectors:

The three streams are then fed to 2D region-permanence / evidence accumulation modules, which are termed "objectors". These report regions in their output only if the incoming regions have accumulated enough evidence for existence through their persistence across frames, and also assigns unique IDs for regions reported across frames, by performing a matching between regions in two consecutive frames (i.e. answering the question: Does this particular region of the current frame correspond to any previously observed regions? If yes, it will continue having their ID, if not, it will be assigned a new unused ID).[5]

Stereobjecters:

Finally, the output streams of the objectors are passed to the "stereobjecters". These modules perform matching of regions across cameras (not across frames as was the case in the 2D objectrs). They also calculate the robot-head-relative 3D depth of the region seen, and in the case of the stereobjecter for hand regions the approximate fingertip location of the hand region is also calculated[6].

[5]For more information, look at [Roy et al.2004].
[6]More details can be found in [Cheng2006].

6.4.2 Visor

This module receives visual evidence from the 2D objecters / stereobjecters, and also receives the current GSM state. On the basis of the visual evidence it proposes object creations, deletions, or updates in the GSM, or hand / face updates. Various versions of the visor exist, depending on which information is used. Originally, when the vision system was mono (and not stereo), two simplifying assumptions were used for positioning:

a) Each object is in contact with the table, unless it is within the gripper (mouth of robot).

b) Each face is positioned on a cylinder that has its perimeter along the edges of the robot's circular table.

These assumptions (constraints) enabled 3D positioning given mono information. In the current system, only assumption b) is used, hand positioning uses full stereo information, and whether a) or stereo information will be used for objects is selectable. Internally, various algorithms occupy the visor's code. For example, a matching process matches visual evidence with GSM objects and proposed deletions, updates, creations (described in 5.2.2). Also, object unprojection/projection is performed using geometrical libraries etc.

3D voxelized shape models through multiple views

In the first stages of the development of Ripley, the visually-sensed objects on the table of Ripley were assumed to be spherical, and the shapes of the ripley and human body parts are apriori given. However, some demonstrations of representing richer shapes through multiple views have already been implemented by the author and demostrated on Ripley. First, as commented upon in section 6.3, the $SIMPLE\_OBJECT$ type has been augmented with the following experimental property:[7]

P6 - 3D voxelised occupancy grid: for 3D model acquisition

The stochastic layer of this property corresponds to a 3D occupancy grid, with each voxel of the grid corresponding to a bin in the 3D histogram. The object is initially thought of as occupying uniformly the whole limits of the histogram, and then, each subsequent view of the object from a different angle, effectively "cuts out" voxels. This is done in the following way: for each view, the object segmenter regions (and not blobs) output is utilized[8]. These regions consist of a set of pixels, which are belonging to the object under examination. Due to projective geometry, each of these pixels can be thought of as belonging to a ray starting at the robot's camera and extending infinitely. All voxels of the occupancy grid that fall within one of the rays are kept, and others discarded. Thus, the first view will carve out a "modulated cone" from the originally fully occupied rectangular occupancy grid. The next view will provide a second "modulated cone", starting from a different viewpoint; this will effectively be intersected with the previous cone on the occupancy grid, and so on. In this way, the final occupancy grid for the object (its voxelized shape representation) will consist of whatever remains from the

[7]Thus, for every object, there exist two parallel models: both a spherical blob model and a voxelized occupancy grid.

[8]In these experiments, only one vision channel was used, thus we have one regionmap per view.

(a) (b)

Figure 6-9: Shape reconstruction from multiple views leading to a voxelized shape model

consecutive intersections of the cones in the occupancy grid after multiple viewpoints are taken. See figures 6-9(a) and 6-9(b) for an example of the real-world performance of the implemented system, taken from the video [Mavridis2006b]. More details can be found in [Mavridis2004b].

Although the system is operational, and 3D voxelized shape models can be acquired and placed in the GSM as seen in figures 6-9(a) and 6-9(b) and the video cited above, these shape models have not been tied to linguistic descriptions through appropriate categorical classifiers yet; also there are many other possibilities for enhancement and extension, which are discussed in section 9.1.1.

6.4.3 Proprioceptor

The proprioceptor updates the robot model within the GSM on the basis of angle encoder information, coming from the robot. A forward-kinematics model is used for that purpose.

6.4.4 Speech Recognizer

We use the Sphinx 4 continuous speech recognizer [Lamere et al.2003] to convert incoming speech into text transcripts. A trigram-based language model is used, which effectively enforces some local syntactic requirements in the following way: In the modules where the utterances are serviced (imaginer, inquirer, rememberer), keyword-based semantic frames are used to parse speech transcripts. These have no syntactic requirements - strings are treated as simple bags of words, and word order is not taken into account at that stage[9]. Because of the simplicity of the serviced language (look at appendix A), this is a viable solution. But then, how is locally correct syntax enforced, even if almost no syntactic information is taken into account? The syntax is actually enforced through the specially-trained trigrams at the front-end of the speech pipeline (the speech recognizer). I.e. "give me the red ball" passes through the speech recognizer, but "me ball red the give" does not, as "me ball red" has zero probability in the language model, while "give me the" has non-zero probability.

[9]With the exception of the extraction of temporal referents, where the utterance is broken in two parts).

6.4.5 Utterance router

After passing through the recognizer, utterances are then classified in terms of their speech act type: REQUEST speech acts such as questions ("Where is ...", "What color is ...", etc.) or action requests ("Touch ...", "Look at ...", etc.), INFORM speech acts providing information about the situation ("There is ...") etc. Tense information is also extracted[10]. REQUEST speech acts in the present are routed to the Inquirer module, REQUEST speech acts in the past are routed to the Rememberer, and INFORM speech acts are routed to the Imaginer. For multilingual support, a simple majority-of-words based language classifier is used, that switches between vocabularies. The same language classifier is used at the text front-end of *Imaginer*, *Inquirer* and *Rememberer*.

6.4.6 Imaginer

The imaginer has two functions:

a) Create/update objects on the basis of linguistic information, i.e. service statements of the form: "there is a small green object at the left" with the creation/update of objects within the GSM.

b) Ex-nihilo create the pre-existing objects that exist at the "beginning of time" for the robot: i.e. create the robot's model (which will be updated later when sensory information arrives through proprioceptor etc., create the human model (again to be updated later), as well as the table model.[11].

For the purpose of a), the basic GSM processes of subcycle 1.3 described in section 5.2.1 are used. There are two points to notice here: the first has to do with what we mean by "imagination" when it comes to the current implementation of the Imaginer module[12], and the second has to do with our choice of including ex-nihilo creation of objects as part of the Imaginer module. Both of these points are discussed in further detail in section 8.2.4

Data fusion and sensory confidences

At this stage, having introduced three possible sources of object information (vision, touch, language), one must deal with the following question (mentioned before in section 5.3): What happens if the information coming from different sources is conflicting? The approach taken here is simple, and is adequate due to the simplicity of the overall system and tasks. If one hears about an object ("there is a small red one on the left"), and later his sensory expectation does not verify what was heard (looks at the left and sees a blue one or does not see anything), then the eyes have precedence over the ears; and furthermore, if the seen object differs significantly (matching metric) from the expected (i.e. a large blue one seen but a small red one expected) then the "imagined" objected

[10] Indirectly, through the existence of temporal referents such as ("...when your head started ...").

[11] All of these ex-nihilo created models are not created on the basis of linguistic input in the current implementation - see discussion in section 8.2.4.

[12] In the current implementation of Ripley, "imagine ..." has the same semantics as "there is ...", but extension is straightforward - see section 8.2.4 and 9.1.4.

is deleted, and a new object (with different uniquely identifying ID) is created. Other, more complicated solutions are possible as future extensions.[13]

6.4.7 Inquirer

This is the module servicing questions in the present ("where is the red one?") or motor action requests in the present ("hand me the small blue object"). Here, the main processes taking place are request type classification (on the basis of wh-words / action verbs), object referent resolution, (i.e. going from "the small red one" to the ID of an object populating the GSM") and request servicing. The other type of referent resolution, namely temporal referent resolution (i.e. going from "when your head started moving" to a time index in the history), is not needed here and takes place in the Rememberer.

Object reference resolution:

Reference to an object can be resolved to any part of the three main agents of the situation model: me (robot), you (human partner) and others (objects on the table). It might be resolved to one, many, or no such parts. It might be referred to either through "part names" (my head, your arm) or through "definite descriptions" (the small red one, the large ones at the top), or through hand pointing (this one). The simple objects (body parts) of the robot and the user are usually referred to by part names, while the objects on the table (others), are referred through attributive descriptions. Consider the question "Where was the blue object when your head started moving?". In this case, both part names ("your head") as well as attributive descriptions ("blue object") are used, one for each object referent. The robot might either ask a disambiguating question (supplemented with deictic pointing by the robot) until it narrows down to a single referent, or it might carry out the requested action in succession on all the referents fitting the description. The course of action taken depends on the action requested, on whether it can accept groups of objects as arguments, and also on whether plural or singular was used.

For example, assume that three objects are on the table - a small red sphere, a large red sphere, and a blue sphere. If the human requests "Touch the red one!", the robot will answer "do you mean this one or that one?" while pointing to the two red spheres in succession. Then, the human can narrow down by saying "Touch the small red one". Else, if the human had requested "Touch the red ones!" then the robot would touch both red spheres in succession. These behaviors are selected via a decision tree which is driven by the number of matching referents, the plural or singular number, and the possibility or not of carrying out the specified action with multiple referents. For more details, look at figure 6-10[14]:

[13] For example, assigning different continuously updated confidences to different sources - some humans are not so reliable in their verbal reports, whenever I hear something from them my vision does not verify it etc.

[14] A slight exception exists for the case of more than one referents for a "pickup" command, in which the robot says: "cannot do that", as it can only pickup one object at a time (only one object fits in the gripper, and without putting it down first no new objects can be picked up).

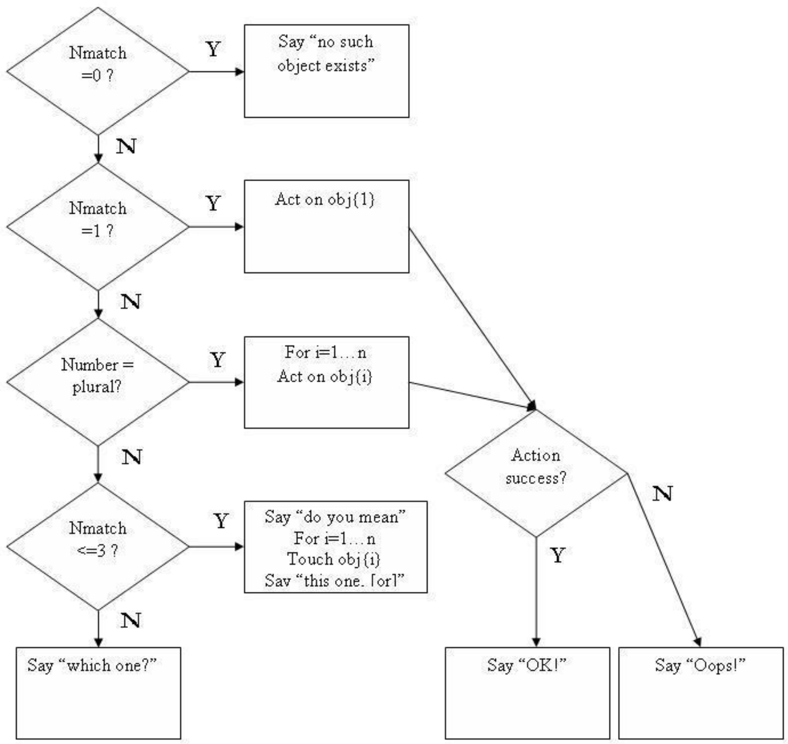


Figure 6-10: Command execution flowchart: Nmatch = number of object referents matching description, Number = grammatical number (singular / plural)

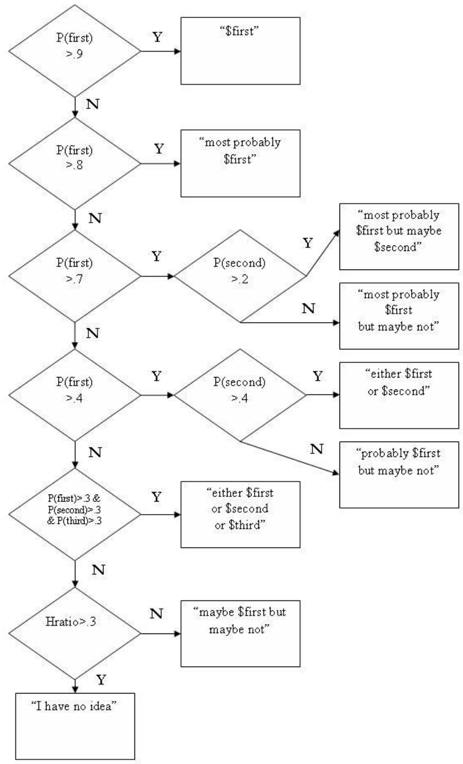


Figure 6-11: Decision tree for verbalisation of gradations of certainty

Verbalisation of gradations of certainty:

A decision tree, which accepts the categorical layer as input (i.e. the probability distribution on the verbal categories), produces the appropriate verbalisation. This decision tree can be seen in figure 6-11. In the case of a clear winning category, the verbalized output is just the category label, for example: "red". In the case of limited but existing uncertainty, the output becomes: "most probably red". In the case of two almost equally probably categories, an either/or verbalisation is used: "either red or green". Finally, in the case of highly spread out probabilities (evaluated through an entropy criterion), the verbalisation becomes: "I have no idea".[15]

6.4.8 Rememberer

Here, request type classification, event classification, object and temporal referent resolution, as well as request servicing takes place.

[15] Notice that the thresholds used in the decision tree could be in principle empirically tunable, through the observation of human verbalisations. Regarding other alternative verbalisations of gradations of certainty, also look at the comments in section 8.5.9.

Temporal reference resolution:

In the case of questions or actions involving the past, temporal references must also be resolved. Their existence is detected through the keyword "when". After "when", an event description involving object referents should follow. Consider the meaning of the phrase "when your head started moving". This amounts to going back in time until a matching event is found, and resolving to the time of this event. The verbalisations of the referable event classes can be found in the appendix A. The participants are either the objects, the user, or the robot itself. In the case multiple candidate events are found, only the most recent is reported. If the requested action is not a question, then one further condition should hold: the referred object should still exist, so that it can be acted upon.

6.4.9 Action Router

Speech action requests are routed to the Speech Synth module, Motor Action requests to the Motor Control module.[16]

6.4.10 Speech Synthesizer

The AT&T truespeech synthesizer is used, wrapped around a PVM-based custom interface. For details on PVM, look at [PVM].

6.4.11 Motor Control

The lowest-level software runs on the robot's embedded PC, and communicates with a dedicated IP link to higher-level[17]. The higher-level code, running on another PC, provides the action primitives listed in figure 6-12.

As no generic GoTo(position, pose) command was available, an inverse kinematics system was created, based on table-look up and optimization for solution refinement, documented in **??**. As commented upon before (in section 6.1) and as can be seen in the cited document, unfortunately the reachability of the robot in terms of (position, pose) pairs is severely limited.

6.5 Comparison to other existing conversational robots

Below, a short review of existing robots with conversational abilities is given. The approaches taken towards connecting language with perception and action will be briefly examined, as well as their behavioral repertoires. Directly afterwards, in the following

[16]Only when abstract action requests become used will this module become really important. Then, an abstract "refer to object" action, might either resolve to pointing or to speaking, through this module. See for example [Appelt1985].

[17]It was found that delays and instabilities of PVM deemed it unsuitable for such highly time-critical motor loops, and so low-level IP was used instead.

	Motor Action Primitive	Comments
A1	GoTo(position)	Pose = looking straight down
A2	GoTo(position,relativepose)	Head-relative pose
A3	Touch(position)	Descend, touch, and go back
A4	HandMe(objposition, humanposition)	Pick up and give
A5	PickUp(objposition)	
A6	PutDown(objposition)	
A7	Weigh(objposition)	Pick up, swing to weight, put down
A8	Relax	Enter zero-gravity mode
A9	HoldPoint	Exit zero-gravity mode & hold point

Figure 6-12: Supported Motor Action Primitives

subsection, the novel capabilities of Ripley as compared to the examined systems will be explicated.

In [Crangle and Suppes1994], the authors propose a natural-model semantics which they apply to the interpretation of robot commands, in two robotic aids for the disabled. As the above robots are not equipped with perceptual systems, a model of the environment consisting of 3D object positions and properties is entered manually into a knowledge base. Total confidence and complete knowledge is assumed. In [McGuire et al.2002], a Bayesian network interconnects visual to verbal information about objects. The system can interpret gestures, and includes visual attention mechanisms, but can only handle action requests. In [Sofge et al.2003], an occupancy map built by range sensor data plays part of the role of a GSM. Objects are individuated, and spatial relations are exploited in answering questions and interpreting action requests.

Cynthia Breazeal's Leonardo robot [Breazeal et al.2004], is her central experimental vehicle towards building humanoids that can act as cooperative partners for humans. Leonardo currently uses a cognitive architecture built on top of the c5 codebase, an extension of c4 [Burke et al.2001]. A centrally located "Belief system" module interconnects speech, vision and action. Hierarchical structures called "percept trees" classify sensory inputs to "snapshots" which are fed to the belief system, which decides whether to create or update beliefs about objects and their properties. Also, the system models human beliefs with representations having the same structure, in a similar way that we do using embedded GSMs. Furthermore, the system models attentional focus, which our system currently does not. However, our existing implementation on Ripley has three novel abilities compared to all of the above mentioned systems, as we will now see.

6.5.1 Novel capabilities of Ripley

Ripley has at least three types of novel capabilities as compared to the other conversational robots:

A) Imagining situations described through language:

The robot can understand commands such as "Imagine an object at the left", or descriptions such as "There is a small object at the right". Such speech acts are translated into representations that may later be related to sensory input.

B) Remembering and resolving temporal referents:

The robot can keep track of salient past events and talk about them. This enables the robot to answer questions such as "What color was the object that was at the center of the table when the red one appeared?" or respond to commands such as "Pick up the object that was at your left when your head started moving".

C) Quantifying and expressing confidence in beliefs:

When asked about the location of an object that the robot hasn't seen for a while, it might answer, "Probably at the left, but maybe not", expressing uncertainty since the object might have been moved while the robot was not looking. Similarily, when its verbal information does not enable specific categorical knowledge, it might respond to: "what color is the object on your left?" with "I have no idea".

The GSM-based design has been *instrumental in attaining these abilities*, and as will be discussed later (chapter 9), we think will also prove *instrumental towards many extensions* and provide significant leverage for such extensions on the way to truly cooperative Situated Conversational Assistants.

Furthermore, notice that all of the examined systems are far from being able to exhibit the behavioral repertoire required for passing a human psychological test such as the "Token Test", which is test situated language abilities and is normally administered to three year old children (to be introduced in the next chapter in section 1.1.1). In contrast, the GSM proposal can cope with the requirements of this test. First, the language comprehension abilities of the currently implemented system (Ripley the Robot), are comparable to those implied by the Token Test, and in some directions also surpass those abilities (as will be discussed in chapter 7). Furthermore, a theoretical GSM-based design that can pass all of the Token Test has been developed[18].

6.6 Recap

In this chapter, we described the operational implementation of the GSM proposal, embodied in the Robot Ripley. We started (in section 6.1) by describing the embodiment and hardware of Ripley, then moved on in 6.2 to the behavioral specification on the basis of which its GSM was designed. In 6.3 we saw the customized representations that were coded for Ripley, as well as the processes and modular implementation architecture in 6.4. These representations, processes as well as the architecture were based on the specifics of the GSM proposal that were described in chapter 5, and were customized

[18]Which remains to be implemented - see future extensions section 9.1.8, on the basis of the design method of appendix D.

in order to meet the behavioral spec. Finally, in 6.5 we discussed how the implemented system compares with other existing conversational robots, and in 6.5.1 we pointed out three novel achievements of Ripley as compared to all of the other systems. In the next chapter, we will discuss the topic of the evaluation of SCAs, and also see under what conditions psychological tests for humans might serve a useful purpose in that respect.

Chapter 7

Evaluating SCAs

In this chapter, we will consider the subject of the evaluation of SCAs in general and our implemented system (Ripley the Robot) in particular. Apart from general issues and proposals regarding the evaluation of such systems, this chapter contains both a *qualitative / descriptive evaluation* as well as a *quantitative evaluation* of the implemented system on Ripley the Robot, and also a *direct comparison of Ripley's current abilities to those implied by the Token Test*, a psychological test administered to human children. The descriptive evaluation is closely tied to the example operation videos of that system that have been published and are also cited here. The quantitative evaluation contains an error-analysis of real-world operation, performed at both the task- and module- levels.

We will first present a detailed example of the operation of the implemented system on Ripley the Robot in section 7.1. The video that corresponds to this example, as well as many other videos of Ripley, can be found at [Mavridis and Hsiao]. Then, in 7.2, we will discuss various different methods for evaluating SCA's in general. Later, in 7.3, we will present the quantitative evaluation that addresses both the task- and module-levels of evaluation. In 7.4, we will introduce the "Token Test" for children, a psychological test administered to children, and discuss its relation to our implemented system on Ripley. Later, in section 7.5, we will discuss the use of such tests for evaluation of SCA's as well as for acting as design specifications for SCA's. Most importantly, we will propose specific prerequisites (the three "commandments") for the meaningful use of such tests both for design and evaluation without "cheating".

7.1 A detailed example of operation of the implemented system

For the extent of the behavioral repertoire of the current system, the reader should consult appendix A. Here, we will illustrate aspects of the abilities of the system, through an example of operation, which is part of an accompanying video, as mentioned above. The accompanying video can be found at [Mavridis and Roy2006b]. The full text of this video can be found in section 1.1.3 of this thesis. Here, we will describe in detail part II of the text contained in that section, thus centering on the "inform" statement, and on how verbal information can produce sensory expectations within the GSM, which are

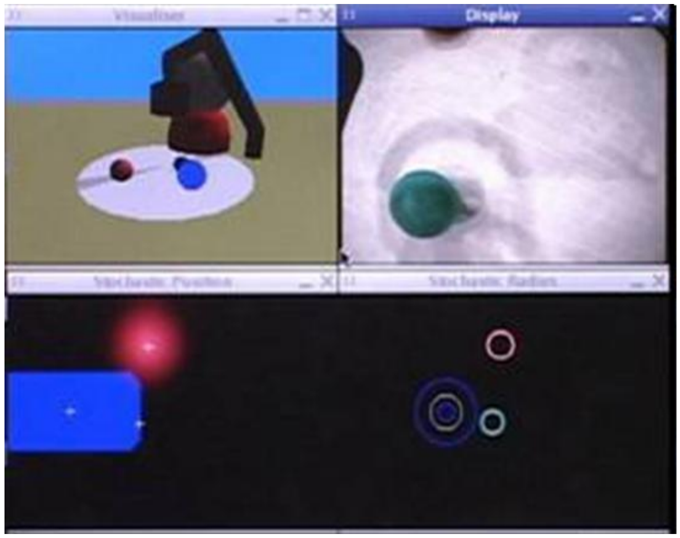


Figure 7-1: Ripley the Robot operation example, part A: GSM contents after the robot is told that, "There is a blue object at the left"

later matched and possibly updated with sensory information.

In our example, initially the state of the situation is as follows: the robot is looking down towards the center of the table, has previously seen a red ball at the top of the table, and now a green ball is directly in its view. Then, a user informs the robot that "there is a blue object at the left" (which is fed to the imaginer). This is the moment depicted on the left in figure 7-1. In more detail:

On the upper left window, one can see a visualization of the robot's estimate of the most likely state of the situation - this visualization is fed through the categorical (single-valued) layer of the robot's GSM.

On the upper right window, one can see what the robot's camera output at the moment is: the robot can only see the green ball - neither the red ball that it saw a while ago nor the blue ball that it was told about is within the field of view yet.

On the lower left window, one can see the contents of the stochastic layer of the "position" property of the balls. Look at the position distribution for the red object (red circularly symmetric cloud at the top), and compare with the position distribution for the green object (a single crossmark at the center): the spread of the red object's distribution of possible positions is much larger than that of the green objects's - which effectively has a zero-variance "delta" spike as its distribution. Thus, the robot entertains much more certainty about the position of the green ball (which is directly in view at the moment) than about the red one (which was seen a while ago and might have moved in the mean time). Now compare the position distribution of the blue object (blue rectangle at the left) with that of the red (red spherical cloud at the top). The uncertainty about the red object's position arose due to the sensory detachment of it - it has not been visible for a while. Thus, the diffusion process took over (section 5.2.2), and slowly diffused the distribution (progressively enlarging the cloud), which distribution originally was a "delta" while the red one was still in view. The uncertainty about the blue object's position arose due to a different reason - the underspecification of the linguistic statement "there is a blue object on the left". We know it is on the left, but where exactly? We have to spread our bets according to a prior - here, instead of an empirically acquired prior about "left", we spread our bets homogeneously on the region that our "left" classifier would classify as left (see section 5.2.2). Again, this region will progressively grow due to diffusion, until we actually see the blue ball that we have heard about.

On the lower right window, one can see the contents of the stochastic layer of the "radius" property for the three balls. The three concentric circles drawn for each object correspond to the radii ($\mu-\sigma$, μ, $\mu+\sigma$). First, notice that we entertain a lot of certainty about the radius of the green object. This is to be expected, as we have accumulated a lot of recent sensory information about it - the green object is and has been visible for a while. Then, notice that we also entertain quite a lot of certainty about the radius of the red object, although it is not visible anymore, and its distribution is under the influence of a diffusion process. Notice that its position distribution has diffused quite a lot (red cloud on the lower left window). Why is this not the case for the radius distribution of the red ball as well? The answer is that the diffusion process for the "position" property has a different time constant than the process for the "size" property. This is so, because through empirical observation we tune the constants in order to correspond to how much we expect each property to change, based on the previous experiences on the tabletop. And indeed, positions of objects change quite frequently; but for the case of Ripley's plastic balls, their radii do not change (and in general, sizes change much less frequently than positions). Thus, as we have commented upon in (section 5.2.2), different properties get different diffusion time constants on the basis of their expected growth of variance with time. Now, having noticed the radius distributions of the green and red balls, both of which demonstrate certainty, notice the radius distribution for the blue ball: here, the three concentric circles do not overlap, and clearly there is considerable distance between ($\mu-\sigma$, μ, $\mu+\sigma$), i.e. the variance (sigma) is large. Why is this the case? Remember, that we have been informed about the existence of the blue object not through vision, but through speech: we have heard that "there is a blue object on the left". But this statement did not include any information about the size of the blue object; thus, the blue object's radius distribution is again filled in with a prior.

Now, having talked about the stochastic layers of the position and radius properties, let us move on to the categorical layer. First, the green object: we expect "green", "at the center", and "small" to have dominant probabilities. Thus, if somebody asked: "what color is the object that is at the center?" he would get the answer "green", and if he asked "where is the green object", he would get "at the center". Now compare with the red object's categorical layer - this will contain "red" and "small" with dominant probabilities, but when it comes to position, although "top" will have the highest probability, there will be some spillover to other position categories, such as "center" - notice that the red cloud at the stochastic layer (lower left window) is pretty diffused, and also spills over into regions that are under the domination of the "center" classifier. Thus, when the stochastic layer is fed to the categorical (see L1-to-L3 in section 5.2.2), although the category "top" will have the highest probability, other categories will contain some too. Thus, if we ask "where is the red one", the answer will be "most probably at the top" and

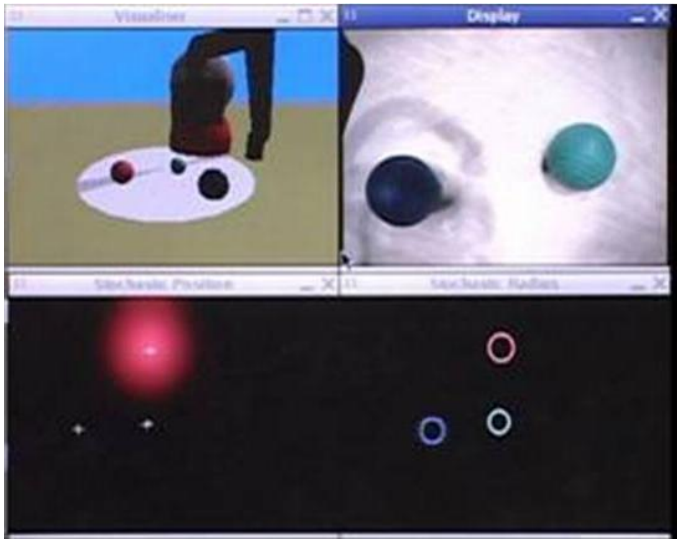


Figure 7-2: Ripley the Robot operation example, part B: GSM contents after the robot moves its head and sees the blue obj.

not "at the top"[1]. Now, let's move over to the blue object, that we've heard about but not seen. At this moment, the categorical layers of the blue object are filled with the values corresponding to the verbal categories given, i.e. "left" for position, "blue" for color, and all categories equiprobable for size. Thus, if the robot is asked "What color is the one at the left?" it will answer "blue", even though it has never seen the object yet, and doesn't know exactly what shade of blue it is. However, when the robot is asked "How big is the one at the left?" it promptly responds "I have no idea" given that this information was not linguistically transmitted and that all size categories are a priori equally likely for blue objects at the left. Later, as we shall see, when it will see the object, it would answer "It is small", as it has captured adequate sensory information.

Now, let's move on to figure 7-2, which follows temporally after figure 7-1. In the right hand side picture of figure 7-2, we have commanded the robot to "look at the left", and the robot has now moved its head, and the blue object that it had previously imagined (after "there is a blue object at the left"), now came into the field of view, and it has now been seen. Compare this to the previous figure (7-1): At the "stochastic position" window (lower left) the blue rectangular area in figure 7-1 has shrunk to a single point (the point under the leftmost cross in figure 7-2). Thus, the robot doesn't only know that the blue object is somewhere at the left, but is much more certain about exactly where it is. At the "stochastic radius" window (lower right), the outer and inner blue circles that existed in Figure 7-1 have now shrunk and expanded in order to coincide with each other, and their current radius happens to be within the "small" category. Thus, when the robot is now asked "what size is the blue one?" it will respond with "small" (and not "I have no idea" as it would before seeing the blue object and after hearing "there is a blue object at the left"). Thus, the robot has effectively updated the information about the blue object that had originally come through a linguistic description ("there

[1]For more details, see the description of the "inquirer" in section 6.4.

is a blue object at the left"), with relevant sensory-derived information. Notice that the blue object continues to have the same unique identifier as it had before, while if the sensory information did not match the language-derived expectation, this would not have been the case. I.e. if the robot moved its head to the left and the object that was seen was significantly different than the expected (i.e. for example red and not blue as expected), then a new $SIMPLE\_OBJECT$ would have been instantiated within the GSM, after the destruction of the old $SIMPLE\_OBJECT$.

This example should have illustrated the following points:

- The *contents of the three layers* of the GSM. Some relevant sections: 4.2.2, 6.3
- The *two "species" of uncertainty*: uncertainty due to linguistic underspecification (when we hear "there is a blue object at the left" we do not know exactly where at the left it is, what shade of blue it is, and how big it is), and uncertainty due to lack of recent information (if we haven't seen or heard about an object's position for a while, it might have moved). Some relevant sections: 1.4, 5.2
- The way that *sensory expectation arising from a linguistic description can be verified and further updated* given sensory information (i.e. how we went from "there is a blue object at the left" to updated, more precise information when the robot actually saw it; i.e. how we matched expectation with incoming sensory information, and updated the instantiated object on the basis of the incoming sensory information - by getting more precise information about its current position, size, and color). Relevant section: 5.2
- The way that *multiple gradations of certainty* are verbalized, such as: "I have no idea" ... "Most probably at the left" ... "At the left". Relevant section: "inquirer" description in 6.4

Furthermore, notice that in the video containing the example [Mavridis and Roy2006b], a session of questions about the past follows - in which the robot resolves temporal referent, remembers, and answers questions about the past, such as:[2]

Q: "How big was the blue one when your head started moving? (i.e. shortly after "Look at the left" was heard, and before the blue object became visible)

A: "I had no idea"

This part of the video illustrates a use of the moments and events representations included in the GSM. Some relevant sections: 4.2.1, 5.1.4, 6.3.

7.2 Levels of Evaluation of SCAs

Having talked in the previous chapters about SCA's and the GSM proposal, and having also described a real-world implementation on the robot Ripley, and presented a detailed example of its operation, the question now follows: In general, how does one evaluate a system as complex as an SCA?[3]

[2]Full text of the dialogue can be found at the section 1.1.3 of this thesis.

[3]Notice that, for the case of conversational robots, almost none of the papers cited in this thesis contain any quantitative evaluations - so the levels mentioned below are not so much a description of what exists, as they are a proposal of what could exist.

We propose that an SCA may be evaluated at various levels:

E1) *Module-level*: One could attempt quantifying the performance of each of the system's components (speech recognition, vision etc.). However, this would not be highly informative about the effectiveness of the system as a whole.

E2) *System-level*: One could evaluate through measurables that involve more than one component. For example, for the composite VisionSystem-Visor-SituationModel chain[4], how much statistical error or delay exists between the "outside world" (external reality) and the "inside world" (situation model)?[5] What is the transfer function between the two?

E3) *Task-level*: For a repertoire of specific tasks, and some criteria of success, how often is the robot successful in carrying out these tasks? For example, in responding to a human saying: "hand me the red one" by indeed handing him the red one?[6]

E4) *User satisfaction*: One could measure human satisfaction while interacting with the system through questionnaires or otherwise.

E5) *Human replacement test*: How close to a human are the SCA's responses, given same situations and utterances heard?[7]

Let's now have a closer look. In the above levels, we started by focusing on the *system and its subsystems*, and proposed an engineering-style subsystems / chains of subsystems evaluation - silently assuming that a good SCA is one whose parts perform well (the equivalent statement for humans would be: a good human assistant is one who has good eyesight, good dextrous skills etc.). Then we focused not on the system, but on the *task*: how often is the required task carried out correctly? This time, we silently assumed that a good SCA is an SCA that carries out tasks successfully (a good human assistant is one that can make coffee, prepare salad, describe an experience etc.) Then, we moved on from task to *user*: we tried to see how satisfied the human users were - a good SCA is one which keeps me happy (equivalently, a good human assistant is one which keeps me happy). Finally, we moved from user to *closeness to human behavior* - a good SCA is one that does what a human would do (a good human assistant is one that does what we would expect other human assistants to do).

7.2.1 How to choose among the levels

Which of all the above levels is more suitable / informative for the evaluation of an SCA? The answer is, that it generally depends - in our view, either user satisfaction or human replacement are most important. In case the purpose of the system is to be commercialized, we would expect user satisfaction (E4) to be closest to a sales predictor. In case it is claimed that the system is able to do what humans can in some way (more of a cognitive science than an AI goal) - than human replacement (E5) seems more relevant. An example of such a claim would be the following:

[4]See figure 5-16in section 5.3.

[5]Some attempts towards that direction are described in [Oney2004].

[6]Many modes of failure to do so exist: speech recognition errors, vision errors, grasping errors etc.

[7]In an extended form, this is almost an embodied version of the "turing test" [Turing1950] for a prespecified set of situations. The original test prescribes communication with the human/machine through a teletype: here, we can see the machine's/human's body and its actions, hear its voice and so on.

"The SCA attaches the same meaning as humans do to the words it hears and utters"

This is quite an interesting claim, which is also of particular relevance to our extended discussion on semantics that will take place in chapter 8. The sections most relevant to this claim are 8.3.1, 8.3.2, 8.3.3, and even more so 8.5.4 and 8.5.5. The interested reader should refer to these sections in order to examine their consequences to the above claim.

7.2.2 The "human replacement" evaluation

Now, let's examine the *human replacement evaluation* in a little more depth. An obvious question arises: when performing a human replacement evaluation, how do we select the tasks/situations that the robot will be exposed to? Also, how do we judge whether its responses were adequately near to a human response?

Using psychological/behavioral tests for *evaluation*

For the above questions, one way to try to decrease the arbitrariness of the possible answers, is the following: let us try to use psychological / behavioral or other competence tests for humans, and use suitably qualified professionals for judging humans as judges for the SCA's. For example, one can use standard situated language comprehension tests that are administered to children. In the next section we will focus on such a test and examine how it relates to our implemented system.

Using psychological/behavioral tests as *design specifications* for SCAs

Another potential use of such tests is in order to provide a sequence of landmarks of behavioral abilities to be reached by SCA's as they grow in maturity - effectively answering the question: what additional abilities for an SCA should we pursue next? I.e. we can use a test administered to a 3-year old as a first design specification as well as landmark/goal. Then, we can extend to a test for older kids etc.

When are we not "cheating" in the test?

However, when one uses such a test both for specifying the required abilities of a system while designing it as well as for evaluating a system, there is an obvious question that arises: aren't we cheating by knowing what the SCA will be tested upon while designing it? If somebody knows the questions of the test before the test is administered, and is also given the answers, how could we possibly believe that his passing the test carries any significance? This question will be dealt with in more depth in section 7.5, and explicit guidelines will be given for the meaningful administration of such test to robots.

We will examine in more detail a specific psychological test whose contents are highly relevant to our currently implemented system on Ripley the Robot, namely the Token Test, right after presenting the quantitative evaluation of Ripley in the section that follows.

7.3 A quantitative evaluation

Two types of quantitative evaluation were performed. The first, was positioned at the *Task-level*, and the second targeted the *Module-level*, according to the levels introduced in section 7.2. I will proceed by introducing four qualitatively different species of possible errors, suggesting specific remedies for further performance improvement, then presenting the tasks and the results of the Task-level and Module-level quantitative evaluations, and closing with a discussion.

7.3.1 Species of errors

A main focus of both evaluations was error analysis; and the following four qualitatively different species of errors dominate in the operation of the current implemented system (arranged roughly in order of frequency):

E1) Speech recognition errors: The said utterance is not recognized correctly by the speech recognizer, and the produced output is different than the said utterance. Notice that this is not necessarily always a catastrophic error when it comes to correct task completion (although quite often it is); for example, if the said utterance is "what color is the small object?" and it is recognized as "what color is the object?" while there is only a single small red object occupying the GSM, then the effect of the command would still be correct: the robot would still answer: "red".

E2) Mechanical manipulation errors: The execution of a motor command (such as "pick up", "touch") fails, due to motor control inaccuracies or gripper slippage. The most frequent type of these errors is gripper slippage, which might only occur during "pick up" or "hand me" commands.

E3) Categorical misclassification errors: A property of an object (color, size, etc.) is misclassified: for example, a green object is classified as being "green", "most probably green", "either blue or green" etc. The cause of this error is most usually the variability of the apparent property: for example, lighting conditions can affect the apparent color of an object significantly, for the case of Ripley, the most frequent error of this species occurs with blue-ish green objects, which are sometimes classified as "most probably green", "either blue or green", or even "blue".

E4) Vision system errors: The vision system output differs from the contents of the apparent view, and thus consequently the GSM contents differ from external reality. The most frequent type of this error occurs when an object disappears from the segmenter output for a number of frames, which most often occurs either when an object is at the edge of the field of view, or when it is very near another object, in which case both might appear a single region at the segmenter output.

Suggested remedies for further performance enhancement

Some possible cures for the four species of errors introduced above are the following:

E1) Speech recognition errors: Retraining / adjustment of the speech recognition parameters, speaker-specific training / adaptation, noise cancelation through sound source separation or otherwise, using the situational / task context for disambiguation

E2) Mechanical manipulation errors: Better precision of the motor control system, autocalibration, intelligent adaptive gripping.

E3) Categorical misclassification errors: Better training set for the categorical classifiers, lighting conditions compensation through the apparent color of a fixed object (tablecloth for example) or otherwise, generic property drift compensation through similar techniques, real-time negotiation / adjustment of categorical boundaries through human feedback.

E4) Vision system errors: Better tuning of the segmenter / alternative segmentation techniques, more intelligent methods for evidence accumulation for creation / deletion of objects.

7.3.2 Task-level evaluation

A typical task was designed (task T1), containing a range of different types of utterances (Questions, Motor Commands, Inform speech acts), and exposing all of the three main novelties of Ripley as compared to other state of the art conversational robots (Handling informs and imagining, remembering and resolving temporal referents, uncertainty handling / verbalization) (see section 6.5.1)

The task that was used was the following (human actions and expected robot actions listed):

H1: [The human sits by the table facing the robot and slightly to the left. He places a green spherical object of small size (O1) roughly at the center of the table.] "Look at the table!"

R1: [Robot moves, looking down towards the center of the table] "OK!"

H2: "Where is the object?"

R2: "At the center"

H3: "How big is the object?"

R3: "Small"

H4: "Touch the green one!"

R4: [Robot moves, touches the green object, and returns to original position] "OK!"

H5: "Imagine a red object at the left!"

R5: [Robot instantiates imaginary object within the GSM with red color (unknown shade), position at the left (unknown exact position), and unknown exact size] "OK!"

H6: "How big is the red one?"

R6: "I have no idea"

H7: "Look at the red object!"

R7: [The robot moves towards the left and looks down. The seen object is matched with the sensory expectations arising from the previously imagined object, and then the object's properties are updated on the basis of what is seen] "OK!"

H8: "How big is the red one?"
R8: "Small"
H9: "How big was the red one when your head started moving?"
R9: [The robot resolves the temporal referent, and remembers] "I had no idea"
H10: Where are the objects?
R10: "One is at the center, and the other is at the left"

Notice that in the above task, the only acceptable behavior from the robot is not only the behavior prescribed by R1-R10. There are also other acceptable alternatives for R2, R8 and R10. For example: for the case of R2, if there has been a small time interval between R1 and H2, then sufficient evidence about the object position has not been accumulated: thus, "most probably at the center" is the right answer, and not "at the center". Similarly for R8 ("most probably small") and R10. These alternative acceptable answers, are counted as correct in the results that follow.

Task-level evaluation: Results

The above task was executed 11 times by an untrained human operator (other than the author) in the thirties, who was unfamiliar with the project and a native speaker of English. When the expected outcome of a command did not occur, the speaker repeated the command. Errors can be classified into three types in terms of severity: *catastrophic* (task cannot continue), *repairable by repetition* (task can continue by repetition of the command) and *non-interruptive* (a slight deviation from the expected result occurs, which however does not prohibit the normal continuation and successful completion of the task).

The shorthand notation that is used in the error analysis that follows in figure 7-3 is: SRE (for *Speech Recognition Error*), and CME (for *Categorical Missclassification Error*)

Task-level Evaluation
(Task T1)

Error Analysis:
Correct Response: 87.1%
Speech Rec Error: 8.1%
Cat Misclass Error: 4.8%

Task Prolongation:
Median Xtra Task length: 1 utterance
Avg Task Prolongation: 11.5%
Avg Errors per Task: 1.18

Most vulnerable turn:
H-R4: “Touch the green one!”
(SRE and CME probable!)

Task Timing:
Avg Turn duration: ~6 sec
Avg Task Duration: ~78 sec

Successful task completion rate: 100%

Figure 7-3: Task-level evaluation results (Task T1)

An examination of figure 7-3, reveals that all trials resulted in successful task completion (i.e. no catastrophic errors occurred), however due to the need for repeated human utterances because of repairable by repetition speech recognition errors, the trials were prolonged by 1 extra utterance median (11 percent average length increase).

The main species of errors were due to speech recognition (SMEs), which arose in roughly 8 percent of the utterances heard, followed by CMEs, which occurred during the servicing of 4.8 percent of the heard utterances. It is worth noting that the two other less frequent species of errors did not occur in these trials - all the "touch" commands were successful (no mechanical manipulation errors), and no destructive object disappearances occurred during this trial.

7.3.3 Module-level evaluation

Due to the sufficient coverage of speech recognition errors by the task-level evaluation, at the module-level evaluation two of the remaining three species of errors were targeted: *mechanical manipulation errors* (gripper errors in particular) and *property misclassification errors* (color misclassification in particular). The speech recognition system was bypassed by typing commands directly through a keyboard, and thus speech recognition errors were effectively isolated, enabling the precise alignment of the two above tested types of errors with their relevant modules: the *motor control subsystem*, and the categorical classifiers accessed from the *visor*.

The following two tasks (T2 and T3) were used:

T2) Mechanical manipulation error assessment: 11 "touch the object" commands and 11 "hand me the object" commands were issued for each of two different objects: object O1 (green spherical object, heavy) and object O2 (red spherical object, light). The success / failure of the commands was logged.

T3) Object misclassification error assessment: 11 "what color is the object" questions were made, for each of the above described objects O1 (green) and O2 (red). The answers were logged.

Module-level evaluation: Results

Task T2: 82 percent correct for "hand me" [gripper slippage], 100 percent correct for "touch".

Task T3: 64 percent correct for O1 (green object) [misclassified as "blue" or "either green or blue" etc], 100 percent correct for O2 (red object).

7.3.4 Discussion

In general, although the system is not absolutely error-free, this does not inhibit its task completion functionality and allows fluid interaction with the robot.

In particular, for the case of the *task-level evaluation*, although there are errors (almost completely due to speech misrecognition), their ultimate effect was only the prolongation of the task by an extra repeated human utterance on average. Regarding the *module-level evaluation*, the main motor control system performance detriment is due to gripper slippages (affecting the "hand me" comments - can be improved with more careful calibration), but with perfect execution of all "touch" commands. Also, while there exists considerable categorical misclassification for the green object (O1) (due to the borderline blue-ish / green color of the object and local light intensity variations), there is perfect classification for the red object (O2).

Error detection and dialogue repairs

An interesting tangent to the previous discussions, as well as a possible direction for extensions, is the following:

In the task-level evaluation, the human was instructed to repeat a command in the case where the expected response from the robot did not occur. Thus, the human was detecting errors in the dialogue, and he was attempting to repair them through this simple recipe: repetition. And, fortunately, all errors that occured during the evaluation were not *catastrophic*: they were *repairable by repetition.* Now, the question arises: are there other ways to repair such errors during the dialogue?

A quick consideration will reveal that the answer to the above question is positive. However, in order to attempt to repair, first the error has to be detected; either at the human or at the SCA side. In the task-level evaluation described above, only the human was detecting errors, by knowing what the expected response should be. But how can the robot detect errors and repair errors too?

Let us consider some possibilities:

Categorical Missclassification Errors: Let us suppose that the human utters: "Touch the green object", and that a single object is on Ripley's table, which the human classifies as being green, while the robot classifies as being blue. In the current implementation, the robot's answer would be: "No such object exists". But notice that Ripley currently has a single-initiative (*user-initiative*) dialogue system; he can only respond to user utterances, and cannot initiate dialogue pairs (*mixed-initiative system*). If we extend our system, we could alternatively have Ripley respond with: "I think no green object exists", followed by "I think this is blue. Do you think this one is green?". In that case, a positive answer from the human, could have enabled the adjustment of the *categorical classifier* that the robot possesses for the color "green", in order to include this specific color shade within the receptive region of the classifier. Thus, essentially, the robot would then be performing *online retraining* of its categorical classifiers; which is just a special small-scale case of situation model alignment between the robot and the human (see species and some possible remedies for case D5 in section 3.2.2)[8].

Speech Recognition Errors: In case the output of the speech recognizer produces an incomprehensible sentence, the robot currently responds with: "I don't understand. Sorry." This does not mean, of course, that an incomprehensible sentence was uttered; it might have (in which case this was not a speech recognition error),

[8]I am indebted to a very useful comment from Prof. Ted Selker that provided motivation for such a dialogue system extension.

or it might not (in which case, we indeed had a speech recognition error). In the case that we were also keeping a discourse model (storing previous utterances), we could also have detected a repetitive incomprehensible sentence; and then, other strategies might have become applicable for repair. For example, again if we extended to a mixed-initiative dialogue system, the robot could then have attempted to ask a yes/no question regarding possible actions, such as "Do you want me to touch an object?". Such a question, usually expecting a single-word answer (either "yes" or "no), could have been followed by the biasing of the language model of the speech recognizer towards these two words. And such an action, could have resulted in much higher recognition rate even under severe noise conditions. This is in accordance to what humans would usually do in a noisy communication channel: use short yes/no questions, whose answer could have much more easily come through.

Mechanical Manipulation Errors: In case a "pick up" or "hand me" command fails, and the object slips, this is readily detected by the touch sensors of the gripper. In that case, the robot currently utters: "Oops!". Thus, the error is detected. Some possibilities for repair are the following: again, with a mixed-initiative dialogue system extension, the robot could actually ask for human assistance in gripping the object. Alternatively, the robot could use another technique: it could try a slightly different motor control trajectory, that might have enabled successful gripping[9].

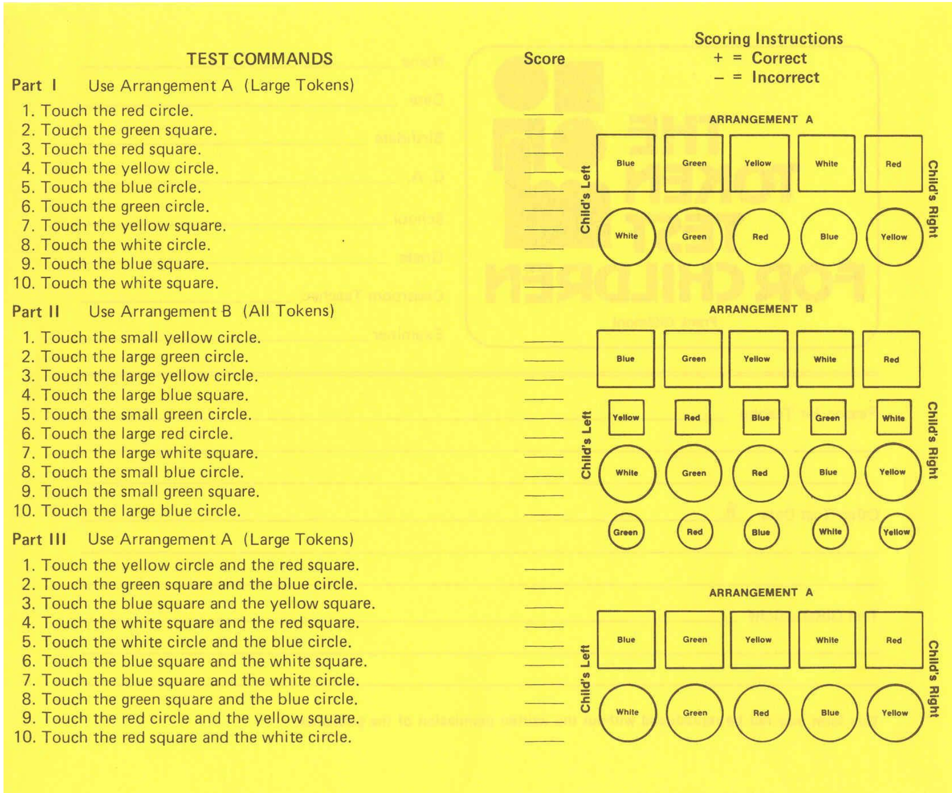
Thus, we have seen some possibilities that could exist for extending dialogue error detection and correction between the robot and the human. These seem to be all fruitful avenues that could well be explored in future real-world systems.

7.4 The Token Test

The Token Test [DiSimoni1978] is a standard situated language comprehension test, that is commonly used to assess language skills of children who exhibit language acquisition difficulties, typically used for three-year old children. To administer the test, the evaluator arranges a set of physical tokens on a table and asks the subject to perform various manipulation tasks ("When I touch the green square, you take the white one", etc.).

The Token Test is a suitable evaluation for our implemented system since it evaluates basic language-to-world mapping skills and does not rely on social or cultural knowledge. The Token Test is divided into five parts ordered in increasing difficulty (see figures 7-4(a) and 7-4(b)).

[9]This is just a special case of a more generic method for retrying failed actions: just vary them a little (similar to "jiggling the key" when the door does not open). The same principle could be applied to speech production (vary the way you say things a little) and so on. I am indebted to a comment made by Deb Roy for this realization.



TEST COMMANDS

Score

Scoring Instructions
+ = Correct
– = Incorrect

Part I Use Arrangement A (Large Tokens)

1. Touch the red circle.
2. Touch the green square.
3. Touch the red square.
4. Touch the yellow circle.
5. Touch the blue circle.
6. Touch the green circle.
7. Touch the yellow square.
8. Touch the white circle.
9. Touch the blue square.
10. Touch the white square.

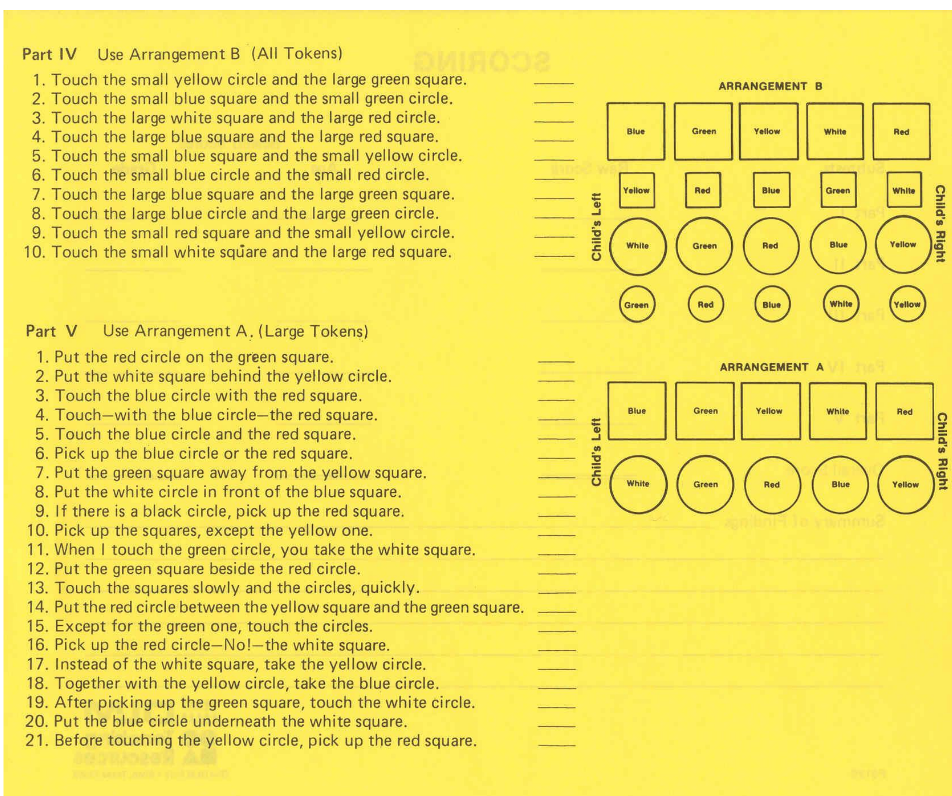
Part II Use Arrangement B (All Tokens)

1. Touch the small yellow circle.
2. Touch the large green circle.
3. Touch the large yellow circle.
4. Touch the large blue square.
5. Touch the small green circle.
6. Touch the large red circle.
7. Touch the large white square.
8. Touch the small blue circle.
9. Touch the small green square.
10. Touch the large blue circle.

Part III Use Arrangement A (Large Tokens)

1. Touch the yellow circle and the red square.
2. Touch the green square and the blue circle.
3. Touch the blue square and the yellow square.
4. Touch the white square and the red square.
5. Touch the white circle and the blue circle.
6. Touch the blue square and the white square.
7. Touch the blue square and the white circle.
8. Touch the green square and the blue circle.
9. Touch the red circle and the yellow square.
10. Touch the red square and the white circle.

(a)



Part IV Use Arrangement B (All Tokens)

1. Touch the small yellow circle and the large green square.
2. Touch the small blue square and the small green circle.
3. Touch the large white square and the large red circle.
4. Touch the large blue square and the large red square.
5. Touch the small blue square and the small yellow circle.
6. Touch the small blue circle and the small red circle.
7. Touch the large blue square and the large green square.
8. Touch the large blue circle and the large green circle.
9. Touch the small red square and the small yellow circle.
10. Touch the small white square and the large red square.

Part V Use Arrangement A. (Large Tokens)

1. Put the red circle on the green square.
2. Put the white square behind the yellow circle.
3. Touch the blue circle with the red square.
4. Touch—with the blue circle—the red square.
5. Touch the blue circle and the red square.
6. Pick up the blue circle or the red square.
7. Put the green square away from the yellow square.
8. Put the white circle in front of the blue square.
9. If there is a black circle, pick up the red square.
10. Pick up the squares, except the yellow one.
11. When I touch the green circle, you take the white square.
12. Put the green square beside the red circle.
13. Touch the squares slowly and the circles, quickly.
14. Put the red circle between the yellow square and the green square.
15. Except for the green one, touch the circles.
16. Pick up the red circle—No!—the white square.
17. Instead of the white square, take the yellow circle.
18. Together with the yellow circle, take the blue circle.
19. After picking up the green square, touch the white circle.
20. Put the blue circle underneath the white square.
21. Before touching the yellow circle, pick up the red square.

(b)

Figure 7-4: The Token Test for children

	Ripley (current state)	Token Test
General form of commands	"<action> the <object_description> [when <event_description] +questions, imagine statements etc.	"<action> the <object_description> [and the <object_description>]"
<action> belongs to:	{touch, pick up, hand me, look at}	{touch}
<obj_des>:	Contains some of: <size> <color> <object> <locus>, with singular/plural object	Consists of: [<size>] <color> <shape>
<size> belongs to:	{small, medium, large}	{small, large}
<color> belongs to:	{red, green, blue}	{red, green, blue, white, yellow}
<shape> belongs to:	{-} (full 3D objects)	{square, circle} (2D flat objects)
<locus> belongs to:	{center, left, right, top, bottom ...}	-

Figure 7-5: Comparison of Ripley's current behavioral repertoire with requirements of first four parts of the Token Test. *Yellow*: Abilities that the Token Test requires that have not yet been implemented on Ripley. *Red*: Abilities of Ripley that surpass the relevant Token Test Requirements

7.4.1 The current state of Ripley vs. the Token Test

Using the GSM-based real-world implementation on Ripley the Robot that we have described, we have attained language comprehension abilities that are comparable to those implied by the first four parts of the Token Test, and in some directions also surpass those abilities[10].

Let us now examine the above claims in more detail: First, in what way are Ripley's language comprehension abilities comparable to those implied by the first four parts of the Token Test? And second, in what ways do we surpass the abilities implied by the whole of the Token Test?

Let us start with the first claim, and perform a comparison (see figure 7-5).

By comparing the token test requirements with the relevant abilities of Ripley, we see that:

- Regarding the *form of the acceptable commands*:

While Ripley cannot currently handle the "and" conjunction in order to connect two consecutive commands[11], he is capable of *at least three levels of generality more than the Token Test requirements* (genarilty regarding the form of his commands): First, both sin-

[10]Also, a theoretical systematic design of a GSM-based system that can pass the token test has been devised by the author, and is discussed in appendix D.

[11]Which however would be an easy and straightforward extension for the current system - and anyway the user can still easily give the two commands one after the other instead of connecting them with "and".

gular/plural descriptions can be handled appropriately, while the Token Test only has singular; second, event-related descriptions referring to the past such as: "The object that was at the left when the green one appeared" can be handled; and third, apart from requests for motor acts, Ripley is also capable of answering questions and handling "inform" speech acts.

- Regarding the *set* of acceptable *motor actions*:

Ripley *clearly surpasses* the relevant Token Test requirements: apart from "touch", Ripley can also handle and perform "pick up", "hand me", and "look at".

- Regarding the *combinatorics* of acceptable *property descriptions*:

Ripley can accept any combination of <size> <color> <object> <locus>, while the Token Test requires <color> <shape> with optionally also <size>. Thus, regarding the combinatorics of the property descriptions, *Ripley is more flexible*.

- Regarding the acceptable *property dimensions*:

Ripley cannot yet support shape descriptions. However, notice that although verbalized shape descriptions are currently not supported (although they could be implemented with ease, as commented upon in 9.1.1), Ripley currently supports full 3D objects through the voxelizer[12].

- Regarding the acceptable *property categories*:

Ripley can handle one more size than those required by the Token Test (medium), and although he can handle three colors (red, green, blue), he does not currently support the two extra colors required for the Token Test (white, yellow). However, this would be a most straightforward extension - retraining of the color categorical classifiers as well as the object segmenter is all that would be required to achieve this. Finally, Ripley can handle various locus categories (such as center, left, bottom etc. and combinations), which is again one more ability surpassing the relevant Token Test requirements.

Thus, it should now be clear that Ripley's current abilities are not only comparable to those implied by the all except the last parts of the Token Test, but also that Ripley's current abilities clearly surpass those implied by the whole of the Token Test in many important ways - motor action repertoire, locus descriptions, combinatorics of descriptions, 3D shapes, plural number, and most importantly: support for the past and events, support for question answering and inform statements.[13]

Furthermore, it should be clear by the above comparison and discussion that most of the required extensions for covering the whole of the Token Test are pretty straightforward, and that they leverage greatly on the existing system and the GSM proposal. Finally, remember that a theoretical design for covering the whole of Token Test has already been devised in a systematic manner, we will see in appendix D.

---

[12] See section 6.4 (as demonstrated by the video [Mavridis2006b]), an ability surpassing the Token Test requirements (which requires only two species of 2D "flat" objects - squares and circles). Also, Ripley can currently handle locus descriptions (such as "the object at the left"), an ability which again surpasses the Token Test requirements.

[13] Our implemented system thus goes well above the here-and-now situatedness (see section 8.1.4) of the language assessed by the token test - it extends spatiotemporally all the way to the realms of past experience ("where was the red one when your head started moving?") and of imaginary situations ("imagine a blue object on the left"), and also extends the speech act horizon of the token test: from handling only REQUESTs for motor action (commands) to being able to also service REQUESTs for speech (questions) and INFORM speech acts ("there is ... ").

Now, having discussed the Token Test in relation to our implementation, we will move on to a question that was left unanswered in the previous section: how can we use such psychological tests for humans not only for evaluation, but also as design specifications for SCA's / robots?

7.5 Behavioral tests for humans serving as design specs and as evaluation tests for SCAs

7.5.1 What consists cheating in a test and what does not?

When trying to use a behavioral test designed for humans as a design-driving specification for an SCA, one main difficulty arises, which we will discuss here. The human to be tested is supposed not to know exactly the contents of the test before it is administered. However, we assume that throughout his experience so far in his life, he has received adequate explicit or implicit training stimuli in order to perform well. When administering a behavioral test to an SCA, we certainly should not allow the designer of the SCA or the trainer to have given explicit response specifications for each of the specific tasks that comprise the test. Allowing such a state of affairs would be equivalent to having trained the human for the specific tasks/questions of the test: having shown him the correct responses and having made sure that he performs them well. This would certainly count as cheating in most human testing cases. So the question arises: if having supplied the "questions" and the "correct answers" to the specific test, and having made sure the human performs them adequately counts as "cheating", what would be considered allowable?

7.5.2 The wider domain versus the specific questions

Let's examine the human case again. A high school student preparing for his SAT tests is exposed to intense training. This training takes the form of explicit tuition on the related subjects, practice tests etc. Then, one day, he has to sit his actual exam. Why doesn't the training count as cheating? Because it is general enough in order to cover the wider area to which the questions of the test to be administered belong to, but is not specific enough in order to cover only these questions. It would have been highly unlikely for the training to cover only the specific questions of the administered test, without any pre-knowledge or "leak" of the exam questions. So what is the moral for the SCA case? Any proposed design solution that passes a human behavioral test should be able to pass any test that belongs to the wider area to which the test belongs, and not only the specific test in question. Of course, the delineation of the "width" of the area to which the specific tests belongs should be carefully justified. Notice that an analogous situation holds in the pattern recognition evaluation literature. In supervised training problems, we are given a training set and are asked to design a classifier that performs well on a (supposedly unknown) testing set. The training set should be general-enough in order to cover and represent densely enough the area where the testing set might belong to. However, it must again not be specific enough so that it suspiciously only includes the testing set.

7.5.3 The test design problem

After all, one must ask a wider question: what is the purpose of testing? What was the test designer trying to achieve? Usually, the test design procedure is posed as follows:

a) Decide on the limits of the wider area of the material to be tested

b) Select specific questions out of the wider area, such that:

b1) they can be practically tested in limited time (few in number, short answers)

b2) the answers given to the selected questions can have high predictive value towards the answers that the examinee would give to all the questions comprising the wider area of the material to be tested.

In short, the test designer must chose a few easily testable questions that are however highly indicative of the examinee's mastery of a wider area of knowledge that is being tested.

7.5.4 The "three commandments"

The above discussion has clear implications towards using behavioral tests for humans as design specifications (or alternatively as training material) for robots:

C1) First, the test must be *reverse-engineered*: given a specific test we must try to delineate the wider area in which the tasks of the test belong, and for which the specific test acts as a representative sample of. In case we have explicit knowledge of the test designer's coverage intentions, we can adapt them readily. Else, we have to "grow" the wider area by using the specific questions as a "seed". The extent and generality of the wider area must be chosen.

C2) Second, we must use this *wider area* as a design specification, or sample it randomly for the generation of a *training set*. We are not allowed to use the specific questions comprising the test as the only target domain.

C3) Third, we must clearly and *explicitly justify* all of our above choices.

Minimality at odds with generality requirement prescribed by "width"

Finally, it is worth noting that the designs that can satisfy the requirements of the wider "grown" areas will often prove to be more costly than those that would only satisfy the seeds. Nevertheless, only a design that can pass the "wider" requirements is of any real value, and is furthermore more general and easily expandable in the future. Of course, the right level of generality must be decided and explicated on a case-by-case basis, as we will do in the next chapter, when presenting an example of the use of the "Token Test" as a design specification.

7.6 Recap

In this chapter, we considered the subject of the evaluation of SCAs. We first presented a detailed example of the operation of the implemented system on Ripley the Robot in section 7.1. The video that corresponds to this example, as well as many other videos of Ripley, can be found at [Mavridis and Hsiao]. Then, in 7.2, we discussed various different methods for evaluating SCA's. In 7.3, we presented the quantitative evaluation that addresses both the task- and module-levels of evaluation. In 7.4, we introduced the "Token Test" for children, a psychological test administered to children, and discussed its relation to our implemented system on Ripley. Later, in section 7.5, we discussed the use of such tests for evaluation of SCAs as well as for acting as design specifications for SCAs. Most importantly, we proposed specific prerequisites (the three "commandments") for the meaningful use of such tests both for design and evaluation without "cheating". The interested read can now move on to the presentation of a sketch of a design methodology (appendix D), that aims towards systematizing the design of SCAs given behavioral specifications or psychological tests, that might be used in place of the specification.

Now, let us see where we are in the large-scale structure of the thesis: having started by the vision of SCAs and relevant background notions in chapter 1, having progressed from a very high-level view to a specific implemented system in chapters 2-6, and having discussed the evaluation of SCAs in chapter 7[14], we will now take a higher-level view again, and in the next chapter, through an extended itemized discussion, we will attempt to reflect on various theoretical implications and connections of the presented GSM proposal.

[14] The systematic design of SCAs is discussed in appendix D.

Chapter 8

A discussion of various topics in light of the GSM proposal

In this chapter, we will consider a variety of subjects related to semantics, grounding, AI and philosophy of mind, in light of the GSM proposal. Many tangents towards a wider view will be given, as well as directions for extensions, arguments against possible classical attacks against the proposal, and discussions of the relevance of ideas found elsewhere with our proposal.

The overall organization of this chapter consists of five large blocks (each assigned to a section):

Block 1: Language grounding, Situated language

Block 2: Abstract entities, Imaginary situations

Block 3: Semantics, Situations

Block 4: Logic, Ontologies

Block 5: Stance X and its relation to GSMs

Let us consider each block in detail.

Block 1: Language grounding, Situated language

We will start by discussing topics relevant to language grounding and situated language (introduced in sections 1.3.1 and 1.3.2). In section 8.1.1 we will initiate our discussion by talking about possible meaning spaces, that are external to the world of words - sensory, sensory-motor, and goal-oriented meaning spaces. Then, in 8.1.2, we will briefly touch upon two metaphors that are useful towards the theoretical unification of speech with sensing and acting. We will then address the question: "How (and if) can we ground the whole of language?" (section 8.1.3), and then discuss the existence of multiple levels of detachment from the "here-and-now" of situatedness (section 8.1.4).

Block 2: Abstract entities, Imaginary situations

Then, we will delve into a discussion of abstract entities and imaginary situations: we will first ask the philosophical question about the mode of existence of abstract entities (section 8.2.1), then discuss where the GSM proposal stands when it comes to this question (section 8.2.2), and then propose a way for representing abstarct entities within "concrete" GSMs (section 8.2.3). Moving on from the abstract to the imaginary, we will deal with inform speech acts, imagine statements, and the "imaginer" module (section

8.2.4), and then discuss the possibility of "pure" imagination in light of the GSM proposal (section 8.2.5).

Block 3: Semantics, Situations

Later, having dealt with abstract entities and imaginary situations, we will touch upon the semantics of certain parts of speech within the GSM proposal, and also upon two aspects of situations: similarity and enumerability. We will start by discussing the semantics of verbs and adjectives within the GSM proposal 8.3.1, and then discuss extensions through successive approximations of meaning models in 8.3.2. Spatial semantics and comparative adjectives will be touched upon in 8.3.3. Similarity of situations and the possibility of enumerating possible situations will be considered in 8.3.4 and 8.3.5.

Block 4: Logic, Ontologies

We will start by discussing the bridging of the world of GSMs with that of First Order Predicate Calculus - for which numerous tools and algorithms exist - in 8.4.1. Another possible bridge - this time between ontologies and grounded semantics - will be envisioned in 8.4.2.

Block 5: Theory/stance X and its relation to GSMs

We will start by discussing embodiment and the GSM proposal, and also consider how dependent the "mind" (agent model) of an agent is on his embodiment, and what needs to be changed when "implanting" minds in new bodies (section 8.5.1). Then (section 8.5.2), a question that we have most frequently encountered by visitors who see our robot will be addressed: how is this different from SHRDLU? Then, we will ask: how are GSM-based systems different than observer-only systems (section 8.5.3)? After connecting in this way with the past landmark systems of AI, we will then move on back to semantics, and view our proposal from the viewpoint of procedural semantics (section 8.5.4). Then, a proposal for another semantic theory will be introduced, a theory which is of quite a different nature than either procedural or model-theoretic semantics, which we shall term "empirically sound semantics" - in section 8.5.5. Subsequently, philosophy will enter the picture: Dennett and the "intentional stance", the classic homunculus attack (sections 8.5.6 and 8.5.7). And finally, in 8.5.8 and 8.5.9, last but certainly not least, we will see the ways in which some of Lawrence Barsalou's ideas regarding "Perceptual Symbol Systems" and Minsky's ideas from the "Society of Mind" and the "Emotion Machine" are related to the GSM proposal.

Now, having examined the map, let the journey start!

8.1 Language grounding, Situated language

8.1.1 Meaning spaces

In the first chapter, while discussing language grounding, a number of questions came up, that we promised to address here. These questions were: First, if we have to ground the meaning of words to something external to words (i.e. not words themselves), what should this external thing be? Second, what are the limits of this approach - can we cover the whole of natural language, and if yes, how? We will address the first question here, *Meaning Spaces:*

1. Sensory ("red", "loud")

2. SensoryMotor ("jump", "search")

3. Goal-oriented ("it is hot in here")

Figure 8-1: Some Meaning Spaces

and the second a little later in this chapter.

Let us first deal with the first question: What are possible "meaning spaces" consisting of things external to words, that we can ground the meaning of words to? Here, we will propose three species of such spaces:

Let us consider each of them in turn:

MS1) Sensory spaces

Consider the meaning of "red" or "loud" - both are concerned with sensory-derived qualities. If we effectively "cut" our optic nerve or our acoustic nerve, we would expect "redness" or "loudness" to correspond to patterns of signals passing through the nerve. In our proposal and the Ripley implementation, the adjectives addressed were actually effectively grounded out to a multi-dimensional "sensory space", where each property contributes a number of dimensions - 3 for color, 1 for size and so on.

Sensory spaces are the most frequently encountered in the literature - see for example [Roy1999] [Regier and Carlson2001] [Siskind2003] etc. However, one must not forget, that they can only cover one aspect of meaning, and for some words they are clearly insufficient, as we shall see. When grounding out to sensory spaces, we are effectively reducing the agent to a "passive observer" of the world - no possibility of action or purpose is implicated.

MS2) Sensorymotor spaces

Consider the meaning of "jump". There are at least two parts of the meaning of this action verb: the recognisive part (second/third person jump) and the generative part (first person jump). The first belongs to a sensory space: how can I recognize that somebody is jumping - what in my sensory streams corresponds to somebody else jumping? For example, it might be approximated by a classifier that reports the existence of jumping people in video streams. The second, belongs to a sensorymotor space: how can I actually perform a "jump" myself? What sensorymotor control loops can achieve that - i.e. what signals should I send to my muscles, and how should I regulate them in terms of visual and proprioceptive or other feedback, in term to perform a successful "jump"? Notice that, at a first approximation, this can be just a motor routine - without any sensory feedback whatsoever: i.e. to "jump" simply send such and such a signal to your muscles. However, at a second approximation, control loops with sensory feedback are almost always implicated. As a second example, consider "search". This verb clearly involves "dialogue" with the environment through acting / sensing / acting sequences - i.e. some form of a control loop or a similar construct is a prerequisite for its execution.

In the literature, the need for sensorymotor spaces has been addressed by a number of proposals. Kai-yuh Hsiao at the media lab has demonstrated simple action learning for Ripley through HMMs, x-schemas have been proposed by Narayanan [Naryanan1997], although these have not been implemented yet on robots, up to our knowledge. I have also addressed some related questions on a paper [Mavridis2005d] proposing extensions to the Human Activity Language proposed by Aloimonos [Guerra-Filho et al.2005]. Again, one must not forget, that by adopting sensorymotor meaning spaces, although we can cover purely sensory meaning spaces as a subcase of them, we are just adopting a certain viewpoint towards meaning - they are not in any case "best" or "most-inclusive" or "universal", as we shall see. Also, one should keep in mind, that when adopting sensorymotor meaning spaces, one takes the viewpoint of the agent as an "actor" in continuous dialogue with the world - thus, sensing and acting are implicated, but no notion of purpose whatsoever is directly involved.

MS3) Goal-oriented spaces

As we have mentioned before in section 2.3.2, there exist two flip-sides to the notion of meaning: referential meaning vs. functional meaning. Referential meaning views language as referring to things - for example "apple" as referring to a physical apple in external reality, or to some pattern of neural activation in the brain etc. I.e. referential meaning answers "what" questions. On the other hand, functional meaning deals with the purpose of utterances: for example, if I want the window of this room opened, there is a number of actions I can take to achieve this - go and open it myself, say: "can you please open it?" to a friend that is sitting next to it etc. Notice that these two ways are both actions, but of a different species: one is a motor action, and the other a speech action. In the first case I move my feet and hands, in the second I move my lips. Both of these ways are interchangeable though, in terms of purpose; they will eventually have the same result and help me reach my goal. Thus, the meaning of "can you please open it" can be specified as an operator that forms part of a plan - its meaning consists of its effect on the world. This is the "functional" side of the language coin - utterances are seen as purposeful actions, that form parts of plans. Thus, I can ground words out to a space of planning operators - where also motor action operators might reside.[1]

The bigger picture

Now, having discussed sensory, sensorymotor, and goal-oriented meaning spaces, one can ask: But where do the above spaces fit in the big picture of the semiotic triangle given in chapter 2? All three reside primarily within the mind - only secondarily, through sensing and acting[2], they connect to external reality. Sensory and sensorymotor spaces belong to the "object" side of the double-sided object/purpose semiotic triangle, while goal-oriented spaces belong to the "purpose" side. One can also generate more species

[1] See for example [Cohen and Perrault1979].

[2] and through M1 and M2, in the terms of chapter 3.

Unifying Speech with Sensing and Acting

Two metaphors:
→ Seeing as asking
“I asked the chair what color it is
by moving my head, eyes & attention, and it answered me”
QUESTION GENERATION AS ACTIVE SENSING

→ Saying as moving things in minds
“I can change the position of a can with my hand,
I can change the beliefs of my friend about the can with my lips”
MIXED SPEECH/PHYSICAL ACTION PLANNING

Figure 8-2: Two metaphors: speech as sensing / speech as acting

of meaning spaces, on the basis of the dual-sided semiotic triangle. For example, one can try to project a "sensory" space (residing in the mind) to its possible external causes (residing in the world). This would amount to effectively moving from the a set of sensations in the mind that correspond to a word, to the so-called "intension" of the word: the set of all possible things in the world that a word could describe. In this extended framework, one can now start discussing the degree of overlap of "intensions" across speakers, whether meaning should be positioned in the world, or in a linguistic community, or an individual etc.[3]. Our discussion on alignment of situation models also becomes relevant here (section 3.2).

Anyway, we will not pursue this path any further within this document. Now, having seen three species of meaning spaces, and before moving on to the question of the "grand plan" of extending grounding to the whole of language, we will briefly discuss another interesting notion that has arisen while discussing "goal-oriented" meaning: the interchangeability of motor actions with speech actions. We will thus now ask: what other such analogies between speech and generic action can be found? How can one view speech, sensing and acting in a more unified framework?

8.1.2 Unifying speech with sensing and acting

Consider two metaphors:

First, let us discuss sensing and speech. If one moves from the notion of passive, continuous-feed sensing to active sensing[4], then one can decide what to look for, and when to look for it. For example, when I want to update my information about the position of a ball that some kids were playing with, I can choose to turn my head, fixate, and obtain visual information about the position of the ball. This is an example of an active sensing act - and in the context of computer vision, a considerable amount of relevant work has existed for quite a while: see for example [Blake and Yuille1992]. Topics that

[3]See for example [Putnam1973].
[4]One of the proposed future extensions for Ripley - see section 9.1.1.

have been dealt with include problems such as: in order to reconstruct the shape of an object through multiple views, what is the next most informative view of it that I can take? [Pito1999]

Now, let's return to the position of the ball estimation problem. I also have other options to obtain the same information: I can ask one of the kids. In that sense, question generation can be viewed as active sensing. From the viewpoint of the GSM proposal, there is a clear interchangeability: both the physical world and the embedded GSM of other minds resides within the GSM. When I perform an active sensing action - through vision, touch etc. - then I attempt to reduce my uncertainty about some aspect of the physical world directly. When I ask a question, I attempt to reduce my uncertainty about some aspect of the answerer's mind directly - something in his GSM estimate that is embedded in mine. Indirectly, I also reduce my uncertainty about some aspect of the physical world - if I trust his answers and his beliefs.[5] In this sense question generation can be viewed as a special case of active sensing - I can sense their mind by their answers, and indirectly I can also effectively extend my sensors through the sensory organs of others - in a human-mediated way, though.

Now, let us discuss physical action and speech. Again, consider the window example given before. The window is closed - I want the window to be opened, and a friend is sitting next to it. I can either open it on my own (through a motor action) or ask my friend to do it (through a speech action)[6]. As mentioned before, in terms of physical effect, the motor and the speech act are interchangeable in this case - and they both can be used as parts of a plan to reduce the temperature of the room. Taking the viewpoint of the GSM, both "change" things: the motor act changes the position of an object in the physical reality description of the GSM, while the speech act directly changes an aspect of the mental realm (the estimated embedded GSM) of the other person, which resides within my GSM again. The indirect effect is again the change of the position of the window panes. In this sense request generation can be viewed as a special case of physical action - I can change something in the mind of others, and indirectly I can also effectively extend my effectors through the effectors of others, again in a human-mediated way, though.[7]

Having briefly discussed the two metaphors of speech as active sensing and speech as physical action, let us now try to answer the promised question: How (and if) can we eventually cover the whole of language through the grounding to external "stuff" approach that we have taken?

[5]For an example of mixed physical/speech act planning in a limited domain, see [Mavridis and Roy2004].

[6]There is quite some freedom of choice here, in terms of the realisation of the speech act. I can use a direct request: "Open the window", or an indirect speech act, such as: "Could you please open the window?", "Oh . . . it is quite hot in here", in order to conform with social etiquette.

[7]Inform statements can also be viewed as "changing" / "moving" things within people's heads - within their embedded GSM. Consider: "the book has fallen from the shelf". Here, no direct physical effect on the book is produced by the speech act; but in the embedded situation model of the hearer that the speaker posseses, there is a resulting movement of the book.

8.1.3 Grounding the whole of language

In Ripley, only a subset of English is effectively grounded - adjectives such as "red", "small", some action verbs etc. One obvious question that arises is concerned with the generality and the scalability of the approach: In principle - and in practice - is it possible to "ground" out the whole of language to such spaces that are external to the words themselves?

Let us see why this might be difficult. There are at least two problems that arise that we have not addressed at all:

First, consider a word such as "red". This has a meaning that can be effectively grounded to a sensory space - some aspect of the pixels of an incoming visual stream correlate with "redness"[8]. Now consider a word such as "dark", used as a modifier for "red": i.e. used in a phrase such as: "Imagine a dark red ball". The question now becomes: how can we learn the meaning of "dark" as a modifier? First of all, as what kind of object do we model it? Suppose that we have modeled red as a classifier on a color space - say RGB. Then, dark becomes an operator that acts on the classifer "red", and effectively creates a new classifier: "dark red". How can we learn this operator through empirical examples? Say that we have acquired the meaning of "red", and then of "dark red" as a totally new category - i.e. trained classifiers for both of them. Then, how can we "calculate" what the meaning of "dark" is, by comparing the classifiers of "dark red" and "red"? How can we later, having seen "green" things, but without ever having seen a "dark green" thing, be able to create a sensory expectation for it? In essence, the question is: how can we empirically learn the "compositional semantics" (the function f) that would enable us to express:

Meaning("dark red") = f(Meaning("dark"),Meaning("red"))

Also, what is the right "meaning space" for "dark" when used as a modifier for colors?

This question has never been addressed to a satisfactory extent, although it is essential for the progress of the "grounding grand plan" - extending grounding to the whole of language.

To get a feeling of the second problem, consider a word such as "loud". This seems to have a meaning that can be grounded to a sensory space quite easily - the "amplitude" of the acoustic stream should pretty much correlate with what we call "loudness". Now consider a word such as "contrast". This seems not to be modality-specific: we can say that "loud" contrasts to "quiet" or "dark" to "bright". But then, what is the right meaning space for a word such as "contrast"? If we call the meaning of "loud" or "dark" a first-order meaning (one that is directly connected to the senses), contrast becomes a "second-order" meaning - it is built upon first order meanings, and does not connect directly to the senses, as it deals with relationships between first order meanings - in this case, the relationship of "loud" to "quiet", which are both regions of an auditory space

[8] As a first approximation, "red" can be thought of as corellating to some aspect of the pixels of the objection that is being characterized as being "red". However, as is well known from results in color perception, this is *just a first approximation* - the nature and the context of the object has to be taken into account too. Such *successive approximations* of meaning that start from "absolute" views and move on to object-specific and context-dependent meanings can be accomodated within the GSM proposal, and are discussed in section 8.3.2.

Our Proposed Solution for extending "grounding" to all of language

Equip machines with:

- **a *starting vocabulary* whose meaning is acquired by "being in the world" (meaning("apple"), meaning("red") etc.)**
- **as well as with *a way to compose* meanings. (meaning("red apple") = f (meaning("red"),meaning("apple"))**
- **Then, by analogy, *extension to abstract* concepts could potentially be addressed. (meaning("justice")=? meaning("opposite")=?)**

Figure 8-3: Extending "grounding" to all of language

- which happens to be analogous with the relationship of "bright" to "dark", which are both regions of a visual space. Thus, analogy enables us to move across spaces and also build upon spaces in order to reach more abstract second-order notions, that are not directly sensory grounded anymore[9]. This is a very important process, that effectively enables us to transcend the limitations of direct sensorymotor grounding and move up towards abstract concepts. Considerable work has been done on analogy, such as [Gentner and Markman1997], however no computationally effective and general enough algorithms are really yet operational, and no implemented practical application of analogy to the domain of language grounding has yet been done - up to our knowledge.

Thus, having seen two major problems that must be solved in order for the language grounding approach to "scale" to the whole of language, and having partially hinted towards ways to address them, we can now propose a "grand plan" for grounding all of language[10]:

In terms of the above proposal, currently our community is still at the first stage, and actually in the initial stage of it. It is our hope that soon, work towards the next two stages should start in parallel, in order to enable us to get a better sense of the problems and limitations of the "grand plan" for grounding all language.

8.1.4 Levels of Detachment from the "here-and-now"

In section 1.3.2, while discussing the discrete designer-imposed stages of the development of the robot Ripley, we noticed that those were roughly aligned with the empirically evidenced progression of linguistic abilities of humans: we start with language about the "here-and-now", then extend to language about the further spatiotemporally detached, and then we start being able to talk/understand utterances about the imaginary. Here,

[9] Here, one might argue that all words have an original sensory-groundable meaning, and then we extend them through analogy - so no purely "second level" words exist. One can argue that this might be the case in terms of language development, but notice that later on the original concrete meaning is often just a small part of the extended meaning, and that it is later abandoned. Consider: "democracy" as the sensory experience of people throwing papers in boxes - this is what the meaning might start with, but then a more sensory-detached component based on multiple analogies is added upon it, and finally the physical-only component is detached - throwing papers in boxes might not constitute voting.

[10] Emotions and affect can also be incorporated into this picture - see paragraphs on affect in 9.1.5. Once we have introduced affective variables in the system, we can then treat them as sensory data arising from the "milieu interieur", and ground the meaning of words to them through the categorical layer etc.

we will try to examine some theoretical stages of detachment from "situatedness" in a little more detail, and while doing so, we will be led to the interesting topic of "imagination", which will be discussed in the next subsection.

The *first stage*, corresponds to being limited only to the "here-and-now, existing concrete things". Words connect to things directly accessible to the senses at the present moment. If there is a chair behind me, although I might have seen it before, I cannot talk about it - "out of sight" means "non-existing" in this case.

At the *second stage* ("now, existing concrete things"), we can talk about the "now", but we are not necessarily limited to the "here" - where here means currently accessible to the senses. We can talk about things that have come to our senses previously, that we conjecture still exist through some form of "object permanence" - i.e., we are keeping some primitive "mental map" of the environment.

At the *third stage* ("past or present, existing concrete things"), we are also dropping the requirement of the "now" - in this case, we also posses some form of episodic memory, enabling us to talk about past states.

At the *fourth stage* ("imagined or predicted concrete things"), we are dropping the requirement of actual past or present existence, and we can talk about things with the possibility of actual existence - either predicted (connectible to the present) or imagined.

Finally, at the *fifth stage* ("abstract things") we are not talking about potentially existing concrete things any more, but about entities that are abstract. But what is the criterion of "concreteness"? A possible proposal is the following, which is based on the discussion of the "extension to abstract concepts" step in the previous subsection: a concrete thing is a first-order entity (one that is directly connected to the senses); an "abstract" thing is built upon first order entities, and does not connect directly to the senses, as it deals with relationships between them. Take, for example, the concept of the "number three": it can be found in an auditory example ("threeness" in the sound of three consecutive ticks); it can also be found in a visual example ("threeness" in the snapshot of three birds sitting on a wire). Thus, threeness seems to be an abstract thing (not directly connected to the senses).

Currently, the GSM implementation on Ripley achieves instances of all the first four stages of detachment from situatedness; the fifth seems to still be out of reach - or isn't it? This question we will try to tackle in the next subsection - after a slight detour.

8.2 Abstract entities, Imaginary situations

8.2.1 Do abstract entities exist in a primary manner?

Even before tackling the question of the achievability of the fifth stage within the GSM proposal, an interesting question arises: do "abstract things" really exist? Do they have a primary, or secondary existence, as compared to "concrete things"? Then, in the next section, we will move on to a second version of the original important question that arises: in the GSM proposal, can "abstract things" ever be represented as a different species? Can "abstract things" populate a situation model?

The answer to the first question (do "abstract things" really exist?) can arguably not be proven empirically - but is more of a question of philosophical metaphysical belief. What are some options for answers? One can assign primary existence to one of the following three (among others): external material reality (and physical objects / events), sensory data (and sensory objects / events), or abstract things - "forms" or "ideas" in the Platonic sense. Then, depending on which of the three domains is chosen as the one possessing primary existence, the other two can be granted secondary existence (arising out of the chosen domain). Let us consider all three possible options:

The first option (*external objective reality* is what really exists) seems to be coherent with scientifically-shaped commonplace metaphysics - only atoms exist (a naïve materialist-reductionist view), and our sensory experiences as well as any abstract entities are just a consequence of atoms and their regularities.

The second option (*subjective sensory experiences* are what really exists) corresponds to a more "sensationalist/solipsist" view: external reality does not exist on its own; it is just manufactured by us for a purpose: external reality becomes just a useful conjectured model for predicting future sensory experiences: if a have the sensory experience of seeing a coke can in front of me, and the I turn my head, and turn it back (a sequence of two actions), then I can use the conjectured model of reality in order to predict my forthcoming sensory experience (i.e. seeing the coke can again). In this viewpoint, external reality might just be an illusion created by a demon that feeds our senses - pretty much the case of what is described in the opening of Descartes celebrated "Meditations" or in the film Matrix - and thus external reality is granted secondary existence, with primary existence given to our sensory experiences[11]. But what about forms in this case? Again, forms become regularities of sensory experiences - thus, derivative to them, and granted secondary existence (dependent on them).

Finally, let us consider the third option: *eternal, unchanging forms* are the only thing that "really" exists; all sensory reality is just a transient illusion, caused by imperfect corrupted and ever-changing shadows of the forms - in the platonic tradition, or similarly all sensory reality is just the veil of "Maya" in the Indian tradition - and the "Atman / Brahman" is hidden from the sight of those that have not reached the stage of awakening. In more mathematical terms, it is eternal invariants that primarily exist; all the changing phenomena are just secondary manifestations of the eternal invariants, which are hidden from direct sighting[12]. Thus, sensory reality is derivative of the forms - and external reality is again just a speculation, which carries only explanatory / predictive value - and does not have primary existence.

So, let us reconsider the starting question (do "abstract things" exist in a primary manner?), in the light of the three above stances: in a simplistc interpretation, in both the external reality as well as the sensory reality viewpoints, abstract things are secondary - they exist as derivatives / regularities of the concrete (external or sensory). Abstract things seem to have primary existence only in the platonic stance; the forms themselves seem like "abstract things". Thus, one can have a strong argument that "ab-

[11] For a similar in some senses exposition on the construction of a geometry out of the regularities of sense data, also see [Nicod1970].

[12] With the possible exception of "third eye" sighting.

stract things" primarily exist mainly within the platonic stance - and not the objective external reality / subjective sensory stance.

8.2.2 Can abstract entities exist in a primary manner in GSMs?

But how does all the above discussion align with the GSM proposal? What metaphysical stance are we silently or explicitly taking?

In our analysis of situation-viewpoints in section 3.2.1, as well as in the preliminaries of chapter 3, we explicitly accepted a version of the "external reality" stance, which however, upon closer inspection, was an "informational" and not necessarily "material" external reality - it was bits that we were referring to (when talking about the partial ordering of descriptions), and actually we were not even talking about reality itself; just about descriptions of reality[13]. Furthermore, we used this "external reality made of bits" assumption because of a utilitarian cause, and not because of any deep-rooted metaphysical stance: we just accepted this assumption, because it served the purpose of the analysis being made in the relevant chapters.

But how about the two other stances - sensory reality is primary / forms are primary? Well: notice that when looking upon the world through the "eyes" of the SCA, one can think that only sensory reality primarily exists; however, by us having pre-designed a structure for the situation model, to which all sensory data contribute, again we silently assume some form of an absolute "external" reality, but this time not a homogeneous and abstract mathematical "informational" reality - but rather, a reality parsed according to the "folk ontology" of language - i.e. made of agents, objects, body parts, properties, beliefs etc. - thus, again, this seems like a variant of an "external objective reality" viewpoint, and not a "primary sensory reality" stance.

Now, what about any hints of a platonic stance in the GSM proposal? Now, one can argue that, by interpreting all sensory data through a situation model, we are silently accepting the existence of a number of "eternal forms": the primitives that comprise the situation model: the notion of agent, the notion of property, the properties themselves. However, these are not exactly what is usually meant be a "platonic form"; and furthermore, and most importantly: say that we need to represent an abstract entity - then we must include as one of these primitives; there of course exist many abstract entities that we have not apriori been included as one of these "primitive" forms, and within our proposal we neither deal with the possibility of extension of these primitives, nor the question of redundancy among them etc. In short: the primitives we have chosen are just fixed, implied through natural languages, and cannot be extended to include other "abstract things" viewed as "platonic forms". Thus: we cannot currently see how to easily incorporate the "platonic stance" in a wide-enough way within the GSM proposal.

Thus, to sum up the main results of this discussion: We asked: "do abstract entities really exist?" Then, we introduced three metaphysical viewpoints, and answered: "Generally abstract entities are assumed to exist in a primary manner only within a platonic

[13] We actually also silently assumed the possibility of an idealized description of reality that has "infinite granularity" - and thus contains as much information as reality needs to be totally described [citeWhitmanInfiniteGranularity.

viewpoint". Then we asked: "What viewpoint are we silently assuming in the GSM proposal?" And we answered: "an external objective reality viewpoint (neither subjective sensory nor platonic); but a special kind of external objective reality: informational, and not necessarily materialist; and also, structured in terms of the folk ontology of language - thus also including mental terms. Furthermore, we do not assume this stance because of deep dogmatic metaphysical belief - we do so because of modeling convenience given our purpose". And thus the question follows: "So, does it seem that abstract entities can exist primarily, according to the viewpoint adopted be the GSM proposal?" And the answer, according to our previous discussion, become: "No". But then, the question obviously follows: "Two sections ago, we talked about language about abstract things. But then, how can this ever be accomplished within the GSM proposal? How can we represent abstract entities within situation models?". We will now deal with this question.

8.2.3 Representing abstract entities within situation models

Now, after some lengthy theoretical speculation, let us finally discuss the important practical- can abstract entities directly populate situation models? As discussed above, the position taken in the GSM proposal is that the answer to this question is negative: "No, they cannot directly populate situation models". But then, how can they be represented? An answer follows: "They can only do so indirectly, through concrete things that stand for the abstract entities by analogy or as exemplars". I.e. when thinking about the number "two", one can instantiate in his situation model an exemplar situation possessing "twoness": two red objects, or two start moving / stop moving events in sequence, or even an object having the form of the written character "2" (which has "twoness" assigned to its semantics). One can NEVER populate a situation model with "twoness" per se; only with these exemplars that possess twoness in some way.[14][15]. Thus, it seems that indeed we have not yet reached the fifth stage of "detachment from situatedness" on Ripley - but the above paragraph hopefully gives some indication of how we might be able to reach it in the future.

8.2.4 Inform speech acts, Imagine statements & the "Imaginer" module

Having supported the opinion that abstract things can only populate the GSM through concrete exemplars which form imaginary situations, we will now try to clarify the difference between "inform" speech acts (such as "there is a red thing at the left") and "imagine" statements (such as "imagine a blue thing"), in light of the GSM proposal, and also discuss the possibility of the "initialization" of the world through linguistic descriptions (the "Word"). In the next section, we will then ask the question: can "Pure Imagination", free of sensory ingredients, really exist?

---

[14]This is also the opinion assumed in Johnson-Laird's Mental Models [Johnson-Laird1983].

[15]Alternatively, and depending on the purpose, one can use analogy, and package "two" as a single sphere - maybe associating it with the color green, while association "three" with red etc.

Let us start with our first task. In section 6.4, while discussing the "imaginer" module, we stated that it has two functions:

a) Create/update objects on the basis of linguistic information, i.e. service statements of the form: "there is a small green object at the left" with the creation/update of objects within the GSM.

b) Ex-nihilo create the pre-existing objects that exist at the "beginning of time" for the robot: i.e. create the robot's model (which will be updated later when sensory information arrives through proprioceptor etc., create the human model (again to be updated later), as well as the table model. There are two points to notice here: the first has to do with what we mean by "imagination" when it comes to the current implementation of the Imaginer module, and the second has to do with our choice of including ex-nihilo creation of objects as part of the Imaginer module. Let us discuss both points here in more detail.

Let us start with the first point. Notice that the current "imaginer" services inform speech acts, that might have one of the following two forms: "there is x" (for example: "there is a blue object at the left") or "imagine x" (for example: "imagine a small red sphere at the right"). One would expect both to result in the instantiation of objects within the situation model - the first one an object having some shade of blue and being somewhere at the left, the second an objects having small size, some shade of red, and being somewhere at the right. But the crucial difference between the commonplace interpretation of the "there is x" statement as compared to the "imagine x" statement deals with sensory verification and not instantiation. In the first case, having heard "there is x", we expect that our senses should agree with the information we were given; if we look at the left, we expect to see a blue object. In the second case, having heard "imagine x", we do not expect sensory verification; an "imagine x" statement gives information about a fictional situation, not about the actual present situation that we are embedded in. Thus, the "imagine x" statement should ideally be handled by spawning a second, fictional situation model, in parallel to the actual one, and instantiating a red object within the fictional model - not expecting to have any sensory verification whatsoever. This is not the case in this first implementation on Ripley - the two statements are treated in the same way. In essence, we are giving to "imagine x" the commonplace semantics of "there is x". Of course, extensions that would enable the system to handle "imagine x" as it should be are straigthforward (via spawning of a second "fictional" situation model, which would not connect to the senses)[16].

Now, let us visit the second point. Why have we included ex-nihilo creation as one of the functions of the imaginer, and have not created a separate module to handle it, which for example might be called the "initializer"? Even before answering this question, a second question arises: why do we need to ex-nihilo create some objects at the "beginning of time"? The answer is twofold: first, the current vision system cannot perform self-imaging - there is no segmenter and no visual shape recognition support for the body of Ripley, the human body, and the table - we can only detect human faces, hands, and the objects on the table. Second, even if we had the ability to detect regions corresponding to Ripley's body through vision, how do we know they are Ripley's body

[16]See also story understanding in section 9.1.8.

to start with and not objects external to its body? The problem of body-discovery (i.e. which part of my sensory data arises from my body) is pretty hard on its own, and has been a subject of study elsewhere. This might be a direction for future extension, which however is tangential to the purpose of the system (being a cooperative SCA).

Now, having dealt with the second question that arose during the discussion of the second point, let us deal with the first question: Why have we included ex-nihilo creation as one of the functions of the imaginer, and have not created a separate module to handle it, which for example might be called the "initializer"? First, for historical reasons - this is how the system happened to be coded initially. But also, for a second reason: as a future extension, the ex-nihilo creation of the body and the environment might come through a pre-written set of verbal commands, coded in a text file, to be "heard" by the imaginer at the "beginning of time" of the system: "There is a (body description), There is a (table description), etc.". This would provide systematicity and easier modification[17].

Thus, we have so far hopefully clarified the semantic difference between "there is x" and "imagine x", and having discussed the possibility of initializing the situation model through linguistic descriptions, we will move on to a discussion of the possibility of "Pure Imagation" in view of the GSM proposal.

8.2.5 On the possibility of "Pure Imagination"

The question is: Can "Pure Imagination", where "pure" is meant in the sense of being "free from sensory ingredients", in view of the GSM proposal? The proposed answer is: "No. Imaginary objects, can only come through a recombination / interpolation (however complex) of sensory-derived fragments".

The supporting argument goes thus: Imaginary objects ("concrete" and not "abstract"), need to be able to create sensory expectations - so that their existence could be verifiable in some world, even though they might not arise in this world due to the constraints of physical law[18]. But the only way that we have to create sensory expectations within the GSM proposal is through instantiated objects within the GSM. Furthermore, the sensory expectations created, depend on the "categorical classifiers"; which in turn, arise out of exemplars which are sensory fragments.

For a more concrete example: Imagine that we trained Ripley to learn "red", "green", "small" and "large" through two examples: a "small red" object and a "large green" object. Then, we can issue a statement like: "there is a small green object" (which Ripley would have never seen before!). Then, we can show him an object - and Ripley will be able to verify whether his expectations are fulfilled - and whether the shown object was indeed a "small green" object. Thus, Ripley was able to imagine a "small green" object although he has never seen such an object before; he has only seen "small red" and "large green" objects, and through these he has learned the meanings of "red", "green", "small" and "large". Thus, the novel imaginary object ("small green" object) arose as

[17] Interestingly enough, it would also be reminiscent of the idea of creation by the "Word" that is prominent in the Abrahamic faiths: "And He said: Let there be light! And there was light ... ".

[18] Or through the fact that they have not been reached yet or are unreachable by evolution - consider the celebrated "Unicorn" case.

a recombination/interpolation of previous sensory-derived fragments - supporting our stance. And through the above argument, it should have become clear that no imagined situations can have non-sensory derived ingredients; the recombination might be novel, but the materials came out of the senses[19]

Now, having explored the big picture for a while, journeying through meaning spaces, then moving on to the theoretical unification of speech with sensing and acting, and then finally arriving at the "grand plan" for grounding all of language, and discussing detachment from "situatedness", abstract entities and their representation, and the possibility of "pure imagination", let us now move from the general to the more specific and discuss the semantics of adjectives and verbs within the GSM proposal, successive approximations of meaning models for sensory meaning spaces, as well as the grounding of comparative terms, before we move to the "big picture" again and discuss similarity metrics for situations and the viability and benefits of enumerating all possible situations.

8.3 Semantics, Situations

8.3.1 The semantics of adjectives and verbs within the GSM proposal

Having examined the GSM proposal, one reasonable question for a reader to ask would be: what are the meaning models of adjectives and verbs within the GSM proposal - for example, where and in what form are the semantics of "red" or "touch" represented within Ripley's code? And, what should one change, in order to change the semantics of these? Let us now try to answer this important question.

Notice here, that within the presented processes, two types of "black boxes" have been mentioned: *"categorical classifiers"* and *"event classifiers"*. The first are fed with an n-D vector, and produce a category number, while the second are fed with sequences of $MOMENT$s, and produce $EVENT$ frames (of the form described in section 5.1.4). A third type of black box is part of the motor control subsystem (see figure 5-16) in section 5.3 - which we will here term *"action controller"*. This is fed with a set of action parameters, and activates an action control loop.

These three boxes essentially encode the potentially directly trainable part of meanings of adjectives and verbs within the GSM proposal. The meaning of adjectives then corresponds with the corresponding "categorical classifier"[20], while the meaning of verbs decomposes into two parts: the recognisive part (corresponding to the "event classifier" - see section 5.2.3) and the generative part (corresponding to the "action controller"). Thus, using our previously developed terminology of "meaning spaces", for the GSM

[19]In the case of abstract objects, those should have again primarily arisen out of regularities of sensory objects, and then maybe generalized from their sensory-modality bindings - as in the case of "contrast" discussed in section 8.1.3. Thus, they too, have arisen as higher-order byproducts of sensory material - and the above argument still holds. Furthermore, as mentioned before in section 8.2.3, abstract objects can only populate the GSM indirectly, through specific situations - and thus again we remain through the sensory-expectation realm, in which case the formed expectations again come out recombination / fluid mixing of originally sensory-derived material.

[20]See L1-to-L3 in section 5.2.2.

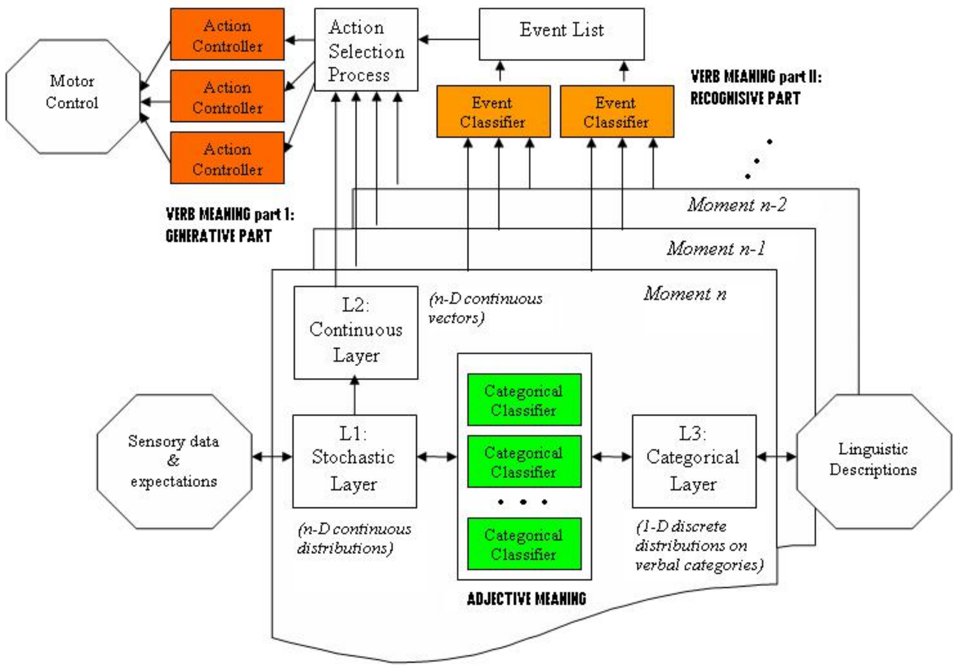


Figure 8-4: The functional position of the three black boxes corresponding to the meaning of adjectives ("categorical classifiers" - green) and verbs ("event classifiers and action controllers" - light and dark orange) within the operation of a GSM-based SCA. Action selection and reference resolution details as well as other components omitted.

proposal our model of an adjective's meaning lies within a "sensory space", while the model of a verb's meaning decomposes to two parts, one of which within a "sensory space" and one within a "sensorymotor space".

But one might ask: is it enough to assume that the model of an adjective's meaning is only its categorical classifier? Well, it depends. If one wants a more complete model of meaning, then for the case of adjective meanings, apart from the categorical classifier, other parts of the ExternalReality-to-SituationModel pipeline (section 3.2.1) should be included too - for example, the property extractor (P1b - the process whose input is a segment of the sensory stream and whose output is property vectors - for example: for the case of "color", the process that, given the sensory stream that corresponds to a segmented region, to calculate the color of the object) etc.). Ideally, for wider generality, these processes should be learnable too, and not hard coded.

8.3.2 Successive approximations of meaning models

In Ripley's implementation and our GSM proposal, we have assumed that there exists a classifier that helps map the stochastic layer into the categorical layer - the so-called "categorical classifier"[21]; as commented upon, this classifier is crucially implicated in the sensory grounding of property category adjectives. For example, consider the one-dimensional property "size": this classifier accepts continuous values (such as "3.141")

[21] See section 3.2.1.

and outputs a category where the value would belong (for example 2="large"). In the above approach, it is assumed that "small" (or any other such property label) has an absolute meaning - its categorical boundaries are fixed, irrespective of what noun is being modified by "small". We will call this approach F1: the first-order approximation to meaning, having fixed absolute thresholds (or regions for multi-D) for the classifiers, irrespective of the noun that follows. This threshold might have been arbitrarily chosen, or tuned through human experimental data, or based on a statistical argument such as: one sigma above mean value should be called "large", one sigma below "small" etc. Thus:

F1 (*absolute*) 1st order approx:
category = f(property value)

However, this is only a first approximation of the mapping between value and category label; for example consider the following case:

"small elephant" vs. "small ant"

It is clear that the categorical boundaries of "small" for elephant, would be different than those that would hold for an ant. Indeed, the size of a small elephant is probably the size of a supernaturally giant ant. In order to deal with this, we need to extend the first-order approximation to second-order: one way to do so, for the case of "size", is to keep track of noun-dependent property statistics - i.e. keep separate histograms of sizes: one for elephants, one for ants etc. and tune (μ,σ)-dependent thresholds, or empirically acquire noun-dependent classifiers: one for small elephants, one for small ants, etc.[22] This we will call the second-order approximation F2, the noun-dependent case. Thus:

F2 (*noun-dependent*) 2nd order approx:
category = f(property value,noun)

But even this might not be sufficient in some cases. Consider the following:

"the small cup" when surrounded by other fairly large cups vs. "the small cup" when surrounded by miniature cups

Here, we need to take into account not only the noun that is modified by the adjective, but also the situational context - the other cups around it. In that case, the model of "small" should be dependent on the sizes of the other cups currently around it, i.e.:

F3 (*situationalctxt-dependent*) 3rd order approx:
category = f(property value, noun, values of same property for other relevant objects)

Again, this is just one more approximation, and one could envision extending towards other directions too. Also, the above apply to other property dimensions too, such

[22] But what about the case where one has not seen an ant before but is asked to imagine an ant, or where he has only seen one and can not estimate the variance? When encountering a new "noun", i.e. while still learning the models, one might transfer assumptions from previously seen things until he gathers enough experience with the new category.

as "color" and so on - consider "red apple" vs. "red sky", and "the red one" among different sets of objects. Here, for simplicity we have chosen to deal with the first approximation only - however, one could easily extend to the others too, in the future[23]. It might well be the case that also during human development a similar trajectory across successive approximations is followed - a conjecture that remains to be tested by empirical evidence.

8.3.3 Comparative adjectives and better spatial semantics

In our implementation as well as in our proposal, the "position" property is treated homogeneously with other properties, such as "color" and "size", and as all others, is only given a first-order approximation of meaning (F1-absolute meaning). However, although this makes our theory neater, in practice this might sometimes differ from common language use. For example, although when we are faced with a table where objects stand, we do use phrases such as: "give me the ball that is at the left" with absolute meaning, we quite often also use relative terms, such as: "give me the one that's to the left of the green one". The question now arises: how can we model the meaning of such terms, like: "to the left of the green one", "bigger than the red one" and so on?

We will call these terms *relative terms.* A model providing a first approximation to their meaning within our framework is the following:

To evaluate a relative term R:
Create an $OBJECT\_RELATION$
among the two $SIMPLE_OBJECT$s A and B that are implicated,
and encode the vector difference between A and B in the relation.

R will thus encode the vector difference between A and B, in the following manner: The stochastic layer of R is then fed by the stochastic layers of A and B by effectively calculating $P(x_A - x_B = x)$ given $P(x_A = x)$ and $P(x_B = x)$. This can be done by iterating through all (x_A, x_B) pairs, calculating the difference $x_A - x_B$, and assigning $P(x = x_A - x_B)$ in the stochastic layer of R to be equal to the sum of all $P(x_A)$*$P(x_B)$ from the stochastic layers of A and B, summed across all those pairs that have difference equal to x[24]. Then, we project from stochastic layer to categorical in the usual way (L1-to-L3 in section 5.2.2. Now, the categories will effectively be "A bigger than B", "A the same as B", "A smaller than B" for sizes, or "A to the left of B", "A under B" and so on for positions.

Thus, whenever such relative terms across two $SIMPLE_OBJECT$s need to be evaluated, one can create an $OBJECT\_RELATION$ and evaluate the right properties in the above way. This has been demonstrated for simple cases in our implementation.

[23] In terms of the structures described in chapter 5, such an extension requires some modifications that however can be sustained within the proposed framework: for example, in order to extend to F3 the categorical layer cannot be fed only by the stochastic layer of a single object anymore, but has to be fed by the stochastic layers of other objects too.

[24] For example: in order to calculate the probability of the difference being equal to 5, I take all pairs (x_A, x_B) that have $x_A - x_B = 5$, and sum up their probability products $P(x_A)$*$P(x_B)$.

Now the question arises: if the two $SIMPLE_OBJECT$s exist, we can use the above method. What can we do if we need to service an inform statement that involves object creation through a relative term, such as: "There is a red one at the left of the blue one"?

In that case: assume object A exists (the blue one). Create object B (the red one), as well as an $OBJECT\_RELATION$ R. Run the processes Words-to-L3 and then L3-to-L1[25] for R: project "at the left of" as a category to a distribution on position differences for A and B. Finally, now on the basis of known $P(x_A = x)$ (encoded in the stochastic layer of $SIMPLE\_OBJECT$ A) as well as known $P(x_A - x_B = x)$ (stochastic layer of $OBJECT\_RELATION$ R), calculate the unknown $P(x_B = x)$, through a process similar to the one described above (take all relevant pairs etc.).

Similarily, one can deal with "There is a red one that is bigger than the blue one" etc.

Thus, we have seen a way to service relative term-based descriptions. Now, we will ask the question: even augmented with relative terms, how realistic are our spatial semantics? They seem to be enough for simple table-top tasks with Ripley the Robot. However, for a more empirical-data-driven approach, the interested reader can refer to [Regier and Carlson2001]. Such models, dealing with non-symmetrical objects that are not viewed as point masses, could also be incorporated in our framework in the future.

8.3.4 Similarity of situations

Any successful attempt towards generalization from specific instances to others, requires some measures of "similarity" or "distance" across situations. Also, different aspects of situations need to contribute more in such measures, depending on the purpose of the measure. The multiple levels of information encoded in the situation type proposed in this thesis enable the creation of a number of different such measures, depending on the purpose.

Let us try to be more explicit through an example. Imagine that a robot is learning to take appropriate actions given a situation, by observing a human's actions when he is exposed to a situation. For example, imagine that whenever a human sees a red object and a green object on a table, he puts the red object in a position where it is to the left, and touching, the green object. The robot has seen some examples of initial situations, and some examples of the subsequent actions of the human. Now, suppose that the robot is faced with a new situation. How can the robot decide whether the new situation is similar enough to the initial situations of the examples it saw, in order for it to be appropriate for taking an analogous action? How should it match the roles of objects in the situations it saw to objects in the new situation?

Let us suppose that we assign three roles to the objects partaking in the situations:

the "ToBeMoved" role (T-role, for the red one in the examples),

the "LandMark" role (L-role, for the green one in the examples),

and the "Indifferent" role (I-role, for all other objects that might have co-existed with the green and red ones in the examples).

[25] see section 5.2.2.

Here, the multiple levels of information, enable multiple hypothesis to be generated and tested. For example: what is important for an object to be assigned the T-role? First, it might be the unique identity of the object (not any red object will do, but only one specific red object might have played the T-role in all examples). Second, it might a specific property of the object (its color, its position), considered up to multiple levels of granularity: at the linguistic discriminability level (where two different shades of "red" count as being the same), or at the continuous or stochastic level (where deterministic or probabilistic distances count) etc. Third, it might be a relation of the object in question with other objects (for example: the object that has the property of being on the left of another and on the right of another).

Thus, we can assign distance/similarity measures, and generate and test hypothesis along all these levels, which are represented readily in the GSM. Of course, the combinatorial explosion of the matching problem is not solved in this way; however, notice that in most cases, at least for the role of cooperative helping hand robot like Ripley, situations consist of a small number of objects, so the problem remains tractable. Even when this is not the case, as is often the case in the real world, the relevant objects and objectifications are explicitly or implicitly introduced in the beginning of the demonstration, and clarifying questions can always be asked - and thus what has to be taken into account is partially explicated.

Now, having briefly discussed similarity across situations, and matching / assigning roles in situations, let us move to another aspect of situations: enumerability.

8.3.5 The Enumerability of Situations

In this section, we will consider the possibility of enhumerating all possible situations for a given GSM and a given number of objects. By doing so, we will be able to transfer from the domain of words to the domain of integers - and thus meanings of words can effectively be made to correspond to statements about integers, such as divisibility, remainders and so on. Finally, the metaphor of language as a "camera" performing projective transformations on situations will be briefly introduced, together with a discussion of the implicit transmission of the speaker's "viewpoint" when performing partial descriptions.

First of all, consider the contents of the categorical layer at a single moment. These are essentially a set of discrete 1-D distributions corresponding to each object / relation, one distribution per property. Still, they contain more information than is lexicalized: as explained in the "inquirer" of section 6.4, a decision tree chooses between a set of different lexicalisations, such as: "red" / "most probably red" / "either red or green" / "I have no idea" - which is depicted in figure 6-11. At this level, after we have thrown away the specifics of each categorical distribution and kept only the information necessary for lexicalization, the set of possible states of the world becomes enumerable - and furthermore, for the case of a fixed pre-known number of objects/relations inhabiting the situation, it also becomes) finite. For example, consider a toy-GSM with two objects (O_1 and O_2), each of which has two properties (size and color), each of which has two categories (small/large and white/black), and where only three gradations of uncertainty are distinguished ("white" / "probably white" / "I have no idea"). This situation can only be in one of the following possible states:

$$\{O_1 \in \{"small","probablysmall","noidea","probablylarge","large"\} \times$$
$$\{"white","probablywhite","noidea","probablyblack","black"\}\} \times$$
$$\{O_2 \in \{"small","probablysmall","noidea","probablylarge","large"\} \times$$
$$\{"white","probablywhite","noidea","probablyblack","black"\}\}$$

Thus, language can only distinguish between 5x5x5x5 = 625 such situations, a finite result. If we allow n objects, this figure will become 25 to the n-th power, still finite. Thus, even for possible infinite n, we still can enumerate uniquely the resulting situations (i.e. bring them into a 1-1 correspondance with the integers). This can be done easily be assigning a unique "serial number" to each possible situation in the following way:

Let $(Siz(O_i), Col(O_i))$ be the state vector of each object, where:

$$Siz(O_i) = \begin{cases} 0 & ("small") \\ 1 & ("probablysmall") \\ 2 & ("noidea") \\ 3 & ("probablylarge") \\ 4 & ("large") \end{cases}$$

$$Col(O_i) = \begin{cases} 0 & (if"white") \\ 1 & (if"probablywhite") \\ 2 & (if"noidea") \\ 3 & (if"probablyblack") \\ 4 & (if"black") \end{cases}$$

Let the integer state of each object be:

$$St(O_i) = \begin{cases} 5 \cdot Col(O_i) + Siz(O_i) + 1 & (\exists O_i) \\ 0 & (otherwise) \end{cases}$$

And finally, let: $SerialNumber(Situation) = 2^{St(O_1)} \cdot 3^{St(O_2)} \cdot 5^{St(O_3)} \cdot 7^{St(O_4)} \cdot \ldots$

(I.e. the primes to the power of the state of each object O_i.)

Thus, even in the unspecified number of objects case, one can always enumerate the possible situations, distinguishing them up to the level that language does. One can also do a similar trick for the case of events, and enumerate them too.

But what is the utility of such an enumeration? First, it provides a very compact representation of the states of situations or sequences of states. Second, let us try to examine how words correspond with the serial numbers that represent the situations:

(Let S be the serial number of the situation).

"Is the first object certainly black?"

translates to the statement:

In the prime decomposition of S (abbreviation: pd(S)), is the power of the first prime (i.e. 2) minus one such that the integer part of itself divided by 5 is equal to 4?

More generally:

"Does the i-th object have Siz = j and Col = k"

becomes:

In pd(S) is the (power of the i-th prime) minus one such, that:
- the remainder of itself divided by 5 is equal to j, and
- the integer part of itself divided by 5 is equal to k?

Thus, under the viewpoint of serial numbers of situations, each property "partitions" the situations, with one partition corresponding to each category. For example, consider the set of situations with maximally 2 objects. We can partition them according to the size of the first object: the 625 become 5 groups of 125 situations. Similarily, we can partition them according to the size of the second: a different partition of 5 groups of 125. We can now also view the statement: "one object is black" in that way: here, we get a dichotomiser - there are 9/25*625=225 cases where the statement holds and another 400 where it does not.

Thus, in this new viewpoint, we have the following correspondances:

- Statements about the situation (example: "one object is small") become dichotomisers (a set of situations where the statement holds and a set of situations where it does not)

- Each property (example: size) partitions the situations into n parts, where n is the number of category-certainty combinations for that property

- Each property category-certainty combination (example: "probably small") selects one of the above partitions, and unites the remaining partitions, thus forming a dichotomiser

- A description consisting of a string of category-certainty adjectives (example: "small black"), corresponds to the intersection of the dichotomisers that correspond to each of the adjectives

Language as a camera

In this sense, language is similar to a "camera projection" of situations. A camera projection exposes a single "view" of the 3D world, providing only partial information about it, which is dependent on the view taken. The projective transformation is not invertible - and only through a series of multiple views can one uniquely reconstruct 3D reality from 2D views. Now consider each question such as: "What color is it?" as a view. Only through multiple views such as "What color is it?" / "What size is it?" and so on can we reach finest-grained specificity[26] about the situation. Else, we only have information up to a "co-set".

[26] Up to the level offered by language.

Also, every time that we are offering a partial description of a situation, such as: "there was a red ball and a blue square on the table", we are also informing the hearer about our (conscious or not) choice of viewpoint: here, the speaker's "language camera" took a viewpoint through which only color and shape were visible - size was ignored. Many times, this not explicitly transmitted information (the "language camera" viewpoint) might be even more interesting and useful than the explicit semantics of the utterance - such viewpoints can offer an insight on aspects of the personality of the speaker on the long turn, or more practically and in the short term, on his current goal (what he has decided to mark linguistically shows me what he has noticed, which shows me what he is looking for, which tells me why he might be looking for this - what he might want to do).

Thus, we have seen how one can enumerate situations and assign "serial numbers" to them, and how statements and words then resolve to partitions / sets of integers, properties of integers etc. We will not explore this direction further in this thesis - this was just a diversion towards an area that might or might not become fruitful. Finally, in the light of our discussion, we introduced the metaphor of "language as a camera", and talked about the implicit transfer of the speaker's "viewpoint" through his choice of partial descriptions.

8.4 Logic, Ontologies

8.4.1 Bridging the categorical layer with FOPC

In our proposal, the categorical layer contents as well as the event structures contain the information that is directly connected to language. Most traditional NLP systems, usually represent meaning using FOPC (First Order Predicate Calculus) statements, such as:

$$\exists i, e, w, t : ISA(w, Moving) \wedge Mover(w, Object1) \wedge IntervalOf(w, i) \wedge \\ EndPoint(i, e) \wedge MemberOf(i, Now)$$

The above statement would roughly translate to: "Object1 is moving". The question that naturally arises is: is there a way to translate the information contained in the GSM to such a representation using FOPC, in order to take advantage of existing algorithms / code bases? Let us discuss translation from FOPC back and forth to categorical layer / event contents. In the previous section we have shown that situations and events become enumerable, if we decide to quantize the gradations of uncertainty (create categories), in the manner that natural language does. Thus, at that level, where situations become totally discretised, we can try to translate to FOPC and back. We just need to write translation rules, to translate between situations and FOPC, and between events and FOPC. Then, we will be able to effectively do:

Objects O1("small", "no idea") and O2("probably large", "black")

$\Longleftrightarrow$

(equivalent FOPC statement)

$$\text{Event(type = started moving, participant1 = agent1object2(="Ripley's hand"),}$$
$$\text{participant2 = none, start\_time = 5, end\_time = none)}$$
$$\Longleftrightarrow$$
$$\text{(equivalent FOPC statement)}$$

Of course, doing so is not trivial. One needs to make a number of design choices, including how to express gradations of certainty in FOPC, how to encode beliefs (for embedded GSMs) in FOPC, what type of temporal logic to use etc. However, this is a well studied subject, and the interested reader will be able to find many useful sources. For example chapter 14 of [Jurafsky and Martin2000] provides a short introduction, and also contains descriptions of many useful standard algorithms that operate on FOPC semantic representations, and can thus be combined with GSM's through this translation.

8.4.2 Grounding ontologies

In the first chapter, while discussing the shortcoming of current NLP approaches towards meaning, we mentioned semantic networks such as "wordnet" [Fellbaum1998], which can also be viewed as a special type of an "ontology". The question now arises: indeed, when a semantic net is left on its own, any definition of meaning within it becomes circular. But what if we are able to "ground out" to some external meaning space some of its nodes? Could there be any use to such a system? It turns out, that there can. And actually on the way to doing so, another question arises: It seems feasible to ground out the nodes (vertices). But what about the existence of the relations (edges) then? After all, all the information that is encoded in wordnet, is contained on the connectivity of nodes to edges, not on nodes on their own - thus, in order to have something to gain from wordnet, after having grounded some nodes, we need to somehow "translate" the edges into something that can inform us about the meaning space of the nodes. Thus, it turns out that we effectively need to "ground the edges", i.e. ground relationships such as IS-A, HAS-A and so on, in order to have something to gain from "grounded ontologies". Some more early thoughts on this interesting diversion can be found in [Mavridis2006a].

8.5 Stance X and its relation to GSMs

8.5.1 Embodiment and GSMs

Two important questions that we have not touched upon explicitly yet are the following: Q1) where is the "body" within the GSM proposal? And Q2) how easy is it to move across bodies? Let us try to briefly address these important questions here.

So, let us first deal with the first question: where is the body within the GSM proposal? We propose that it appears in at least three different ways (look at figure 8-5):

First, the actual physical body appears directly in the physical / informational boundary between the agent's mind and the physical world - i.e. the red areas of figure 8-5, which reflect in the models M1 (ExternalReality-to-Sensor model) and M2 (Action-to-ExternalReality model) as defined in chapter 3.

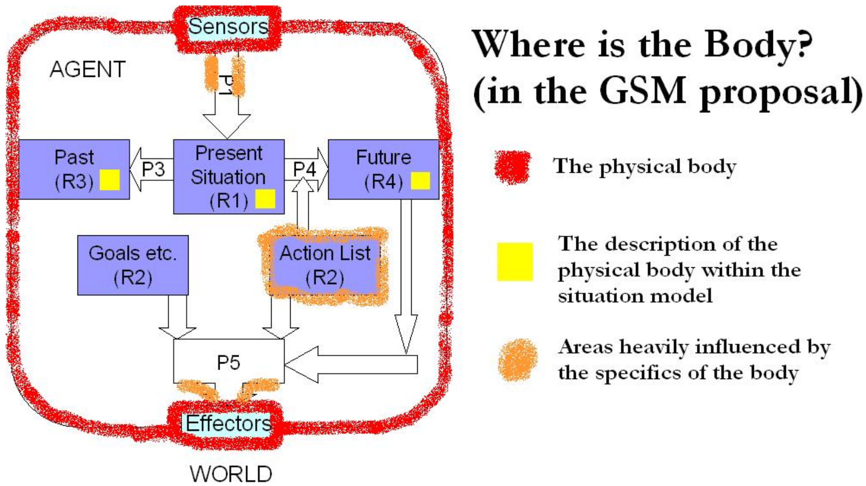


Figure 8-5: The body within the GSM-based Agent Model: the actual physical body (red), the representation of the body within the situation model (yellow), areas heavily influenced by the specifics of the body (orange).

Second, a description of the physical body of the self appears within the situation model type, which is used for both past, present, as well as future situation descriptions - the yellow areas of figure 8-5. (see $AGENT$ body in section 5.1.1).

Third, there are some areas of the agent model that are heavily influenced by the specifics of the body of the agent: for example, the action list (R5), the first stages of the sensor-to-situationmodel process (P1) as well as the last stages of the action selection process (P5) (see chapter 3. These areas are marked with orange in figure 8-5.

Now, having answered the question of where the body appears within the GSM proposal, let us turn to the second question: how difficult is it to change embodiments? The answer to the previous question can be illuminating. When changing a body (i.e. changing the red area - effectively also M1 and M2), we must also: First, change the physical model of the body which resides within the situation model (yellow area). Second, change the action list R5, the last stages of the action selection process P5 (to comply with the new action list), and the first stages of the sensor-to-situation model process P1 (to comply with the specifics of the new sensors) - i.e. change the orange areas.

Thus, we reach two conclusions: First, in order to quickly plug-and-play with a GSM library, any physical body should come together with an accompanying "body customization" software package, which should include: a model of itself to be placed within the situation model, an action list, and gluing stages for P1 and P5. Optionally, this software package can also include models M1 and M2 for theoretical purposes or for ToM of others with the same body. The second conclusion is: there can be significant reusability of GSM designs and parts irrespective of the body, as long as we have accompanying "body customization" packages for different embodiments.

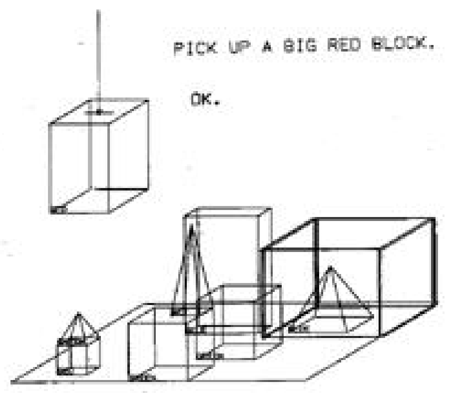


Figure 8-6: A snapshot from SHRDLU's operation

8.5.2 SHRDLU and GSMs

Many more technically oriented visitors, when exposed to a system like Ripley, often get reminded of a milestone system in the history of AI, that had the weird-sounding name "SHRDLU". SHRDLU is a program for understanding natural language, written by Terry Winograd at the M.I.T. Artificial Intelligence Laboratory in 1968-70. SHRDLU carried on a simple dialog with a human user through a teletype, about a virtual world of idealized objects shown on an early display screen. The important question now becomes: in what ways is the GSM proposal different or similar to SHRDLU?

The most important difference should be obvious: SHRDLU exists in a virtual world, while the GSM-based Ripley exists in the real world. This difference has a number of very important consequences: First, in SHRDLU's virtual world, there is no need for sensing; the contents of the world are continuously and totally accessible at the object and property level; i.e. there is no ExternalReality-to-SituationModel process (see section 3.2.1) in our terms - no complications with vision, segmentation, property extraction and the such. The Situation Model IS the only Reality in SHRDLU's case - and this reality is totally accessible and not partially hidden or noisy or only partially visible at the time. Then, we can easily strip off the stochastic layer (section 4.2.2) from a situation model from SHRDLU - we just need continuous and categorical. We can also strip off all the "objecter"-related (section 6.4) object permanence code - the omniscient "eye" of SHRDLU knows all objects with their unique IDs, and they will never disappear from sight. Furthermore, there is no concept of "sensory expectation" and verification of sensory expectation - it can never be the case that reality does not match its estimate (the situation model), because reality IS the situation model. Also, there is no learning whatsoever of semantics of words; "red" is not attached to any sensory space semantics through examples, "pick up" is never taught by demonstration. Thus, there is no "imagination" in the sense used in this document too - no arbitrary recombination of sensory-derived elements (section 8.2.5), no sensory expectation formed that can be verified or disproved[27]. Finally, to close a small part of a non-exhaustive list of

[27]On the positive side for SHRDLU, it contains a planning system that is also able to attribute causes

differences, actions in SHRDLU always have their predicted effects - another major simplification, and SHRDLU's model of the world does not contain the user or the self and does not have embedded models.

When it comes to the similarities, some are obvious too: both perform some sort of "understanding" of natural language and engage in dialogue with the user about a world (virtual in SHRDLU vs. real in RIPLEY). In both cases, one can see visualizations of the beliefs of the system about the state of the world through a 3D rendering. Both systems keep sequences of events and so on. But apart from such generic similarities, we have seen that the differences are indeed numerous and great. SHRDLU was an amazing system in its times; but in order to get to the next level, and build SCAs in real environments, we believe the GSM proposal is crucial.

8.5.3 Observer-only sensory meaning spaces and GSMs

Here, we will consider the following question: Is there an important qualitative difference between systems that rely on observer-only sensory meaning spaces, such as the video event grounding system presented in [Siskind2003], and systems that are both observers and actors, such as the GSM-based Ripley robot?

First, one must note that in systems such as [Siskind2003], we are in one respect one level closer to Ripley than SHRDLU was: we are not in a purely *virtual world* anymore, but we are actually processing video input coming from the *real world*[28]. However, we cannot act on the real world; just see. Some of the consequences of this difference include the following: First, our *sensory expectations* are more accurately just *temporal predictions*: we do not expect to sense X if we do Y (as we cannot do anything but wait), but we can only expect to sense X after waiting for time t. Thus, we are never learning input-output models; just input-after-time-t models. Second, generally systems like [Siskind2003] usually assume continuous (even if partial) visibility of objects; thus, a part of the need for *object permanence* (the part which arises because of the movements of the "eye" of the robot), does not exist anymore. Third, we usually can never acquire the *first-person* meaning of motor verbs: for example, we can only perceive events related to other-person objects "moving" or "falling" - as the camera is static, there is no experience of first-person "moving", "falling" etc. However, aside from these limitations, there exist observer-only systems that have reached amazing results within their space of capabilities - most notably, the highly elegant and sophisticated approach of [Siskind2003] and its continuation in Siskind's later work.

8.5.4 Procedural semantics and GSMs

One viewpoint that is often taken when one discusses the meaning of words that are comprehended by an AI system, is that of procedural semantics - in the words of Terry Winograd, the creator of the historical SHRDLU system:

to actions, and talk about them, which Ripley doesn't. However, as will be discussed in section 9.1.5, this would be a straightforward extension for Ripley.

[28] Which is more clearly the case in Siskind's later work.

One of the basic viewpoints underlying the model is that all language use can be thought of as a way of activating procedures within the hearer. We can think of an utterance as a program - one that indirectly causes a set of operations to be carried out within the hearer's cognitive system.

This viewpoint, if translated to a GSM-based design such as our implementation of Ripley, reduces to:

Q: What is the meaning of words for Ripley?

A: The meaning of a heard utterance can be identified as the sum of the processes that will be initiated when it has been received by the system (a procedural semantics view). These processes are in turn also determined by the current state of the system, and are thus in turn determined by its structure, its training, and the recent heard utterances ("legomena") and perceived events ("symvanta"). Thus, words are connected indirectly to the world, through these processes, the central hub of which is the mental model of the situation (i.e. the Grounded Situation Model - GSM).

Although the above might be a truthful statement, it is not very helpful in terms of deriving models of meaning. One can try to enlist the processes that are elicited when specific words are included in an utterance, but due to the dependency on the coexistence of other words in the utterance and also on the current state, the answers that one would get will become giant branching-structures with lists of processes, that might not be so informative, and thus although this path exists it has not been followed here.

8.5.5 A proposal for Empirically Sound Semantics

Another path that can be followed towards devising a theory of semantics, is the decomposition of the meaning of a generic action into "production conditions" and "reception effects" - which creates the possibility of generic empirically derivable semantics through the long-term observation of agents that are just acting and interacting. This approach I introduced in [Mavridis2005d]. Let us first see what the approach is, and then how it relates to other models of semantics and to the GSM proposal.

What is the "meaning" of an action, which is not necessarily a communicative action? First, what is a solely communicative action? Here I will adopt the view that a solely communicative action is one that has negligible direct physical effects - the effects it accomplishes are always mediated through others. For example, saying "open the window" has negligible direct physical effects: some air molecules moved. However, the indirect human-mediated effects might be substantial: the hearer might open the window. Thus, the action of saying "open the window" is a solely communicative action by the above definition. On the other hand, the action of me physically opening the window with my hands (and not saying "open the window") is not a solely communicative action. Of course, there almost always is informational content to any action; somebody observing it might gain information from it, for example he might understand that I am feeling hot; also, I might have used it as a secret sign - but nevertheless, there is a direct physical effect too - thus, opening the window, is not a solely communicative action[29]

[29]Note that as mentioned above, any action, even the ones that conform to the above definition of being "solely communicative" has a physical effect too - when I say "open the window" I am creating sound

Before we tackle a proposal for the meaning of a generic action, let's see how classical semantic theories tackle the meaning of utterances. A brief glance through semantics theories for spoken languages will expose many varieties: "realist" semantics trying to associate words with things in an objective external world, "mentalist" versions associating words with mental states, third-realm proposals such as Frege's etc. Most of them have at least one thing in common: they pre-suppose a model that is pretty much language like, and they connect for example the word "Nick" with an externally existing entity that is called "Nick". Thus, we presuppose a language-like meaning space, in order to assign meanings to it. But how can we have minimal assumptions about the meanings of signs? If my friends and I create a secret code of signs, and we only use it within the context of a specific plan, and it happens that although I can try to explain the meaning of the signs using natural language, they only have a within-plan meaning that would better be described for example mathematically, then the assumption of natural language as a good language for describing meanings of signs does not hold anymore. And anyway, in many cases (as in the robot language learning case), as we have seen before, we cannot presuppose the existence of a human interpreter reading the descriptions of the meanings of the signs in natural language. Thus, the problem remains: How can I get empirical models of the meanings of signs, that would not resolve to natural language statements, and which can be acquired by pure observation of a set of agents that act and interact over a period of time? Let us unpack thus question, and get some more precise requirements for what should be considered a satisfactory answer:

The central question is:

What is a meaningful definition of the "meaning" of an action, which definition fulfills the following three requirements?

R1) It is empirically sound. Meanings defined in this way should be estimatable in a computational manner, on the basis of external observations of a set of agents acting and possibly interacting. Note that this is in total contrast to the way most current semantic theories construct meanings: they rely on the manual encoding of meanings - meanings are usually just written down by human semanticists, after they have introspected on what they think the meaning of a word should be.

R2) The meanings should not be defined on the basis of a set of predefined high-level natural-language "concepts". For example, we should define meaning like a traditional semanticist - we should not be satisfied by saying that the meaning of the utterance "John ate the pizza" in model-theoretic semantics or FOPC is concerned with the individual "John", the object "pizza", and the action of "eating" which are supposed to be well-defined notions (which notions, by the way, only humans can understand - and not machines, and also humans often without agreement among them). Rather, the meaning should be defined in terms of an agreed upon

waves etc. Thus, as an alternative definition, instead of requiring "negligible" physical effects in the above definition, one could require it to have no intended physical effect - and in this way the definition becomes relative to the actor's intentions, which might or might not serve as a good definition, depending on our purpose of use of the term "solely communicative".

description of externally observable states of the world, which world is modelled in some non-language like model (for example, a physical model).

R3) The meaning should contain both a internal/mental as well as an external/physical part. Of course, the actual content of both of these parts depends on our clearly explicated choice of models: we first choose a "mind model", then we choose a "world model", then we observe actions and interaction of agents, which we describe in terms of the "world model", and then we can estimate "meanings" in terms of either the conjectured "mental component" of them (how they connect with mental states, given the assumed mind model and how it connects to the world) or the directly observable "physical component" of them.

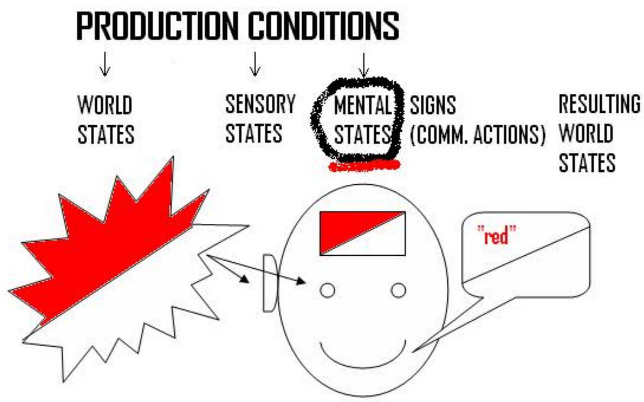
In order to tackle this problem, I will here try to take a stance that is simplistic but which I believe to be concrete and most importantly empirically sound, which I have introduced in [Mavridis2005d]. The "meaning" of each action will consist of three parts: "*production conditions*", "*direct physical effects*", and "*reception effects*". Solely communicative actions (which can be viewed as signs) have no direct physical effects - thus they are bipartite. Production conditions are those internal states of the producing agent that will lead to the sign. Of course, these correlate with the sensory history that the agent has been exposed to so far. On the other hand, reception effects are the changes in action production in the future, compared to the actions that the agent would produce if he had not received the sign[30].

Production conditions

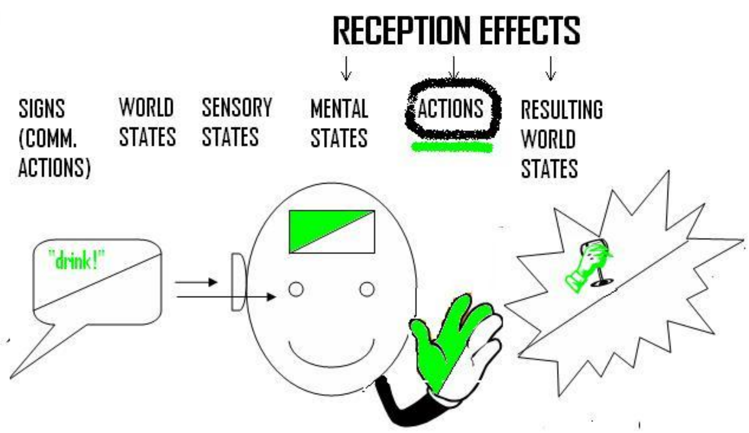
Thus, whether the sign is produced at a given instant depends on two things: the part of external reality in the temporal interval between the agent's birth and the possible production of the same that has been accessible to the agent, and the innate specifics of the agent himself. I.e. in folk terms: whether the sign is produced depends on the character of the producer, as well as what he has experienced so far. I.e. the sign always fundamentally signifies something about external reality in the past, and usually the recent past, and in particular the way that local external reality has affected the producer. For example, when I see the train coming I might say to you: "The train is coming". When I tell you: "Please give me the pen" conventionally (and usually) this means that the current situation has made me desire it, I have seen that you are a helpful person, and I have decided to try to exploit the social affordance that you present to me[31]

[30] One can also talk about the changes in the probability distributions of these actions.

[31] Notice that here we are not excluding the possibility of lies or unconventional language generation - for example, due to some abnormality in the speaker's brain. But still, even in that cases, the sign does tell us something about the speaker's mental state primarily, and secondarily, about the state of external reality. Production conditions for a given action are not required to be universal across individuals (agents producing the sign/action), and thus in our definition given later, we define ProductionConditions(sign,agent) at a given time.



(a)



(b)

Figure 8-7: Production Conditions (A) and Reception Effects (B): Production Conditions as a dichotomizer on mental states that result to the production of the sign (and consequently also dichotomizing the world states that caused the mental states), Reception Effects in terms of the change of future actions given the reception of the sign, as compared to not having received the sign

Reception effects

On the other hand, the overt part of the reception effects will unfold in the future, when the receiving agent modifies his actions due to the fact that he has received the sign. The modified action effects will again reflect in external reality, that will be changed (or unchanged) due to the modified actions. Thus, the reception effects will show, during sometime in the future, when my behavior will be changed due to the fact that I have received the sign. For example, after I hear: "the last tower has been taken", I might initially produce no covert effects. But later, when I will be questioned: "was the last tower taken?" I will answer: "Yes, it was" instead of "I don't know", which is what I might have answered if I never received the sign. Thus, the reception of the sign has modified by subsequent behavior. Of course, received signs only carry the potential for doing so; I might have never been asked in the future, and never used this information. Then, the actual reception effects differ from the potential reception effects.

Of course, from one viewpoint, we are not saying something extremely deep: of course the causes of a communicative action have to lie in the past, of course something in the local external reality through the agent's filter has caused the sign. And of course the consequences of receiving the sign have to lie in the future, and of course they will amount to something in the local external reality, and will depend on who the recipient is. But this externalized spatiotemporal locality as well as the producer/receiver dependence gives us a clear theoretical view, which can even be exploited statistically:

***Meaning*(given sign, at a given instant)**

=

Production Conditions, Reception Effects across all agents

***Production Conditions*(sign, agent)**

=

Which of all possible past experiences will make this agent produce the sign, which past experiences will always ultimately ground out to states of external reality

=

A classifier on all pasts, which dichotomizes them to those that will result to the production of the sign, and those that won't

***Reception Effects*(sign, agent)**

=

How will the receiving agent's behavior be modified for all possible futures, which will in turn further constrain all subsequent futures (states of external reality) after the agent's actions

=

A further constraint on all possible futures after the reception of the sign

Translation to the visuo-motor language

But how does this translate to a visuo-motor language that consists of physical actions too, like the one proposed in [Guerra-Filho et al.2005]? Let us see. The meaning of an action from the producer's end will be the production conditions of the action, i.e. the information that the action carries about the past as experienced by the actor. Thus, on the one hand, for an external observer, every observed action "summarizes" the past as experienced by the actor. The other half of the meaning, from the recipient's end, will be the constraint on all possible futures after the reception of the sign. Thus, on the other hand, again for an external observer, every action that is known to have been received (noticed) by a recipient helps predict the future.

And this is the total meaning of the observation from H1 of the performance of an action by H2 that was noticed by H3: If H1 is ideally positioned outside as an observer incapable of action, the observed action carries information about the past as experienced by H2 and the future as it shall be affected by H3. Of course, one might argue, when I see somebody performing an action, I might be able to tell what he's going to do next: but this is a secondary consequence, and it depends on the purposeful coherence of the actor's action language. Indeed, after the actor produced the action he might become irrational or amnesic and interrupt his meaningful action discourse. The primary information carried by the sign resolves to external reality and consists of the two parts given above:

information about the past as seen through H2, and information about the future as affected by H3.

And this seems to be a strongly anchored definition of "meaning", although at first sight impractical. However, given the usually small spatio-temporal windows, in many cases it can be quite applicable. Small-scale experiments using agents in simulated environments have been devised, and more are underway.

Recap

Thus, in the "production conditions" / "reception effects" framework for action semantics that I have introduced, every performed action says something about the past, and every action that has been noticed by somebody can act as a predictor of the future for somebody else that knows that the other party noticed. This is the primary meaning of an action, and of course there exist secondary extensions. And that being so, actions always carry information towards knowing the past and predicting the future - and this information is their natural semantics (primary meaning)[32].

Empirically sound semantics and GSMs

So, how does this all fit with the GSM proposal? Empirical semantics can definitely be used as a theoretical tool to talk about GSM-based systems, and also in order to do some human-mediated analysis. But, how about a robot trying to acquire empirical semantics by observing two humans interacting or a human and another robot? Within the above framework, this would be theoretically viable. However, they are some practical difficulties that make it difficult at this stage and for an adequate subset of language. The simple simulation-based examples of acquisition of empirical semantics assume very simple

[32]Through physical law, and through the physical / social chain reactions caused by an action, we can also have further effects: but this are modeled here as chain effects, i.e. secondary meaning.

environments (blockworlds), and very simple minds of agents: just stimulus-response (level-0 in the terms of chapter 2) organisms. When the environments becomes more complex, and the organisms starts to have extensive and structured internal state, such as is the case on the GSM-based Ripley the Robot, the toy techniques that are used in blockworlds are in desperate need of not-so-obviously-derived extensions. And the viability of such extensions has not been researched yet.

8.5.6 The intentional stance and GSMs

In GSMs, through the use of embedded GSMs, we are trying to estimate the contents of other agent's GSMs. The question thus follows: how do we know that other agent possess GSMs? Even more fundamentally, how can we know whether another agent posseses any sort of internal representation, or any sort of a belief?

The philosopher Daniel Dennett has offered the following answer to this question, which he has called the "intentional stance" toward an entity that is thought of as possessing a belief. According to the "intentional stance", ascribing a "mental" object such as a "belief" or a "desire" to an entity, is just a calculative device that enables prediction of the entity's behavior. Thus, "beliefs" and "desires" are just hidden variables that play a role in our models of the behavior of others - they might not correlate with any actual physically existing "carrier" of their contents - but still, might be very useful as components of our models of others.

In the GSM proposal, in the case of other agents, we postulate a "mental" part, which includes an embedded GSM, among possibly other things. In the simplest case - as a first approximation of the mental part of others (as we commented in the $AGENT$ MIND paragraphs of section 5.1.1), this embedded GSM has both the same structure, the same contents, and is fed through vision / language in the same way as the main GSM - i.e. we assume that other agents share the same belief contents with us, and they carve out the world in the same way. As a second approximation, one can still assume the same structure as well as feeding process, but now assume different contents due to the different position of the sensors etc., and at a higher level one can even start postulating different structures (ways to carve out the world), different feeding processes etc., and we saw some of the qualitative ways of difference in section 3.2.2. How does all this relate to the intentional stance? It is actually in obvious agreement with this stance. By postulating embedded GSMs containing estimates of the "mental" part of others, we are not making any claims about the actual physical existence of such a part; furthermore, by giving freedom of choice in the structure and feeding process of the embedded GSMs, we are clearly in accordance with the view that the mental part contents, including beliefs and desires, are merely a calculating device for predicting behaviors of others. Thus, at least in this respect, the GSM proposal is in agreement with Dennett's "intentional stance".

8.5.7 The homunculus attack and GSMs

The layman's description of the GSM as a "theatrical stage" in the robot's mind quite often attracts a classical philosophical attack against it - the so called "homunculus" attack. That attack goes as follows:

If you say that there is a "theatre" in the agent's mind where things happen, who is watching that theatre? Thus, your explanation depends on the existence of some kind of a magical miniature man inside the agent's mind - a homunculus (in Latin). Of course, this cannot be the case; thus, there's something absurd with your model. Or else, if you say that no, there might be another agent in there watching, and this other agent too has a "theatre" in his own mind, and so on - the a never-ending infinite regress follows - again, absurd. So, there cannot be a "theatre" in the agent's head - this whole GSM proposal is faulty.

Fortunately, the attack can be easily repulsed. Here, we are not supposing any unspecified "actors" or "viewers" in the "theatre of the mind"; instead, we have been very explicit in specifying exactly what processes will read and write the GSM contents, and how they will do so (section 5.2.2). The fact that we can externally visualize the contents of the GSM, peeking into the robot's head, is not essential and does not change anything. The play will go on without us, and without any other supposed internal viewer - except for the carefully detailed processes that do the reading and writing of the GSM. Thus, there is no need for any "homunculus" in the robot's head, even more so for one that contains yet another homunculus inside it and so on, and this superficial attack is quickly repulsed.

8.5.8 Perceptual Symbol Systems and GSMs

Lawrence Barsalou in [Barsalou1999] proposes an elaborate and interesting precomputational framework for a perceptual theory of knowledge, developed in the context of cognitive science and neuroscience. He sharply contrasts his proposal with amodal symbolic theories, which have become dominant in cognition during the last decades. The main elements of his theory are what he terms "simulators", which produce simulations of perceptual components that are extracted from experience and stored in memory. The conceptual system that is derived from these simulators is claimed to support type representation, categorization, production of categorical inferences, productivity, propositions, and abstract concepts - thus, a fully functional conceptual system is implemented. The question now arises: what is the relationship of Barsalou's proposal with GSMs?

First of all, notice that Barsalou is concerned with human cognition, while the GSM proposal is focused towards artificial SCAs. Second, there is a difference in the explicitness of the theory: Barsalou's proposal is a pre-computational framework, while GSMs are a fully specified computational theory. Third, let us ask: are GSMs an amodal symbolic theory or a perceptual theory of knowledge? The answer is that GSMs are indeed amodal, but they are not exclusively symbolic - they also contain the continuous and stochastic layer. Thus, in a sense they do belong to the "opposite team" of modal perceptual symbols, but are not the typical amodal propositional logic member of it. Also, they do overcome many of the problems of the typical amodal theories - let us for example visit the problems of amodal theories that Barsalou enlists in [Barsalou1999]:

Lack of empirical and neuroscientific evidence: As the GSM proposal is an AI proposal for building SCAs, and does not openly claim cognitive plausibility, the cited lack of such evidence from humans is not a problem - GSMs have clearly demonstrated plausibility on Ripley the Robot.

No satisfactory account of transduction process: That is, of the process that maps perceptual states into amodal symbols. But this process is actually completely specified within the GSM proposal - so this problem does not apply anymore.

No satisfactory account of the symbol grounding problem: That is, of the process that maps amodal symbols into perceptual stated. But this process is also actually completely specified within the GSM proposal - it consists of the process that creates sensory expectations, and the processes that matches and verifies sensory expectations, effectively giving a predictive component to the grounding cycle. So again, this problem does not apply anymore.

No satisfactory account of how an amodal system implements comprehension in the absence of physical referents: Well, this is exactly the case of receiving an "inform" statement within the GSM proposal, such as the statement: "Imagine a small blue object at the left". The physical referent may or might not exist and match, but the comprehension consists in producing relevant sensory expectations for it which are in principle testable and possibly verifiable in the future.

Mediating perceptual representation make amodal representations unnecessary: As long as Barsalou's theory has not been explicated computationally and implemented in a real-world system (while GSMs have), the argument that mediating perceptual representations make amodal representation redundant is not prohibitive - we have no hard proof yet that we can do it all just by using perceptual representations without any use of amodal representation.

Thus, it seems that although GSMs still arguably belong to the "amodal" side, the cited problems do not apply. Nevertheless, Barsalou's proposal is quite tempting and really wide in scope, and it would be interesting to see real computational implementations of it in the future, and then recompare with GSMs, and rejudge similarities / differences as well as advantages / disadvantages. It is also interesting to note that in Barsalou's answer to the responses of others to [Barsalou1999], he concludes with:

"The way I see it, three basic approaches to knowledge exist in modern cognitive science and neuroscience: (1) classic representational approaches based on amodal symbols, (2) statistical and dynamical approaches such as connectionism and neural nets, and (3) embodied approaches such as classical empiricism, cognitive linguistics, and situated action ... I believe that each of these approaches has discovered something fundamentally important about human knowledge, which the other two have not ... What I have tried to do in formulating perceptual symbol systems is to integrate the positive contributions of all three approaches. Regardless of whether my particular formulation succeeds, I predict that whatever approach ultimately does succeed will similarly attempt to integrate representation, statistical processing, and embodiment"

And the above seems to be a statement that is indeed pretty close to the spirit of the GSM approach. Of course, the GSM proposal puts heavier emphasis on the first two of the above components. Most importantly, the GSM proposal does not attempt re-enactment of modality-specific sensory experiences as part of its operation. In more detail: In the current implementation of the GSM proposal on Ripley, there exist fixed *"sensory bridges"* that update specific properties of specific objects through information coming in from one or more pre-specified modalities. For example, the human face position is updated through face detector information coming through the visual stream, the robot's pose is updated through proprioceptive information coming through the proprioceptive stream, the table object position is coming through another substream of the visual stream, or again there also exists a possibility for language-based updated of these property etc. However, some combinations are prohibited: for example, currently we cannot get robot body position info through vision, due to inadequacy of the vision system. If we wanted to get closest to Barsalou's proposal, we would have to move away from the idea of modality-independent (amodal) object "properties" that can be fed through multiple modalities as we do here, and push the results of imagination further towards the senses: when imagining a situation, instead of filling up the three amodal layers describing object properties, the system should fill up the modality-specific representations that exist at the sensory end of the "sensory bridges" (and not the modality-independent three property layers that exist at the property end of the "sensory bridges" - as said before).

Thus, we have discussed some claims, as well as some similarities and differences between the GSM proposal and Perceptual Symbol Systems. Although numerous differences exist, there are also many similarities in spirit and purpose; and in my opinion, there might be much to be mutually gained from further interaction of GSMs with Barsalou's ideas in the future.

8.5.9 Marvin Minsky and GSMs

Marvin Minsky has unarguably been one of the most prominent figures in the history of Artifical Intelligence, and his views have often helped shape the field throughout its history. Here, we will try to see in what ways some of his views relate with our proposal - some relevant to the issue of Internal Models and some to the Hierarchy of Multiple Representations:

In the celebrated "Society of Mind" [Minsky1988] provides some brief but very successfully directed points in his brief discussion of internal models.

First, he offers a double-sided definition of knowledge - one behavioral, and one representational :

D1: "When Jack says, "Mary knows geometry", this indicates to us that Jack would probably be satisfied by Mary's answers to the questions about geometry that he would be disposed to ask" (p.301 of [Minsky1988]) [a behavioral definition - internally held knowledge can only be verified when demonstrated through appropriate responses to the appropriate stimuli (not only the right answers, but the right demonstrations that might be comprised by motor actions too), and also knowledge is not absolute but observer dependent]

D2: "Jack knows about A" means that there is a "model" M of A inside Jack's head. (p.303 of [Minsky1988]) [a representational definition - knowing relies on the existence of some internal not directly observable structure]

Then, by combining D1 and D2, he offers a utilitarian measure for the goodness of a mental model:

Jack considers M to be a good model of A to the extent that he finds M useful for answering questions about A. (p.303 of [Minsky1988]) [again, this can easily be extended to: to the extent that he finds M useful for some purpose that serves his goals or the goals of life]

Second, having introduced "mental models" (roughly as "things in the head" M that help us answer questions about other things A, usually outside the head), he introduces "mental models of people" (which in the terminology of this thesis are agent models of humans), as well as "self-models". Then, he introduces "world models" (in his definition, "all the structures in Mary's head that Mary's agencies can use to answer questions about things in the world" p.304 of [Minsky1988]), and illustrates the nesting of models within models (Mary's model of herself within her model of the world, Mary's model of her model of the world within her model of the world etc.). He quickly hints at stopping at some depth as some measure against possible infinite regress in multiply embedded models (no explicit stopping criterion given though). And finally, he points out the rough bi-partite nature of "mental models of people", which consist of two parts: "model of the body" (physical realm) and "model of the mind" (mental realm) (in his words: "This means that Mary's model of her model of herself will have an overall dumbbell shape, one side of which represents the physical self, the other side of which represents the psychological self").

In Marvin's more recent "Emotion Machine" [Minsky2006], there is explicit discussion of the importance of the existence of a hierarchy of heterogeneous representations. In Marvin's words (chapter 8 of the Emotion Machine):

"The sections above have briefly described several kinds of structures that we could use to represent various types of knowledge. However, each of these representation types has its own virtues and deficiencies-so each of them may need some other connections through which they can exploit some of the other types of representations. This suggests that our brains need some larger-scale organization for interconnecting our multiple ways to represent knowledge."

Minsky's suggestion of a possible hierarchy includes many types of heterogeneous representations: some symbolic (narrative stories, trans-frames etc.), some numeric (neural networks etc.).

Now, having briefly considered some relevant views of Marvin, let us see how many aspects of the GSM proposal are related to his views:

Both mental and physical realm in agent models

First, the bi-partitiness of agent models (physical and mental realm) is found in our proposal too. In this document, it is actually derived from the first GSM desideratum, i.e. the "situation model structure that is implied by natural languages"[33]. Furthermore, in our proposal the two parts are augmented with a third part: the interface, connecting the two, as described in the paragraph $AGENT$ INTERFACE of section 5.1.1[34].

World models, models of others, models of self

Also, what we term Grounded Situation Model, is a part of Minsky's "world model", and also contains embedded GSM's, which are essentially what Minsky would call "mental realm of the mental models of other people that are contained within the world model".

When is internal representation good?

Furthermore, Minsky's utilitarian measure of goodness of GSM's as extended in the preceeding paragraphs is essentially a succinct expression of one of the main results of our discussion on "why internal representation arises?" in chapter 2[35]

SCA evaluation through tests

Finally, our proposal for evaluation of SCA's through psychological tests such as the token test, is reminiscent of his behavioral criterion of knowledge (D1).

Having seen all of the above points of agreement, below we discuss another point that Minsky makes (verbalization of gradations of uncertainty) that has also been considered in our GSM proposal (description of the "inquirer" module in section 6.4), but where our current practical implementations choices differ.

Gradations of uncertainty

Regarding verbalization of gradations of uncertainty, Minsky proposes that they are actually verbalized as (p.302 of [Minsky1988]):

"The red object *is* on the table", "*I think* the red object is on the table", "*I believe* that the red block is on the table"

This is quite different than the verbalization approach taken here ("most probably, probably, etc." in figure 6-11). Our approach is indeed more unnatural in terms of actually spoken language, and the semantics attached to "most probably" are more scientific and normative instead of being empirically derived. However, it should not be difficult

[33] In essence, the bi-partiteness arises from the "folk psychology" implied by natural languages - which might have slight differences across natural languages and cultures, however in most of them this bipartiteness is clearly manifested.

[34] Which is essential from a practical viewpoint in using the agent models for mind reading, as commented upon.

[35] Not our extension, but the original Minsky measure, one can derive by replacing the "goal hierarchy of life" with the more specialized goal of the "perfect answerer".

to change and retune the decision tree used here in order for it to accord with Minsky's proposal, if desired.

Aren't GSMs a single-agent proposal?

Furthermore, one might ask: But isn't the GSM proposal, which seems like a monolithic single-agent AI proposal, totally contrary to Minsky's "pandaemonium", which is a multi-agent model of intelligence? At first sight it seems so, however this is not completely the case. First, notice that although a GSM-based agent externally appears as a single agent, in a real-world implementation multiple agent-like entities might reside within the modules proposed. Also, notice that one can easily relax the strictly modular view taken here, and enable some limited "illegal" connectivity among modules, while keeping pretty much the same overall clusters of concentration. And thus, we are not that far away from Minsky's "pandaemonium"; it is just a more ordered society of daemons in our case, with prescribed territories, and no direct access to actions and effectors for all of them.

Multiple levels of representations

Finally, but most importantly, the GSM proposal is very much in agreement with the need of multiple levels of representations that Minsky insists so much upon. The three-layered design which is at the heart of the GSM addresses exactly this need: to bridge the stochastic, with the continuous, and the symbolic. When one views the different representations that exist within the GSM, he finds many analogies to Minsky's hierarchy: our sequences of events when translated to FOPC are similar to his "narrative stories", our property lists are similar to his "frames", the predictive part of our proposal could contain "trans-frames", and our stochastic-to-categorical classifiers could easily be "neural networks". Indeed, there is a deeply shared belief with Marvin that a huge reason behind the shortcomings of current systems was their insistence on using one type of representation universally - usually either connectionist or symbolic.

8.6 Recap

In this chapter, we considered a variety of subjects related to semantics, grounding, AI and philosophy of mind, in light of the GSM proposal. Many tangents towards a wider view were given, as well as directions for extensions (some of which to be explicitly considered in the next chapter on future extensions), arguments against possible classical attacks against the proposal, and discussions of the relevance of ideas found elsewhere with our proposal.

The overall organization of this chapter consisted of five large blocks (each assigned to a section):

Block 1: Language grounding, Situated language (8.1)

We discussed meaning spaces, speech as sensing / acting, grounding the whole of language, and levels of detachment from the "here-and-now".

Block 2: Abstract entities, Imaginary situations (8.2)

We discussed the primary existence of abstract entities, implications by the GSM proposal, the representation of abstract entities within GSMs, inform speech acts vs. imagine statements, ab-initio creation, and the possibility of "pure imagination".

Block 3: Semantics, Situations (8.3)

We discussed semantics of adjectives and verbs and extensions through successive approximations, comparative adjectives and spatial semantics, and the similarity and enumerability of situations.

Block 4: Logic, Ontologies (8.4)

We discussed the bridging of GSMs with logical representations, and the grounding of ontologies.

Block 5: Stance X and its relation to GSMs (8.5)

We discussed embodiment, SHRDLU, observer-only systems, and procedural semantics in light of GSMs. Then, we gave a proposal for "empirically sound semantics", and discussed the intentional stance, the homonculus attack, as well as Lawrence Barsalou's and Marvin Minsky's views in light of the GSM proposal.

Now, having reflected on all of the above important topics on the basis of the GSM proposal and the implemented system, we will try to look towards the future and ask: "what lies ahead?" by discussing possible shorter and longer-term future extensions. We will also try to show how much leverage the existing proposal provides towards these extensions, making them much easier to achieve as compared to starting from scratch, and providing ease of integration in a well-organized system.

Chapter 9

Future Extensions, Contributions and Significance

Having discussed theoretical implications of the GSM proposal in the previous chapter, we will now move on to the final chapter of this thesis, and discuss future extensions in section 9.1, specifically consider the conjectured SCA prerequisites of intention recognition and activity coordination in 9.2, and close with a conclusion, restating the contributions of this thesis, discussing their significance, and ending with an epimetron in 9.3.

9.1 Longer-term extensions

First of all, let me start with a generic positive comment that I hope will be supported through this section: The basic GSM architecture is not only customizable, but also have proved to be easily expandable in various directions. As we shall see, the current proposal provides significant leverage towards all the directions of extension that we will consider; if one tried to pursue these directions without the existing GSM proposal as a basis, he would have not only to rebuild a significant amount, but also, he would end up with a one-sided system that would not afford easy integration of different extensions. But together with this comment, an open question remains: do we expect this architecture to scale to more behavioral capabilities easily, or can one easily foresee obstacles or exponentially-increasing barriers for future growth? In the following sections, I will try to provide evidence for the generic comment, and also a discussion of the crucial open question, which we will both revisit in soon section 9.1.11. But before that, let us first start by briefly listing some directions for immediately applicable short- to mid-term extensions:

X1) Better Visual Abilities:
X1.1 Richer shapes & shape recognition
X1.2 Active sensing
X1.3 Richer spatial relations (containment, support, etc.)

X2) Better Tactile Abilities:

X2.1 Soft/Hard, texture, and temperature sensing
X2.2 Shape by touch

X3) Better support for Mind Reading / Embedded situation models:

X3.1 Eye gaze direction
X3.2 Absence/presence of humans, face recognition, multiple speakers
X3.3 Extend recognition of human-related events
X3.4 Keep history of interactions w each human

X4) Better Language subsystem:

X4.1 Online naming of objects and agents
X4.2 Conditional execution of actions
X4.3 More systematic treatment of vocabulary, speech acts, descriptions etc.
X4.4 Extend inform speech act handling for story understanding / future anticipation

X5) Mixed motor/speech planning, explicit goals, inference, affective state

X5.1 Translate categorical layer to FOPC
X5.2 Perform simple inference on the basis of translated categorical layer
X5.3 Perform situational commonsense knowledge acquisition
X5.4 Devise goal representations and perform simple planning
X5.5 Perform online collaborative planning / action coordination
X5.6 Introduce affective state for the robot fed through goal satisfaction, novelty etc.
X5.7 Introduce affective state for the human fed through expression recognition etc.
X5.8 Affective learning for the robot by observation of human affective responses

X6) Core GSM enhancements

X6.1 Extend basic GSM operations: attentional focusing etc.
X6.2 Better memory subsystem: Consolidation, long-term experience acquisition
X6.3 Incorporate more results from cognitive science research

X7) Integrated Online learning

X7.1 Integrate online learning of adjectives, verbs & params with normal operation
X7.2 Online learning of tasks and situationally-appropriate actions

X8) Passing Landmark tests / Evaluations

X8.1 Pass all of the "Token test" with professional testgiver
X8.2 Perform simple story understanding tests
X8.3 Perform cooperative assembly task tests
X8.4 Perform extensive tests through the levels described in section 7.2

X9) Standardisation / Open Sourcing / Design tools

X9.1 Open Source a generic library of basic GSM representations and processes
X9.2 Build design tools for semi-automated CAD of GSM-based SCAs

X9.3 Build an affordable educational GSM-based robot (One Robot Per Child)

X10) Theoretical
X10.1 Alignment of Situation Models
X10.2 Empirically Sound Semantics

One can observe that the ten extension directions listed above roughly proceed from the *sensory* domains (X1-3: object vision, tactile, human activity) to the *cognitive-affective* (X3-6: embedded situation models, language, plans / goals / affect, memory), then move to the *real-world operation and design* (X7-9: online learning, evaluation, design), and finally close with *theoretical reflections* (X10). One should also observe that although the directions are generally independent, some prerequisite relations among them exist (for example: X5.5 depends on X5.4, X3.3 and so on). Also, one should notice that some of these directions have already been partially explored in our implemented system, and also other systems.

Now let us try to consider these directions in more detail, granting a subsection to each major direction:

9.1.1 Better Visual Abilities (X1)

One can aim towards enabling the system to handle richer representations of object shapes, acquired through multiple views, integrated in an active vision framework. Richer shape capabilities will enable the incorporation of richer spatial relations, such as containment and support, which in turn figure prominently in natural language semantics.

X1.1) Richer shapes & shape recognition

As mentioned in the "visor" paragraph of section 6.4, a simple voxelized 3D shape model acquisition system that generates the models on the basis of multiple views of an object is already operational on Ripley, yet has not been tied to language yet. However, many possibilities for extension of the current system exist: The current method is somewhat crude, and suffers from lots of redundant computations, errors, and non-consideration of helpful constraints arising from the nature of multiple-view geometry. Thus, more advanced methods can be tried out in the future - any standard textbook such as [Hartley and Zisserman2000] can provide useful insights. A better segmenter would also be required for handling more complicated cases. Regarding shape recognition, some crude methods have been tested out in matlab for the output derived from the voxelized method described above. These methods involve canonicalisation transformations to provide translation and scale invariance, and then comparison to some "archetypical" voxelized prototypes for spheres, cylinders and rectangles. More advanced methods could easily be employed. Furthermore, the current method suffers from the poor position / pose reachability of Ripley due to his mechanical design[1]; thus, an adequate set of views can only be taken for objects positioned near the center of the table[2]. As a general

[1] See comment in section 6.1 and document **??**.
[2] See also **??**.

comment, the voxel-based shape representation has on the one hand been found to be simple to handle, but on the other hand is not necessarily optimal - many other shape representations might prove more useful in the future[3].

X1.2) Active sensing

As can be seen in figure 6-8 (chapter 6), our current vision system is basically a long pipeline, continuously feeding through anything that reaches the camera to the visor for comparison with the current state of the situation model. However, this "continuous feedthrough" approach has two inherent disadvantages: first, a lot of computational power is wasted, and second, it is cognitively unplausible. Humans generally don't perceive everything that comes to their eyes, and two types of motivations determine what they will notice: bottom-up (percept-driven) as well as top-down (goal-driven). For an example of bottom-up motivations, one can have a look at models of visual salience (that can also sometimes predict human saccades), such as [Itti et al.1998][4]. For an example of top-down motivations, one can have a look at classic examples of goal-influenced change-blindness in the cognitive science literature [Simons and Rensink2005].

In order to change the current sensory system from "continuous feedthrough" to "active on-demand sensing", one should first decide upon a set of basic sensory processing primitives, that will become available as actions to be chosen by a planner or some other action-selection subsystem (see extension X5.4 later in this chapter). Regarding the choice of suitable primitives, one can have a look at the visual primitives chosen at [Ullman1996], [Rao2002], and one could also look at [Breazeal and Scassellati1999].

X1.3) Richer spatial relations (containment, support etc.)

Given a better vision system and 3D shape representation capabilities (or otherwise), one could attempt to recognize and represent the notion of containment ("the ball is in the cup"), support ("the cup is on the square") etc. in some form. The notions of containment and support are not only particularly important from a practical viewpoint, but also from a theoretical: through a spatial-to-temporal analogy, one could also cover temporal containment, and connect the semantics of "in" with those of "while", for example: "The red one appeared while my head was moving" will then be able to be metaphorically visualized in a single moment, through the objectification of the events (two objects, one for the event "the red one appeared", and the other for the event "my head moving"), and the spatial containment of one object within the other.

[3]See for example **??**.

[4]An implementation of which has been already created for Ripley, albeit is not fully functionally connected yet.

9.1.2 Better Tactile Abilities (X2)

X2.1) Soft/Hard, texture, and temperature sensing

The current touch sensors do not allow anything better than a discrimination between "something gripped" and "nothing in gripper". However, one could equip the robotic gripper with better sensors, that would enable hardness, texture, as well as temperature measurements, effectively grounding out the meanings of "soft", "rough", and "hot" (for example) to a sensory space.

X2.2) Shape by touch

Given a different gripper[5], one could try to measure normal vectors at different points of an object, and thus effectively acquire shape-by-touch. This capability would in turn many enable interesting experiments; for example, in cross-sensory alignment (aligning vision with touch), cross-sensory informational conflicts (see 'imaginer' section in 6.4) etc.

9.1.3 Better support for Mind Reading / Embedded situation models (X3)

X3.1) Eye gaze direction

Eye gaze direction recognition would enable one to play with interesting misalignments between the contents of the robot's situation model, and the estimated situation model of the human. For example, the human might look away for a while, or have access to a part of the table that the robot does not, and so on. As commented upon previously in ($AGENT$ MIND paragraph in section 5.1.1), such misalignments could motivate hearer-specific description generation, or the purposeful issuing of inform speech acts in a planning framework.

X3.2) Absence/presence of humans, face recognition, multiple speakers

Currently, there is an inherent assumption that a human is always present somewhere around the table of Ripley, and the human head is just moved around whenever it is seen. One could extend this in many ways: by allowing the absence / moving away of the human, by enabling multiple humans whose identity is recognized through face recognition to interact with Ripley either one at a time or being all together around the table and so on. Many interesting behaviors can then be experimented upon.

X3.3) Extend recognition of human-related events

Currently, Ripley can parse events corresponding to movements of the human head and hand, and touching events where the human touches an object. These can be extended to a much wider set of human-related events and human-object interaction events.

[5] such as three-fingered gripper with precise force sensors.

As commented upon before in 6.2.2, recognition of human-object interactions is immensely important for coordinated execution of plans (also see X5.5 later) etc. Also, one could try using a whole-body tracker of humans for future humanoid robots[6].

X3.4) Keep history of interactions w each human

It has been found that in general, humans get bored when interacting with robots in the long term - see for example [Ishiguro et al.2006]. But why is human-human interaction not boring in the long term while human-robot interaction currently is? One can imagine a number of reasons: first, the robot's behavioral/conversational repertoire is usually very limited. Second, and quite importantly, robots currently do not keep histories of their interactions with particular humans - and does do not have shared memories and experiences with them, which could form an interesting topic for discussion and consideration, and which could possibly increase the bonding between the human and the robot tremendously. Given extensions X3.2, X3.3, and X6.1, one could experiment in this direction and empirically try to justify the above claim - that when robot and human share common experiences and memories their bonding will increase[7].

9.1.4 Better Language subsystem (X4)

X4.1) Online naming of objects and agents

Currently, body parts have fixed names, and apart from the personal pronouns, the human and the robot have no given names. It should be quite straightforward to extend the current system to support "declaration" speech acts (name-giving actions) such as: "Let's call the blue one Mickey, and the other Mini. Where was Mini when Mickey left?" Furthermore, having X3.2, one could also experiment with greetings and introductions: "Hi! My name is Nicholas" or "Hello Nicholas!"

X4.2) Conditional execution of actions

Currently, Ripley can only handle requests for immediate actions. One could foresee the quite useful possibility of extending temporal referents to the future, in order to support the execution of conditional actions, such as: "When the red one appears, pick it up!", "When I move my head, touch the green one!" As such commands are part of the fifth part of the "Token Test", one method to achieve support of these statements is described in the quasi-minimal design that passes the Token Test which is derived in detail in [Mavridis2005c].

[6]Extending to human-like bodies is a very interesting endeavour on its own, and although many complications arise, for example because of the complexity of motor control, it is amazing to note how much more effective the human-machine coupling can be become, because humans automatically empathize and read human traits in the machine, just because it looks much more human-like. See for example some of the amazing research of Hiroshi Ishiguro's on androids / geminoids [Ishiguro and Minato2005], or David Hanson's robots [Hanson et al.2005] and imagine a day when such natural-looking artificial bodies will be habitated by more advanced minds.

[7]See also [Bickmore2003].

X4.3) More systematic treatment of vocabulary, speech acts, descriptions etc.

Glancing through the current set of behavioral capabilities of Ripley (appendix A), instantly reveals that the current supported vocabulary as well as the supported verbalization methods are not extensive. For example, although there is support for several types of wh-questions (such as "where", "how big", "what color"), there is no support for "which" questions, such as: "which object is at the left?". The capacity for answering "which" questions might require the generation of (possibly hearer-dependent - utilizing the embedded GSM) uniquely-specifying object descriptions, such as: "the small red one", or similarly temporal descriptions for "when" questions. Many such directions for extension exist. Furthermore, the systematicity, parametrisation (and consequently learnability) of both the present language comprehension as well as present language generation subsystems can be very much improved. This would also enable an easier migration towards extension areas X5 and X7.

X4.4) Extend inform speech act handling for story understanding / future anticipation

Currently, the only type of inform speech acts can be serviced contain information about the present situation: "there is a blue one at the left", and cannot include temporal sequencing information, which would enable them to expand to the past / future or to imagined "stories" consisting of sequences of events, such as:

"One day, I saw John. Then a blue ball appeared. Then John asked me to pick up the blue ball. I picked up the ball. After a while, a red ball appeared. It started moving ... "

These past / future / imaginary sequences of events could be later queried through questions, integrated with current experiences, or be anticipated.

X4.5) Richer semantic models

As discussed in relevant sections of chapter 8, the semantic models currently employed in the Ripley implementation are just a first approximation to potentially much richer models. Such extension proposals were discussed specifically at the sections: "Successive approximations of meaning models" (section 8.3.2), "Comparative adjectives and spatial semantics" (section 8.3.3), and more generally at the "Grounding the whole of language" section 8.1.3. As a suggested first step, one could extend/integrate the comparative adjectives proposal.

9.1.5 Mixed motor/speech planning, explicit goals, inference, affective state (X5)

Currently, as commented upon before, Ripley's action selection system is effectively equivalent to a (stimulus, GSM context)-to-response table. An obvious next step would be to experiment with multi-step ahead action planners, as well as partial planning (in the sense of [Bratman1987]), and move all the way to real-time cooperative planning / action execution together with the human user. Simple affective states for the robot, resulting from the outcomes of previous actions, and/or modulating future action selection might also be introduced. Furthermore, sensing of the human's affective[8] state would enable interesting interactions. As a simple example: if the human is displeased from an action of the robot, the robot could acknowledge that and offer to undo the previous action.

X5.1) Translate categorical layer to FOPC

The first step towards planning and inference, would be the translation of the GSM categorical layer contents as well as events to some form of FOPC (First Order Predicate Calculus) sentences, as commented upon is section 8.4.1.

X5.2) Perform simple inference on the basis of translated categorical layer

Having implemented some form of X5.1, one can then try to implement some inference engine operating on the derived sentences. This would also enable the answering of a wider set of questions, whose answer would depend on the inferential results.

X5.3) Perform situational commonsense knowledge acquisition

By noticing empirical regularities arising in the experienced situations (for example, that banana-shaped objects are usually yellow, or that before an object stops moving, it has to be moving for a while), one can try to conjecture the existence of such empirical laws, and entertain them as pieces of commonsense knowledge if they are found to generally hold. It is worth stressing here that these laws do not arise out of any form of logical necessity - there could exist environments where they do not hold. On the contrary, they arise out of empirical / statistical evidence[9].

X5.4) Devise goal representations and perform simple planning

Once we can describe situations and event sequences in FOPC, we can also try to describe goal-states using first order logic sentences. Then, we can try to move from the current situation to a situation that satisfies the goal description by selecting a set of actions - through a multitude of planning approaches, some of which were discussed in chapter 2. Furthermore, in the case of SCAs, there are two species of actions available to the assistant: motor actions as well as speech actions. Then, some form of mixed motor/speech act planning can take place, in the vein of approaches such as [Appelt1985]. For a look at some of the underlying principles of mixed speech/motor act planning, and the use of embedded GSM's, the reader can look back at the section entitled "Unifying Speech with Sensing and Acting" (section: 8.1.2).

For a quick illustration of some basic ideas involved in this endeavour, consider the following: If the robot can estimate the state of the human user's knowledge, and can

[8]For an extensive discussion of the importance and possibilities of affective computing, the interested reader is referred to [Picard1997].

[9]Which in turn, arises also because of what Whitman Richards calls the existence of "natural modes" in the world: tight clusters of highly correlated properties (see for example [Bobick and Richards1986]).

also predict the effect his utterances will have on the human's knowledge, then he can select among possible utterances through a homeostatic stance. So: On the one hand: a physical action affects the properties of an object in the robot's mental model (reflecting the outside world). On the other: for example, an "informative" speech act affects the properties of an object in the human's mental model, an estimate of which is included in the robot's representation of the human agent, i.e. in the robot's mental model. Thus: in the same way that planning can be used to stack three blocks on a table through the robot's physical actions, it can now be used to get the robot to make the human know that there are three objects on the table even if the human cannot see them. Furthermore, for example: Planning can be used to issue commands to the human in order get him to assist the robot in the object stacking task and vice versa.

X5.5) Perform online collaborative planning / action coordination

Collaborating humans plan together, and often re-plan and negotiate during the execution of their plans. Achieving collaborative planning between humans and SCA's, requires step X5.4, plus a number of further capabilities.

For a quick illustration of some basic ideas involved in this endeavour, consider the following: Through a common knowledge of goals, mutual mindreading, and the knowledge of the effects of both physical and speech actions (assuming a collaborative partner), we can get the robot to collaborate with the human in a task in a natural manner.

X5.6) Introduce affective state for the robot, fed through goal satisfaction, novelty etc.

One could envision the implementation of simple frustration, satisfaction as well as amazement variables, which are fed through events such as: unsuccessful completion of current action or goal, non-fulfillment of sensory expectations or detection of novel objects etc. The current values of these variables could in turn promote/demote behaviors, by altering utilities in a utility-based planner or otherwise.

X5.7) Introduce affective state for the human, fed through expression recognition etc.

By enhancing Ripley's vision system through augmentation with a subsystem such as ESP[10], one could try to read human affective state through facial expressions. This could in turn drive behavioral alterations for the robot: for example, actions that keep the human happy or interested should be promoted etc., or during learning sessions the robot might use apparent user satisfaction or confusion for feedback and so on. Some key affects whose recognition would be very useful are agreement, disagreement, and interest.

X5.8) Affective learning for the robot by observation of human affective responses

Also, the robot could try to "tune" its own affective subsystem with the human's: by observing the effect of the outcome of human actions or of novel experiences to the

[10]For more information, see the relevant page of the affective computing group of the MIT Media Lab at http://affect.media.mit.edu/projectpages/esp/.

human's affective state, it could try to mimic these event-to-affect responses for its own affective model. For example, if unsuccessful handling of an object leads to frustration to the human user, then the robot could learn to be frustrated too in similar situations. Furthermore, apart from learning the (situation, goal)-to-affect map (for example: learn when to be happy by observing when humans are happy), the robot can also try to learn the second part of the process, i.e. the (affect, situation)-to-action map as observed by humans (i.e. learn how to behave when it is happy by observing the behavior of humans while they are happy)[11].

9.1.6 GSM enhancements (X6)

X6.1) Better memory subsystem: Consolidation, long-term experience acquisition

As commented upon in section 5.2.3, the robot's episodic memory is currently just a flat, ever-increasing storage space of moments and events. This creates overflow after several hours of operation, and is of course a cognitively implausible simplification. One could start structuring the episodic memory into compartments, by separating recent from long-term storage, and one could even have consolidation processes operating during idle periods of the robot ("sleeping"), that try to parse through old memories, throw out / compress unnecessary moments, find regularities that had existed etc. Also, one could impose some "associative" structures in the long term store, enabling quicker access of components of moments and events through a mesh of relations, leading to associative chains among the components.

X6.2) Extend basic GSM operations with attentional focus etc.

Apart from the standard update / matching process devised and implemented already (see sections 5.2 and 6.4), several other elementary processes can be proposed for the GSM. Consider the operation of "focusing", which in different variants has been proposed in the cognitive science literature [Grosz and Sidner.1986]. For example, when the robot is helping out in an assembly task, the GSM might be filled with data describing the current situation: the bricks, the human, the human's beliefs etc. Given a command to "move the red brick", the GSM's focus might change, bringing that particular brick under the spotlight. Subsequent cognitive operations will give higher preference to the recently focused areas. Also, given a perceived event, for example the human moving the green brick, the focus might follow. The role of such attentional foci has been major in experiments dealing with story understanding etc.

X6.3) Incorporate more results from cognitive science research

Many more results from the cognitive science regarding human situation models could be incorporated within the proposed model - for example, ideas about the five proto-

[11] In analogy to the move from situationally appropriate speech to purposeful speech, the robot could, after having learnt when to display situationally appropriate affect, to also learn how to use the display of its affective state in a purposeful manner, as operators in plans.

typical dimensions etc[12].

9.1.7 Integrated Online learning (X7)

X7.1) Integrate online learning of adjectives, verbs & other params with normal operation

Currently, in order to change the meanings of adjectives and verbs for Ripley, the human user should stop its normal operational mode, and hand-edit (through training data or otherwise) the categorical classifiers and event classifiers (see section 8.3.1). One could envision the integration of learning with the normal operation of the robot - in such a case, a human "teacher" would be allowed to manipulate the external world in order to provide "training samples" in order to tune the robot's categorical classifiers, event classifiers, as well as other parameters. For example, the human might point towards an object, and say: "this is red" and so on[13]. This could even allow the robot to have user-specific preferences for the meanings of adjectives (what Nick calls red might be different than what John calls red). Furthermore, one could envision devising a learning procedure, through which the robot incrementally learns a wider set of parameters, effectively "aligning" its situation model with the human's (as discussed in section 3.2). The notion of active learning should also prove important in such attempts, i.e. moving the robot from the role of the pure observer of a presentation to an active learner that selects examples and asks questions.

X7.2) Online learning of tasks and situationally-appropriate actions

Given a reasonable achievement of X3.3, the robot can try to learn tasks by observing the human execution of the tasks. Given a suitable implementation of a situational similarity metric, one could also try to generalize and produce situationally-appropriate actions for a wider variety of conditions, as described roughly in the "similarity of situations" section (section 8.3.4).

9.1.8 Passing Landmark tests / Evaluations (X8)

X8.1) Pass all of the "Token test" with professional test-giver

A mid-term goal would be to develop the GSM to a stage that enables the robot to pass all five sections of the standard Token Test for Children (see section 7.4. A theoretical design for doing so has been derived by the author in detail, and is described in [Mavridis2005c]. A full real-world implementation as well as an actual instance of administration of the test to the robot by a qualified test-giver under normal conditions remains to be done.

[12] For example, see: [Zwaan and Randvansky1998].

[13] For an example of limited online learning through examples, look at (ref:RoyToco).

X8.2) Perform simple story understanding tests

Given X4.4 as well as other enhancements, one could envision a human-replacement story understanding test given to the robot, in which a story is narrated to the robot, and then questions are asked in order for it to demonstrate its understanding. The demonstration might also include illustration through manipulation of objects, effectively "acting out" episodes from the narrated story as children often do.

X8.3) Perform cooperative assembly task tests

Given X3.3 and a reasonable degree of X5.5 or other X5 stages, one could demonstrate human replacement in a cooperative assembly / object re-arrangement task with human and robot sharing the table.

X8.4) Perform extensive tests through the levels described in 7.2

One could quantify performance of vision, speech recognition, motor skills as isolated subsystems, or one could measure task completion rate, human satisfaction etc. One could also try to characterize characteristics of the "transfer function" between external reality and internal situation model: position errors, time delays etc.

9.1.9 Standardization / Open Sourcing / Design tools (X9)

This is a very important area for extensions. The real value of the GSM proposal will be exposed only if many different designers of SCAs use it and extend it, and the best way to get people to start using these ideas is by providing an open-source implementation of some core components that will save them time and effort when building SCAs.

X9.1) Open Source a generic library of basic GSM representations and processes

This is a mid-term goal, that could easily be accomplished given sufficient manpower within less than a year. The C++ classes for three-layered GSMs that already exist for the case of Ripley, should be cleaned-up, generalized, and packaged in a distributable form.

X9.2) Build design tools for semi-automated computer-aided design of GSM-based SCAs

Apart from the manual method for designing SCAs described in section D.4, one could envision a graphical SCA-development environment, which pretty much in the same way a CAD program enables easy re-use of components from a library, would speed up the development of novel SCAs tremendously. In a sense, this could be some form of a GSM-based extension of products under development, such as Microsoft's Robotics Studio. Of course, the adoption of X9.1 from a community of users seems like a natural prerequisite for such an extension.

X9.3) Build an affordable educational robot on the basis of GSMs (One Robot Per Child)

This would use easy-to-use and learn versions of both X9.1 and X9.2, and utilize cheap hardware - such as a $100 laptop[14] and a $100 robotic arm with vision.

9.1.10 Theoretical (X10)

X10.1) Alignment of Situation Models

The general discussion of section 3.2, can be further extended, refined, and reach a level of maturity that will allow implementions.

X10.2) Empirically Sound Semantics

The definitions and approach described in section 8.5.5, can be applied to toy problems in simulation, as well as more realistic cases, and also current semantic theories can be explicitly positioned in the framework of this section as well as the general framework of chapter 2.

9.1.11 Extensibility and Scalability

Now, let us return to the generic comment made in the opening paragraph of this section, and to the crucial open question. First, the generic comment: is the GSM architecture indeed easily expandable? Did it provide significant leverage for the extensions? Most of the extensions described above, can be easily accommodated within the proposed framework, and furthermore, their implementation can be foreseen to be quite straightforward, through incremental augmentations of the existing system. In that sense we believe that the generic positive comment indeed holds: the GSM architecture is expandable and provides significant leverage for all these directions of extension.

But now, the difficult open question remains: *How much of this will indeed scale, and up to where?* This has to be seen in practice. From where we stand now, it seems that it will - one can see that most of the extensions for directions can be foreseen to have non-trivial end results. If there are some general guidelines to be given, where hopes for scaling can be anchored, they are the following: hierarchical learnability (relevant to X7), and open sourcing (relevant to X9).

First, let us consider *hierarchical learnability:* the only way to really scale is to be able to move from hard coding aspects of the system to empirically learning them through teaching sessions or social interactions with normal human users (and not expert coders). Going back and redesigning an existing system so that several aspects of it can be trainable through teaching by normal humans might not be the ideal PhD project or demo show-piece: in the end, the trained system might be able to exhibit only the same behavioral repertoire as the original hand-coded system was able to exhibit. However, as the desired capabilities increase, only trainable systems and systems that capitalize on their long-term experiences with users will be able to scale.

[14]For more information, see http://www.laptop.org/.

Then, let us consider *open sourcing:* the man-power demands for the development of the system, will also increase with time, and also an adequate user base should exist in order to teach it and in order to gather long-term experiences. Thus, the steps described in X9, are crucial not only due to ethical and philosophical considerations (open and free dissimination of knowledge etc.), but also due to very practical ones: no other to achieve a huge enough developer as well as user base can be easily foreseen.

Thus, this is I believe the most appropriate answer to the scaling question, given where we stand: scalability will only be achieved through learnability / trainability, and through the establishment of a big enough developer as well as user base. No hard barriers can be easily foreseen from where we stand - although some classic problems at the sensory / motor control end of things remain if we want full blown humanoid embodiments in real unrestricted environments: simply - vision, speech recognition, and walking algorithms still need to mature more, and take context and affect into account much more than they currently do. But if we restrict ourselves to a Ripley-type embodiment with pretty much controlled objects, we can still go a long way towards more complicated language and cooperativity (X5.5) - and thus get significantly nearer to the ideal cooperative Situated Conversational Agent.

9.2 Achieving Intention Recognition and Activity Coordination

Having come full circle and having returned to the ultimate vision of this work, namely truly cooperative Situated Conversational Agents, let us now look back at chapter 1.2.5, and revisit our three conjectured prerequisites for the ideal SCA, namely support for *situated natural language interactions*, as well as a capacity for *context-dependent human intention recognition* and *real-time activity coordination.* We seem to be on a good path towards the first prerequisite - a brief glance through the current behavioral repertoire as well as the list of extensions will provide justification; but is this the case for the second and the third too? Does the GSM proposal and the implemented system indeed provide a solid basis for extensions towards human intention recognition and real-time activity coordination? We believe that one feature of the GSM proposal, among many others, acts catalytically towards the fulfillment of both of these prerequisites: the existence of embedded situation models. Let us consider each of the prerequisites in turn.

First, consider *activity coordination* - this corresponds very closely to extension X5.5, which in turn depends to a number of other extensions, such as X5.1, X5.4, X3.3 etc. In more detail: the currently implemented system provides the way to bring the situation context "inside" to the GSM. The current state of the objects relevant to the activity under execution is being tracked; as well as the current state of the human partner, and also his current actions and interactions with the objects (through the recognition of human-related events, such as "human picked up x" etc.). If we translate the categorical layer to FOPC (X5.1), which seems to be quite straightforward, then we can represent goals and use standard planning algorithms (X5.4). If the robot can then dynamically replan motor actions on the basis of the current state of the objects relevant to the activity, then we can have uncoordinated synergistic activity. If furthermore we extend planning from motor actions to mixed speech / motor actions, and either the human or the robot take the responsibility assigment / activity leader role, than we can start the activity by verbally assigning responsibilities, and dynamically replan / inform the other / reassign responsibilities throughout the execution of the activity. Of course, we have not reached that stage yet, and many details need to be worked out, but our current system provided significant leverage, the overall steps to be taken seem clear, and all of the above described steps require extensions that can integrate well within the proposed framework.

Now, consider *intention recognition*[15]. As mentioned before, apart from explicit statements of declaration of intention ("Now I want the dishes to be cleaned"), intention is often silently inferred by observing actions that naturally belong to specific plans given the situational context. For example, in the kitchen assistant task, when I see my assistant looking at the tomatoes that are within my reach, I might easily have guessed that he needs them for the salad he will prepare. What steps do we need in order to get there? First, the observation of the actions taken by the human - extension X3.3 becomes relevant. Then, the situational context, up to enough detail - covered by what we have plus a task-specific version of X1, X2 etc. And then, a way to know how actions compose to relevant plans and/or what are some relevant or usual plans for relevant tasks - covered by some version of X5.4 or X5.5. Thus, once again - although we do not have perfect visibility of the path that will lead us there, we have provided significant leverage, and we can foresee the relevant easily integrable steps towards achieving intention recognition, i.e. the remaining prerequisite for truly cooperative Situated Conversational Assistants.

Thus, the vision of the ideal SCA, discussed in section 1.2.4, might not be as far away as it first seemed. Now, having discussed how we could fulfill the prerequisites for truly cooperative SCAs, we can at least see some possible pathways towards our vision, which will become clearer the closer we approach them on the way.

9.2.1 Contributions of this Thesis

This thesis consists of a multi-part proposal for building Situation Conversational Assistants (SCAs) through Grounded Situation Models (GSMs). The GSM proposal is not just a theoretical proposal for future system - it contains explicit computational models, and has also been implemented and evaluated on a robotic SCA. It covers motivation, definitions, explicit computational models, an implemented system, evaluations, design methods, future extensions, and relations with other theories.

In more detail, it consists of several contributions, including: a *theoretical motivation* of GSMs, a listing of set of *defining characteristics* that any GSM should possess, a *computational model* of the basic representations and processes comprising a GSM specified to the level of pseudocode, a modular *implementation architecture* for building GSM-based SCAs, an example *implemented system* (Ripley the Robot), quantitative and qualitative *evaluations* of the above system as well as a discussion on multi-level evaluations of such systems, a *theoretical design method* for such systems and an example of such a design, a theoretical discussion of the relation of GSMs to prominent

[15] For a discussion on constraints necessary for intention recognition, see [Mavridis2005a].

theories including a *proposal for empirically sound semantics*, and an outline of a multi-year research program for *extensions* of GSM-based SCAs.

In even more detail, the main contributions of this thesis are:

- *Motivation and Requirements for Grounded Situation Models* (chapters 1 and 4) - A theoretical contribution showing how GSMs can be motivated from a variety of open problems (shortcomings of current NLP, systematic SCA design, cognitive science-inspired robots, semantics), and a specific proposal for a set of basic defining characteristics that any GSM should possess.
- *A customizable Representation, together w associated Processes* (chapter 5, appendix C) - A computational proposal for a customizable representation and a set of standard processes, specified all the way down to the level of pseudocode, which can easily be translated to a general-purpose library that could be used on a multitude of embodiments and applications of SCAs, and which also provides significant leverage for a variety of future extensions (section 9.1).
- *A modular Implementation Architecture* (chapter 5) - A practical proposal for breaking down the implementation of a GSM-based SCA into a set of standard intercommunicating modules.
- *An Implemented System: Ripley the Robot* (chapter 6) - A real world Situated Conversational Assistant (SCA) implemented on the embodiment of Ripley the Robot, useful for exploring real-world complication os such systems, evaluation, and further extensions on top of the existing software.
- *Quantitative and qualitative evaluations of Ripley* (chapter 7) - As well as a discussion on evaluating GSM-based systems and using psychological tests which could be used as a guideline to the multi-level evaluation of other similar systems in the future.
- *An outline of a multi-year research program for extensions of SCA-based GSMs* (section 9.1) - Providing short descriptions of more than 30 projects organized in 10 areas, all of which are heavily leveraged by the GSM proposal and the existing implementation and code, and which could potentially bring the state-of-the art of Situated Conversational Agents much closer to the ideal SCA, and provide true cooperativity through human-like natural language communication, intention recognition, and activity coordination.
- *A discussion of the relation and side-benefits of GSMs to prominent AI, cogsci, and semantics theories - including a proposal for empirically sound semantics* (chapter 8) - Which provides interesting insights to a number of central questions that belong to these areas.

Furthermore, this text also contains the following less central possible contributions:

- *A rough method for designing GSM-based systems* (appendix D) - That has already demonstrated its use by deriving a quasi-minimal GSM design that can pass all of the Token Test requirements, and which can be used for designing future systems, given either a behavioral specification or a psychological test.

9.2.2 Expected Significance of this Thesis

The expected significance of this thesis will be discussed here, on the basis of its already exhibited *novelties* above the state-of-the-art of conversational robots, and also on the basis of the *applicability*, *reusability*, and *scalabality* of the contributions to future systems (Situated Conversational Assistants). Expected significance will also be assesed in terms of impact on *theoretical* aspects, and of the use of this thesis as a basis for a multi-year *research plan*.

Novelties

Regarding novelties, the implemented GSM-based system (Ripley the Robot) has at least three types of novel capabilities as compared to other conversational robots (as was discussed in more detail in section 6.5.1):

A) Imagining situations described through language:

B) Remembering and resolving temporal referents:

C) Quantifying and expressing confidence in beliefs:

Furthermore, notice that all of the cited conversational robots are far from being able to exhibit the behavioral repertoire required for passing a human psychological test such as the "Token Test", while as discussed in section 7.4.1 Ripley already exhibits behavior comparable to parts of the Token Test and in some directions even surpasses the required behavioral repertoire. Furthermore, a theoretical derivation of a quasi-minimal design for a GSM-based SCA that can pass the whole of the Token Test has already been produced by the author, and can be found in [Mavridis2005c]. The GSM proposal has been *instrumental in attaining these novelties.*

Applicability

The GSM-proposal has proved its applicability to the domain of table-top manipulator robotic assistants, through Ripley. As discussed in section 8.5.1, and by glancing through the future extensions listed in this chapter (section 9.1, one can foresee the possibility of the application of the GSM proposal and also of a standardized codebase (see section 9.1.9) to many different application domains: household assistants, humanoids, intelligent buildings and cars and more.

Reusability

As can be deduced from section 9.1.11, not only the general ideas of GSMs but also a codebase and the existing implemented system (Ripley), can not only be extended in various directions, but can also provide significant leverage towards these extensions. Furthermore, due to the fact that changes of embodiment require only partial changes within the GSM-based SCA (section 8.5.1), one can foresee reuse of various parts of designs and code across embodiments.

Scalability

As discussed in section 9.1.11, the question of scalability remains to be proved in practice, and the foreseen way of achieving real scaling passes through the prerequisites of *hierarchical trainability* and *open sourcing*. I.e. scalability will only be achieved through learnability / trainability, and through the establishment of a big enough developer as well as user base.

Theoretical significance

This thesis brings together and interrelates ideas from traditionally somewhat disparate areas, and attempts to provide a unifying framework for them. Aspects of semiotics, semantics, artificial intelligence, cognitive science, philosophy and of other areas are interwoven together. Furthermore, specific theoretical proposals are made that might be proven to be quite fruitful in the future, such as the "Empirically Sound Semantics" proposal in section 8.5.5.

This thesis as a basis for a multi-year research plan

A brief glance through section 9.1, will easily illuminate the possibility of the creation of a 5-10 year research plan for the members of a laboratory team in industry or academia, that would provide a practical pathway for bringing the vision of the creation of truly cooperative SCAs closer to its realization.

9.3 Conclusion

We have come a very long way since the first chapter. And all the way through, we have tried to motivate, explicate, and discuss the GSM proposal and its implications and extensions. In chapter 1, we started by introducing Situated Conversational Assistants, discussing symbol grounding, situated language and situation models, and explaining why traditional NLP cannot "plug-and-play" in SCAs. In chapter 2, we took a higher-level view, and discussed the semiotic triangle and descriptions of Signs, Minds and Worlds. Then, we entered the core part of the GSM proposal, and focused gradually from the generics to the particulars: we discussed GSM-based agent models in chapter 3, motivated and introduced GSMs in chapter 4, discussed the specifics of the proposal (representations, processes, modular architecture) in chapter 5, and provided an example of a real-world implementation on Ripley the Robot in chapter 6. Then, we entered the third and final part of the thesis, and discussed evaluation, design methods, as well as theoretical implications (chapters 7-8), and finished with a discussion of future extensions, contributions, and significance in chapter 9.

The ultimate vision was stated upfront: truly cooperative Situated Conversational Assistants (section 1.2.4). The GSM proposal has already been very helpful towards the vision of truly cooperative Situated Conversational Assistants, and it provides easy integrability as well as significant leverage towards the extensions that would enable us to fulfill the prerequisites of a truly cooperative ideal SCA, whose realization should not seem as such a distant dream anymore.

9.3.1 Epimetron

In the introductory section, we stated upfront and discussed the vision of this thesis - SCAs. But taking a higher viewpoint, one might ask: by why do we need SCAs for humanity? Or taking an even higher viewpoint, one should ask: where should SCAs and intelligent robots fit within the Order of the world? This is a question which would deserve deep consideration, and significant future-reaching examination. Thus, I will end this thesis not with an answer, but with a wish:

May the day come when humans and their artificial partners can co-exist in harmony and peace, enabling each other to utilize their full potential, and most importantly, helping each other towards fulfilling their true purpose of existence.

Appendix A

Ripley's Behavioral Repertoire

The system responds to
the following utterance classes:

1. Questions:
(present/past)
<question> is/are the <obj des>
<question> was/were the <obj des> when <event des>

2. Action requests:
(referent description in present/past)
<action> the <obj des>
<action> the <obj des> when <event des>

3. Imaginer request:
Imagine/There is a <obj des>

4. Location-centered "look at":
Look at <location>

5. Viewpoint:
<action> the one on my/your left/right

6. Basic mode switching:
Wake up, sleep, relax, look at me/the table

Types:

<question> belongs to {where, how big, what color}

<action> belongs to {touch, pick up, hand me, look at}

<obj des> contains <size> <color> <object> <locus>

<event des> contains <actor> <event type>

<size> belongs to {small, medium, large}

<color> belongs to {red, green, blue}

<locus> belongs to {center, left, right, top, bottom}

<actor> is either an <obj des> or <agent part>

<agent part> belongs to {my, your} × {head, arm}

<event type> belongs to {appeared, disappeared,
started moving, stopped moving,
came in view, came out of view}

7. Recent extensions:
Indexicals:
<obj des> extended to {this} + fingerpointing
Touch events:
<event des> extended to <actor> <event type> <patient>
<event type> extended: {started touching, stopped touching}
Multilingual support: English/Greek(auto language switching)

Appendix B

The GSM-based agent model in detail

Here we will have a closer look at the components of the GSM-based Agent Model introduced in section 3.1. The block diagram of the model is repeated here for easy reference (figure B-1).

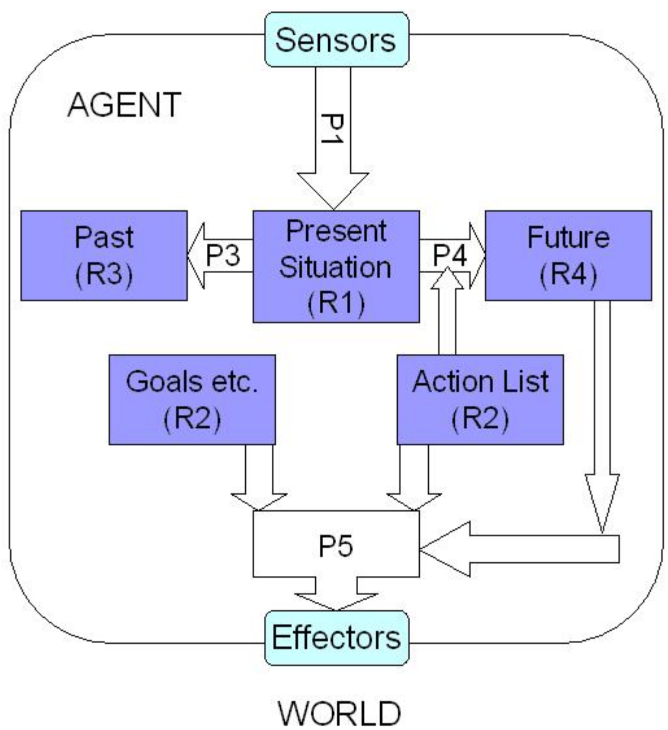


Figure B-1: Block diagram of the proposed agent model

.

Situation model, sensory updates (R1, P1), and ExternalReality-to-Sensor model (M1)

The situation model contains a representation of the current situation, as known by the agent. This representation decomposes the situation into objects (potentially agents), which are described through a set of properties, that have continuous values but are also quantized into categories.

For the discussion that follows, we will provide a quick illustrative example. Consider the situation model contents of an agent equipped with vision and sensing a two-dimensional image, containing a small green and a large red ball on a white background. The illustrative example will be an idealized example; we will not focus on all the well-known problems of computer vision and color perception. Instead, we will focus on how an observer-independent description of external physical reality (such as some sort of an infinitely-precise 3D model), is gradually transformed to the information contained in a verbal description (such as the string "a small green ball and a large red ball"), through a pipeline of four well-defined stages of transformation. We will also explicate the representations that will arise at the output of each stage, which will also feed the next stage[1].

For the construction of the situation model, we propose a three-stage sensor-to-situation model process (or equivalently, a four-stage externalreality-to-situation model process, if we include M1), with the following stages:

Reality-to-SensoryStream (M1): Spatiotemporally local external reality is projected into "sensory streams"[2]] - for example, the physical model of the two balls is transformed to a video stream at the output of the camera of the agent (or a neural stream at the optical nerve).

Sensorystream-to-IndividuatedStream (P1a): Sensory streams are segmented to "objects" - in the example: each frame contained in the video stream is segmented to three sets of pixels (or neural patterns) - belonging in this case to the green ball, red ball, background.

IndividStream-to-ObjectPropertyList (P1b): All of the segments of the sensory streams that resolve to a common object are collected, and object properties are calculated on their basis - For each object, the above pixels are used in order to calculate size, central position, average color. For example: the average color is found to be (156, 24, 22) for the red ball on some color space

ObjectPropList-to-ObjPropCategList (P1c): Continuous-valued or fine-granularity property values are quantized to meaningful categories -Color is quantized to "red", "green", "yellow" etc., size to "small", "large" etc.

Notice here, that the contents of the situation model, having been produced by this four-stage pipeline, are essentially a "personalized" subjective view of external reality. Thus, in general one can try comparing the "personalized" subjective realities of two agents, and ask what sort of relation exists among them (see section 3.2).

[1]The final output of the fourth stage of the pipeline, will contain the information similar to that which is encoded in the "categorical layer" of the triple-layered GSM, while the output of the third layer will contain information similar to that which is encoded in the "stochastic layer" or the "continuous layer". Here for simplicity we will assume total certainty in the "stochastic layer", effectively making it equal to the "continuous layer", and effectively collapsing the discrete distribution over categories of the original "categorical layer" to a single (certain) category.

[2]Often called the "sensory projection" stage.

Not only passive objects, but also agents in the situation model

In some cases, the situation model contains special descriptions for agents. These include not only descriptions of the agents as a physical object (body), but are also augmented by embedded estimated agent models of them[3]. Such estimated agent models of others contain estimates of their situation models, affective state, goals, as well as the personalized processes operating on them. These estimated situation models of others are a catalyst towards effective coexistence and cooperation: on the basis of them, one can decide when he needs to inform the other agent, one can predict the effect of possible actions on the other agent or its future behavior etc. More on embedded situation models in 4.3.2, 5.1, and 6.3, and other sections.

Past States and maintanence (R3, P3)

The process responsible for past storage, stores selected snapshots of agent state, including situation model state, possibly partial or compressed. It might also discard some of these snapshots after some time or further compress them for economy. Furthermore, this process recognizes and maintains a log of "events". In this proposal, by "events" we denote important changes in the commonsense comprising the agent model. These changes might either be instantaneous or have duration. Such changes might correspond to first-time appearances of external objects, specific movement patterns of them etc. An event description contains a temporal descriptor, a general event type, a list of participants (agents/objects), and further relevant information. For example, an event descriptor might be:

time: t=t0 to t=t1
type: sound event (type No. 14)
participants: bell (object No. 37)
parameters: ringing sound, high amplitude

Events effectively landmark / segment the stream of the past, and can be referred to through linguistic descriptions in a similar way that objects can be referred to: "when the green object appeared" is a reference to an event, while "the small blue object" is a reference to an object.

Future Predictions (R4, P4)

By having enough exposure to the physical law of his local environment (throughout evolutionary or personal history), the agent can learn how to make useful predictions for the future given the past. For example, many living beings seem to make successful kinetic predictions for physical objects. However, when moving from physical objects to agents, behavioral predictions become much more difficult, and treating agents as special physical objects is often insufficient. When predictions about future agent behavior must be made, the embedded estimated agent models of others (if any) prove

[3] As introduced in the first chapter.

to be invaluable - modeling bodies of agents as purely passive physical objects with no self-movement usually provides very poor predictions, unless the agent being modeled is "dead" or "unconscious". Predictions about the future actions of the observed agent are necessary, and the embedded estimated agent model could be able to provide these.

Action list and Selection Process (R5, P5), Internal State (R2, P2), and Action - to - External Reality model (M2)

At any given time instant, a decision about whether (and which) actions must be taken must be made. Here, I will briefly review the five levels of complexity of the action selection process introduced in chapter 2, also considering together with them their associated internal state descriptions. Notice that complexity increases greatly as the flexibility and the temporal planning horizon of the action selection process increases. Also notice how intricately bound affective state, action selection and goals.

Level-0: Reflexive

At the simplest end of the spectrum, the other agent might simply use a stimulus-response table for action selection. In order to align such a table with our general agent model proposal, we break it up to two components: an instantaneous stimulus-to-SituationModel table (in P1), and a fixed SituationModel-toaction table (in P5). We might further assume that the situation model has no object permanence. Therefore, stimuli will instantly produce transient situation model contents which will in turn produce genetically hardwired responses that cannot be modified. This is what is normally termed the "reflexive" level, and even some human behaviors belong to it ("hammer" reflex etc.). So in this case the action selection mechanism is a simple current situation model - to - action table, and the current situation model depends only on current sensory input.

Level-1: Reflexive with internal state

Now, the internal state is allowed to play a role in action selection. Effectively, instead of one current situation model - to - action table, we have n such tables (in P5), one for each possible value of the "affective" internal state (R2). Notice that stimuli do not only change the situation model contents, but also change the affective state. The SituationModel-to- InternalState process (P2) achieves this. It is interesting how natural it is for humans to "read" emotions in a simple automaton with a few discrete internal states and appropriately chosen actions. At level-1, the tables are still genetically fixed, and normally cannot change during the lifetime of the individual.

Level-2: Conditionable reflexive

Here the effective (stimulus, state)-response tables are not fixed throughout the lifetime of the individual, but can be conditioned. Notice that some form of memory is necessary for conditioning. Otherwise, when the reward or punishment arrives, the stimulus-action pair that the organism had tried, and which should be promoted / demoted, will have been forgotten. Here, part of the initial (stimulus, state)-response table should encode a genetically-fixed "learning" mechanism; i.e. certain stimuli should be special (the "reward" or "punishment" stimuli), and they should produce internal responses that amount to re-wirings of the remaining (stimulus, state)-response table. There exist some almost universal reward (food) or punishment (electric shock) stimuli, that can help condition almost any such organism.

Level-3: Generally fixed short-term planning

Here the organism doesn't only select current actions; but also future actions, for some time window ahead of him. At this level, explicit goals that might change enter the agent model. Therefore, it becomes worth modeling action selection as a rational decision procedure. Actions in the action list are considered in turn, and their anticipated effects predicted (through P4). Then, the resulting predicted world states are evaluated according to the pursued goal. The actions that result in preferable world states form the selected plan, which is then executed.

Level-4: Flexible longer-term partial planning

This is a more realistic model for human action selection. We make plans, we leave details open until they are fixed later, we partially commit to our plans, but also change or abandon them in case new information demands so. Thus, this level contains the processes of level-4, but also contains the possibility of leaving plan details open and coordinating / revising plans before execution.

A hierarchy of minds

Here, we will briefly propose a hierarchy of levels of complexity of minds, which follows the complexity of the action selection process hierarchy, but also prescribes more characteristics of the agent, such as whether an explicit episodic memory should be expected to exist etc[4].

Level-0: Reflexive, memoryless

Essentially memory-less, or with very short temporal horizon (future/past). Situation model does not have any "object permanence"

Level-1: Reflexive with internal state

Here the recent past is accumulated into "internal states", and no explicit episodic memory exists.

Level-2: Conditionable reflexive

Here there might effectively seem to be some "mental maps" of the environment, and rudimentary memory capabilities.

[4]Notice that this hierarchy is not restrictive; it just proposes what overall capabilities one should expect to find in a living organism that has a certain action selection method. For example, the presented GSM implementation on Ripley the Robot is misaligned with this hierarchy; Ripley has an explicit episodic memory, short-term prediction capabilities, and rudimentary estimated models of others (level-4 characteristics), but still does not use explicit planning (i.e. has level-1 action selection).

Level-3: Generally fixed short-term planning

Relatively detailed situation models. Some form of episodic memory. Short-term prediction capabilities for objects and some familiar agents. However, no explicit embedded estimated agent models of others or the self.

Level-4: Flexible longer-term partial planning

With embedded estimated agent models of others or self.

Recap

In this appendix, we had a closer look at the components of the GSM-based AgentModel introduced in chapter 2. More information was given on the role of the proposed processes and representations, as well as a rough hierarchy of minds conforming to the proposed model.

Appendix C

Pseudocode for GSM processes

In this appendix:

1. Notation is introduced.
2. The processes comprising the GSM operational cycle are analyzed in pseudocode, and various options are discussed

The GSM operational cycle was described in section 5.2.1, and conceptual descriptions of the processes given here in pseudocode can be found in section 5.2.2.

Note: A full Bayesian treatment might also be attempted (not described here).

Notation

Assume a property dimension (for example, color) with M-dim measurements (M=3 for color), and N verbal categories (N=8 for Red, Green, Blue, Cyan, Magenta, Yellow, White, Black)

Stochastic Layer (L1): Continuous-variable M-dimensional distribution p(x), where x is an M-dim vector Might be approximated by discrete-variable M-dimensional distribution, with N1 x N2 bins, covering the expected range of values (rectangular bins by default - other partitions possible too)

Continuous Layer (L2): xcont = M-dimensional vector, continuous-valued

Categorical Layer (L3): Discrete-variable 1-dimensonal distribution P(C), where C belongs to 1 to N

Classifier[1]: Function C = F(x) where C is integer from 1 to N, and x is M-dimensional vector (i.e. Verbal Category C = F(measurement value x) might be explicitly known, or realized as neural net, SVM, nearest-neighbor etc.)

[1]The "categorical classifier".

InitL1L2L3) Initialise Distributions

Stochastic Layer:
p(x) = ?
either:
1)Non-informative prior
For example, uniform distribution covering the expected range of values
or:
2)Empirically derived prior
Accumulate experience from colors of objects seen so far
I.e. estimate distribution by all the incoming x' s of the past sensory experiences

Categorical Layer:
P(c) = ?
Fed by stochastic layer through the stochastic-to-categorical feeding process (see below)

Continuous Layer:
Xcont = ?
Fed by stochastic layer through the stochastic-to-continuous feeding process (see below)

L1-to-L3) Stochastic-to-Categorical feeding process

```
For C=1…N (i.e. iterate through verbal categories)
        P(C) = Integral of p(x) over those x-values such that F(x) = C
        (In the case of discrete-variable binned approximation to p(x),
        this integral can be simply approximated by iterating over all bins,
        and summing the probability contribution of only those bins
        whose center x indeed gives F(x) = C for the particular verbal
        category C under consideration, i.e.
        sum = 0
        For I1 = 1…N1 //(iterate over bins)
                For I2 = 1…N2
                        …
                        If F(center(I1, I2, …) )==C
                                sum = sum + p(I1,I2,…)
                        endif
                        …
                Next I2
        Next I1
        P(C) = sum
        )
Next C
```

L3-to-L1) Categorical-to-Stochastic feeding process

```
For C=1…N (i.e. iterate through verbal categories)
        Either:
        1) Non-informative prior
        Distribute P(C) over p(x) for those x-values such that F(x) = C,
        so that the integral of p(x) over those x-values is equal to P(C)
        (One reasonable choice of how to distribute: distribute uniformly,
        i.e. set p(x) for those x-values such that F(x) = C equal
        to P(C) / Area(region of x's such that F(x) = C)
        In the case of discrete-variable binned approximation to p(x):
        Find number of bins such that F(center(I1,I2,…)) = C, and call it Nc
        And then for those bins assign p(I1,I2,…) = P(C)/Nc, i.e.
        (of course the code found below can be optimized much more!)
        Nc = 0
        For I1 = 1…N1 //(iterate over bins)
                For I2 = 1…N2
                        …
                        If F(center(I1, I2, …) )==C
                                Nc=Nc+1
                        endif
                        …
                Next I2
        Next I1
        For I1 = 1…N1
                For I2 = 1…N2
                        …
                        If F(center(I1, I2, …) )==C
                                p(I1,I2,…) = P(C) / Nc
                        endif
                        …
                Next I2
        Next I1
        Or:
        2) Empirically-derived prior
        Estimate p(x | verbal label = C) from human training data
        Then weigh this estimate by P(C) and contribute accordingly to p(x).
        This might in essence admit partial overlap of the measurement regions
        corresponding to different verbal categories.
        Recall that as the classifiers perfectly partition the measurement space,
        we are silently admitting that such an overlap does not exist or that
        we can anyway tolerate the residual classification error that would
        result from this "fuzziness" of the category boundaries.
Next C
```

Senses-to-L1) Sensory measurement-to-Stochastic feeding process

Let's assume incoming measurement xnew
Before the measurement, we had stochastic layer distribution p(x) = pold(x)
After the measurement, the updated distribution will be p(x) = pnew(x)
Thus, we have to choose an update function fupd such that:
pnew(x) = fupd(xnew, pold(x)).

A simple solution:
Assume "windowing function" wxnew(x) (like parzen windows)
wxnew(x) should be centered at xnew and have integral equal to one.
Choose lambda belonging to [0…1] (confidence in new measurement)
Then, define fupd such that:
pnew(x) = (1-lambda) * pold(x) + lambda * wxnew(x)
In the case of discrete-variable binned approximation to p(x):
Assume discrete-variable binned window function wxnew(I1,I2…)
wxnew should be centered at those I1, I2, …,
such that center(I1, I2,…) approx equal to xnew

```
// Find bin coordinates (Ixew(1), Ixnew(2),…)
// corresponding to incoming measurement xnew
For I=1…M
        Ixnew(I) = MeasurementToIndices(xnew)
Next I

//Create centered window function wxnew
//with center at (Ixew(1), Ixnew(2),…)
//assume prototype window function w centered at (Ic1, Ic2,…)
For I1 = 1…N1 //(iterate over bins)
        For I2 = 1…N2
                …
                //of course, care for non-negative matrix indices
                //should be taken here etc.
                wxnew(I1,I2,…) = w(I1 – Ixnew(1) + Ic1 , … )
                …
        Next I2
Next I1

//Update stochastic layer by weigh-adding new window function
For I1 = 1…N1
        For I2 = 1…N2
                …
                p(I1,I2,…) = (1-lambda) * p(I1,I2,…) + lambda*wxnew(I1,I2,…)
                …
        Next I2
Next I1
```

Other analytical or empirical solutions for fupd(xnew, pold(x))
can of course be used too.

Words-to-L3) Speech derived information-to-Categorical feeding process

Hard update:
Here we are assuming that new verbal information totally overrides existing information (sensory/verbal) totally.

Either:
Assume incoming verbal category Cnew.
Then P(Cnew) = 1
And set all other P(C) such that C!=Cnew
to be equal to zero

Or:
Assume incoming verbal category Cnew
with confidence lambda belonging to [0…1]
Then set P(Cnew) = lambda
And set all other P(C) such that C!=Cnew
to be equal to (1-lambda) / (N-1)

Or:
Assume incoming verbal category distribution Pnew(C)
Then just copy Pnew(C) into P(C) for every C belonging to 1…N

Or:
(if empirical data p(x | verbal label = C) available)
Set this to stochastic. Feed through to categorical

Soft update:
If we assume that new verbal information has to be blended with existing:

Either:
Assume incoming verbal category Cnew,
and confidence to new information = lambda belonging to [0…1]
Then P(Cnew) = (1-lambda) * P(Cnew) + lambda
And set all other P(C) such that C!=Cnew
to be equal to P(C) = (1-lambda) * P(C)

Or:
(if empirical data p(x | verbal label = C) available,
and confidence to new information = lambda belonging to [0…1])
Update stochastic as:
pnew(x) = (1-lambda) * pold(x) + lambda * p(x | verbal label = C)
Feed through to categorical.

Various combinations and variants of the above can be used.

L1-to-L2) Stochastic-to-Continuous feeding process

Some averaging function chosen,
accepting an M-D distribution p(x),
and returning an M-D vector xcont
For example, one can choose mean, median etc.
xcont = ExpectedValue{p(x)}

L3-to-Words) Categorical-to-verbal description feeding process

Here we seek a function from a discrete-variable distribution
to a set of templates.
Let's assume verbalization of categories W(C)
so that for example W(1) = "red" etc.
A decision tree with conditionals on the probability of the two most probable categories and the entropy is used, producing language generation templates such as:

```
If (Pmax1>Threshold1)
          Sentence = W(C such that P(C) = Pmax1)
Else If (Pmax1>Threshold2)
          Sentence = "Most probably " + W(C such that P(C) = Pmax1)
Else If (Pmax1>Threshold3)
          Sentence =  "Most probably " + W(C such that P(C) = Pmax1) +
                        " but maybe not"
Else If (Pmax1>.4 and Pmax2>.4)
          Sentence = "Either " + W(C such that P(C) = Pmax1) +
                        "`or " + W(C such that P(C) = Pmax2)
.... (etc.)
```

Such a tree can be empirically learned,
Or some other empirically-trained generation model can be derived.

Also, an "inversion" of the above mapping can be used in feeding incoming descriptions such as "There is blue object most probably at the left" to the categorical layer (this was covered in the previous section – this is just an addendum to that section). Again one must use a simplifying assumption to overcome the one-to-many aspect of the inversion.

L2-to-actionparam) Continuous-to-action parameter feeding process

Just supply xcont as the action parameter.
For example Lookat(OBJi)
will need position(OBJi)
which will be taken by xcont of position of object i

DiffuseL1) Stochastic layer diffusion process

Assume old distribution pold(x)
And new distribution pnex(x)
Choose update function fdiff such that:
pnew(x) = fdiff(pold(x))

For example, for variance-parametrised distributions, increase the variance.

In the case of discrete-variable binned approximations of p(x),
One solution is to "low-pass filter" the distribution (in signal processing terms).
In order not to get into filter theory, some simple solutions suffice.
For example, one could just arbitrarily choose a rectangular window for the filter mask, and specify the rectangle lengths in bins. Then, the filter mask slides over the distribution, centered on each element (pixel, voxel etc.) in question.
Instead of non-uniform mask coefficients, one can again use a simplifying assumption: all coefficients equal to one over the mask area. Then in essence, the filter performs local averaging in a rectangular neighborhood around the center of the mask.
For example, for the simple one dimensional case with window size = 3:

```
For I=1…N
        pnew(I) = (1/3) * (pold(I-1) + pold(I) + pold(I+1))
Next I
```

Or for two dimensions and a 3x3 window:

```
For I1=1…N1
        For I2=1…N2
                pnew(I1,I2) =  (1/9) * (pold(I1-1,I2-1) + pold(I1,I2-1) + pold(I1+1,I2-1) +
                                        + pold(I1-1,I2) + pold(I1,I2) + pold(I1+1,I2) +
                                        + pold(I1-1,I2+1) + pold(I1,I2+1) +pold(I1+1,I2+1))
        Next I2
Next I1
```

In order to control the filter averaging effect, one could add a lambda coefficient belonging to [0…1].
Then, the diffusion function looks like:

```
For I=1…N
        pnew(I) = (1-lambda) * (1/3) * (pold(I-1) + pold(I) + pold(I+1)) +
                        + lambda * pold(I)
Next I
```

In essence, this increases the contribution of the central bin, and decreases the averaging effect. Of course, in the above example we have not considered the treatment of negative indexes. One should also always take care of boundary conditions; one simple hack solution is to change the weighing when the mask is near the boundary, for example:

```
pnew(1) = (1/2)*(            pold(1) + pold(2))
while
pnew(2) = (1/3)*(pold(1) + pold(2) + pold (3))
```

In our implementation, a variable-size rectangular window is used, and the lambda coefficient is hand-tuned for reasonable diffusion speeds (higher for position, less for color, less for size). Of course lambda as well as the window size or shape (or even each separate coefficient) could be empirically tuned, given data on the statistics of temporal change of property values (position, color, size etc.)

Appendix D

Designing SCAs

Here, as our ultimate purpose, we will try to provide a brief sketch of how a behavioral specification (either "grown" from the seed of a psychological test or provided otherwise[1], can be used as the basis for the design of an SCA, in an incremental manner, which under certain conditions can produce "minimal" or "quasi-minimal" designs. Such a systematic design method for SCAs is useful in many ways: first, it simplifies the manual design of such systems. Furthermore, in the future one could foresee a semi-automatization of the process, in the context of CAD systems for SCAs with reusable components (see section 9.1.9). Also, the "minimality" or "quasi-minimality" of the derived designs is useful for economy and leanness, and can be used as an argument against the adhoc-iness of other approaches. When a designer wants to claim that his design is better than somebody else's, some form of optimality proof always comes in handy.

On our way to discussing the proposed sketch of a design method, we will first introduce a partial ordering of GSM-based systems (section D.1), and then, in section D.2, we will introduce a generic incremental augmentative method for producing minimal or quasi-minimal designs, with quasi-minimality defined explicitly in that section. In section D.3, we will propose a criterion for testing the local minimality of complex systems that can be decomposable into parts, which we will term the "Jenga" criterion. Finally, in section D.4, having gathered the right tools on the way, we will deliver what was promised: a sketch of a practical design method for producing quasi-minimal SCA's, given a behavioral specification (such as the one in section 6.2) or a psychological test (such as the Token Test which was discussed in section 7.4).

D.1 A partial ordering of GSM-based systems

In order for someone to claim optimality of a particular design, given a universe of designs, a testable optimality criterion is needed. This optimality criterion might involve a scalar-valued metric - in that case, the optimal design is the one which is guaranteed to maximize this metric (if the metric is a "fitness" metric) or minimize it (if it is a "cost" metric). But in reality, when searching for an optimal design, the absolute value of the

[1]For example, such a spec might be the one given in section 6.2.

metric is of no particular importance; all that we really care for is the ordering of the metrics corresponding to the designs. Remember: we are just looking for a single "best" design; we don't directly care about how much better it is than others. This ordering needn't even be a total ordering - not all designs need to be comparable with every other design. As long as there is a single design that can stand at the top of the partial ordering lattice, the existence of pairs of designs which are not comparable to each other further down the lattice poses no problem. And we do not even need the whole lattice; again, remember, we are just looking for a single "best" design: as long as we can prove that any other design is worse than the winner, we don't directly care who comes second or who comes third.

Here we will thus take this route - we will propose an ordering criterion for the comparison of two grounded situation models. The possible outcomes of the comparison of any two designs D1 and D2 that differ from each other are three: D1 > D2 (D1 is larger than D2), D1 < D2 (D1 is smaller than D2), D1 * D2 (D1 is not comparable to D2 - thus the partiality of the ordering).

But how are we going to define this partial ordering for the case of two GSM designs? According to chapter 3 and appendix B, we have seen that any GSM-based design can be decomposed into a number of distinct parts (representations Ri and processes Pj). We will define the partial ordering in the following manner in terms of the parts of two different designs D1and D2:

D1 > D2: For every representation Ri of D1 it holds that Ri of D2 is a subset of it, and for every process Pj of D1 it holds that the equivalent Pj of D2 is a subset of it, where by A subset of B we mean that A contains all of B plus possibly more.

D2 > D1: The converse statement: all reps and processes of D1 are subsets of those of D2.

D1 * D2: If neither D1>D2 nor D2>D1.

As we shall see, the crucial point for the meaningful applicability of the above criterion is the notion of "incremental augmention" of designs: Say that we have two design specifications of increasing difficulty: S1 and S2. Now, say that we have devised a design D1, that fulfills spec S1. Now, say that we are trying to devise a design D2 can not only fulfill S1, but also the more difficult S2. By "incremental augmentation" I mean: take the old design D1, and augment it (without changing the existing parts) in order to derive a new design that not only satisfies spec S1, but also spec S2. Notice that in this case, as D2 came out of an augmentation of D1, we are guaranteed that D2 > D1 given the partial ordering introduced above.

Now the question arises: how can we derive a minimal design using the above ideas?

D.2 A generic incremental augmentative method for producing quasi-minimal designs

D.2.1 Incremental design using the natural partial ordering

Consider that we are given a list of design specifications (S1, S2, ...). We start by producing a design D1 for the first specification S1, and then by some argument or test we prove it to be a minimal design for it. Then, we consider all possible augmentations D1i of our initial design that would enable it to cover not only the first (S1) but also the second specification (S1 and S2), and we choose the smallest augmentation D1best (according to the partial ordering criterion - i.e. the augmentation for which D1best < D1i for every i). Under what conditions will the resulting new augmented design (D2 = D1 augmented with D1best) be optimal for the augmented specification that includes both the first and second specification (S1 and S2)? The answer should not be difficult to spot - the resulting new design is optimal as long as:

(Condition 1) There is no design D2' that would include only part of the initial design (i.e. where D1 would not be a subset of D2), and which would suffice for both the requirements of S1 and S2.

But what happens if condition 1 does not hold? Then, we have not retained minimality through the augmentation process - in this case, what we have (D2) is not the minimal design that covers both S1 and S2, but instead the minimal design that can be derived as an augmentation of D1 that covers both S1 and S2: The minimal solution not out of all possible designs, but out of the subset of designs which include D1 as a part of them - thus, this is not the minimal design, but the minimal "augmented D1" design.

Extending this idea and iterating through the list of design specs (S1, S2, S3, ...), once we reach Sn we will have design Dn which covers (S1 ... Sn), and which is the minimal "successive augmentation" of D1. Notice that this design will also be the globally minimal design only if:

a) D1 was minimal for S1

b) condition 1 continued to hold at each augmentation step

Else, it will just be the minimal "augmented Dn-1" design, and not the globally minimal design.

D.2.2 Quasi-minimal designs

Such designs, which were produced by the above procedure of successive augmentations, and where condition 1 did not necessarily hold in all augmentation steps, we will call "quasi-minimal" designs[2].

D.2.3 Generic method for producing quasi-minimal designs

Thus, we propose the following generic method for producing either minimal (or at least quasi-minimal) designs: Given a list of design specs Si, i=1 ... n:

- Derive a minimal design D1 for spec S1
- Augment D1 with minimal parts D1' in order to cover spec S2
- Augment D2 with minimal parts D2' in order to cover spec S3
- And so on, until we reach the desired Dn.

The final design Dn will be globally minimal as long as condition 1 holds at every step, or else it will just be "quasi-minimal".

[2]Similar to "greedy".

But how reasonable is this method? In nature, evolution proceeds by quasi-continuous steps: patches are applied on top of existing subsystems, and / or existing subsystems are gradually modified to order to better adapt to new requirements. Thus, nature seems to be roughly following the proposed method, at least in the case of patches. In the case of GSM-based systems, as we have already decomposed the design into a set of processes and representations, each of which consist of well-defined parts that are augmentable, the above generic method seems to be quite suitable, as will be further illustrated by the design example that is discussed in detail in [Mavridis2005c], where a quasi-minimal design that passes the Token Test which is derived in detail. This design example that passes the Token Test that is described in [Mavridis2005c], is based on the GSM-specific adaptation of the above generic method, which will be given in the next section of this chapter.

D.3 A criterion for the local minimality of a given design

Now, having introduced a generic method that is guaranteed to produce minimal or quasi-minimal designs, we will introduce a criterion that can be used in order to test whether a given design is locally minimal - a criterion that can be used either on its own, or as a method of verification for designs produced by the generic design method described above.

This criterion will be termed the "Jenga" criterion, and it can guarantee that a proposed design is a local extremum in a design space. The criterion is simple to state:

a design is locally optimal if both of the conditions below hold:

a) it fulfills the design specifications

b) removal of any of its parts, will make it fail to fulfill the design specifications.

The intuition behind the criterion comes from the "Jenga" game; the removal of any single brick from the tower might make it collapse; a locally optimal tower (in terms of the number of bricks) is a standing tower with the property that, if any single brick was removed, it would collapse. This criterion will be applied to the final design that will result from the incremental procedure. If the criterion holds than it will prove local optimality; if it doesn't it can be used as a way to "prune down" the design. Finally, it is worth noting that a pathway towards easy further expansion is naturally provided through this methodology. When given new specifications, one just needs to start the incremental design procedure with the previous final design as the new initial design. The old final design will be augmented, in order to cover the new requirements, and will not need to be expensively totally redesigned ex nihilo.

Let us summarize what we have seen so far. A natural partial ordering of GSM designs was proposed. This ordering does not necessitate the explication of cost values for components of designs. It most importantly enables us to use an incremental design procedure that guarantees quasi-minimality of the resulting design as long as the augmentations are the smallest possible, and also to guarantee not only quasi- but also full minimality as long as a condition holds. We then discussed whether the generic method is realistic. Finally, a criterion for local optimality of a design was proposed - the Jenga criterion. This criterion can be applied to a derived design that results from the

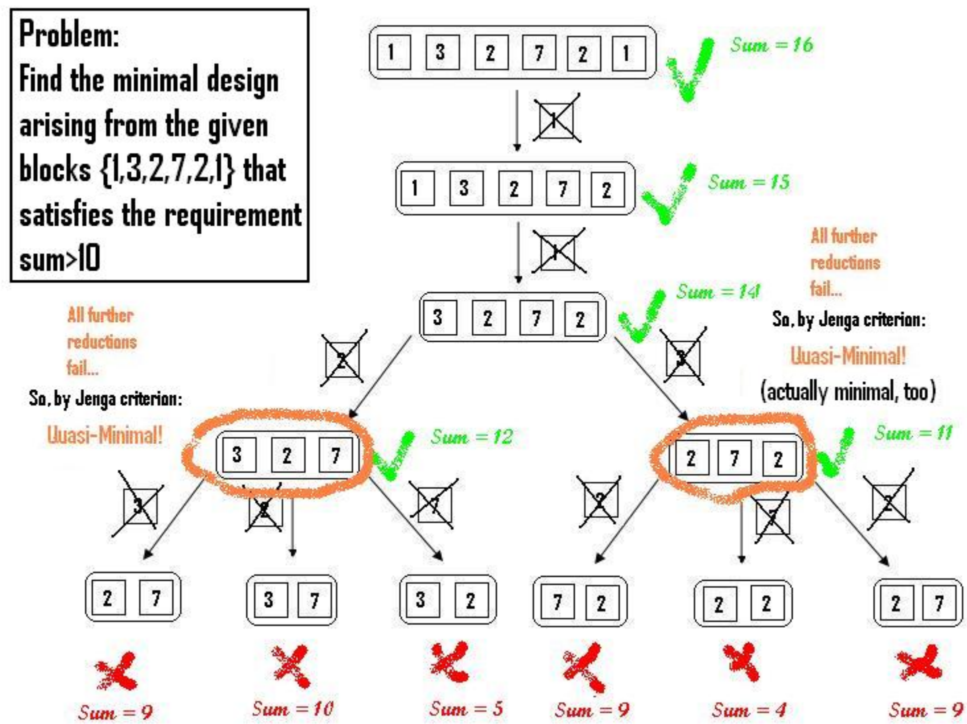


Figure D-1: An illustration of the Jenga Criterion, applied to checking the quasi-minimality of solutions for a toy problem: given the blocks [1,2,3,7,2,1], find the minimal design that satisfies the requirement sum > 10

incremental design procedure, to prove its local optimality, or to help to further prune it down. Regarding the virtues of the Jenga criterion, it is simple to state, easy to apply, and can be used directly for the case of GSM-based designs, as we shall see. When a design is produced through the design method that we will propose (or otherwise), it can then be proved as quasi-minimal through an application of the Jenga criterion - and this is its real value. This criterion will be utilized in the next section, as the final step of the promised proposed methodology that will be used for deriving quasi-minimal GSM designs, given a behavioral spec or a psychological test, which resulting designs are furthermore naturally expandable.

D.4 A sketch of a practical design method for quasi-minimal SCAs

The proposed incremental method starts from either a psychological test or a behavioral specification, and in the end creates a quasi-minimal SCA that passes the test and the wider domain of the test / meets the specification, where local optimality is also tested for the final end-product through the Jenga Criterion defined in the previous section. During the application of the method, the design under derivation is incrementally augmented, and all the intermediate designs are also quasi-minimal.

The steps of the method are the following:

Starting from a psychological test:

S1) Roughly arrange the questions in groups in order of increasing difficulty.

S2) Starting from the easiest, consider each of the above groups in turn:

S3) Use the questions of the group under consideration as a "seed" for growing the targeted area that was targeted by them[3].

S4) Using the general GSM-based agent model presented in chapter 3[4], consider each representation / process specified (Ri/Pi), and augment[5] it with content that enables them to fulfill the requirements of the behavioral area targeted by the group of questions under consideration.

S5) If there is a way to fulfill the requirements of the previous groups of questions AND those imposed by the current with content of less size[6] than the originally plus the augmented, use this way[7].

S6) If any question groups are left, go back to S3. Else, proceed.

S7) Apply the Jenga Test to verify that indeed you have a quasi-minimal design, by considering subtracting any part of the derived design and showing that the smaller design would not pass the test.

Step S5 above[8] is needed only when we want to target full minimality, and not only quasi-minimality.

The above method can also be modified for the case of starting from behavioral specifications, and not psychological tests. In that case, step S3 is not required - the spec does not need "growing", and all of the previous steps apply, by replacing "questions" with "parts of spec".

A detailed example of the use of the above proposed method is given in [Mavridis2005c], for the case of deriving a minimal SCA that passes the "Token Test", which was introduced in chapter 7.4.

D.5 Recap

In this appendix, as our ultimate purpose, we tried to provide a brief sketch of how a behavioral specification (either "grown" from the seed of a psychological test or provided otherwise, can be used as the basis for the design of an SCA, in an incremental manner, which under certain conditions can produce "minimal" or "quasi-minimal" designs.

On our way, we first introduced a partial ordering of GSM-based systems (section D.1), and then, in section D.2, we introduced a generic incremental augmentative method for producing minimal or quasi-minimal designs, with quasi-minimality defined explicitly in that section. In section D.3, we proposed a criterion for testing the local mini-

[3] In the the manner discussed in the previous chapter - the "three commandments" of section 7.5.

[4] The interested reader can also take into account the considerations exposed in the design example that passes all of the Token Test, which is described in detail in [Mavridis2005c].

[5] Augment without subtracting previous content.

[6] Fewer component parts of the representations Ri and processed Pj (see chapter 3 that comprise the system.

[7] If no such way exists, then condition C2 of the previous section holds. Else, we have to go through this step if we want guaranteed minimality and not only quasi-minimality.

[8] Corresponding to condition 1 mentioned before in section D.2.

mality of complex systems that can be decomposable into parts, which we termed the "Jenga" criterion. Finally, in section D.4, having gathered the right tools on the way, we delivered what was promised: a sketch of a practical design method for producing quasi-minimal SCA's, given a behavioral specification (such as the one in 6.2) or a psychological test (such as the Token Test, which was described in 7.4).

Bibliography

[Ahn and Picard2006] Ahn, H. and Picard, R. W. (2006). Affective-cognitive learning and decision making: The role of emotions. In *The 18th European Meeting on Cybernetics and Systems Research (EMCSR 2006)*.

[Appelt1985] Appelt, D. E. (1985). *Planning English Sentences*. Cambridge University Press, Cambridge, UK.

[Barsalou1999] Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22:577–660.

[Bickmore2003] Bickmore, T. W. (2003). *Relational Agents: Effecting Change through Human-Computer Relationships*. PhD thesis, Massachusetts Institute of Technology.

[Blake and Yuille1992] Blake, A. and Yuille, A., editors (1992). *ACTIVE VISION*. MIT Press, Cambridge MA.

[Bobick and Richards1986] Bobick, A. and Richards, W. (1986). Classifying objects from visual information. Technical report, Cambridge, MA, USA.

[Bratman1987] Bratman, E. M. (1987). What is intention? Technical Report 69, CSLI Stanford. This is a full TECHREPORT entry.

[Breazeal et al.2004] Breazeal, C., Brooks, A., Gray, J., Hoffman, G., Kidd, C., Lee, H., Lieberman, J., Lockerd, A., and Mulanda, D. (2004). Humanoid robots as cooperative partners for people. *International Journal of Humanoid Robotics*.

[Breazeal and Scassellati1999] Breazeal, C. and Scassellati, B. (1999). A context-dependent attention system for a social robot. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.

[Burke et al.2001] Burke, R., Isla, D., Downie, M., Ivanov, Y., and Blumberg, B. (2001). Creature smarts: The art and architecture of a virtual brain. In *Proceedings of the Game Developers Conference*.

[Cheng2006] Cheng, M. (2006). Human hand event detection for ripley the robot. AUP Project Report at http://web.media.mit.edu/ nmav/thesis.htm.

[Cohen and Perrault1979] Cohen, P. and Perrault, C. R. (1979). Elements of a plan-based theory of speech acts. *Cognitive Science*, 3:177–212.

[Crangle and Suppes1994] Crangle, C. and Suppes, P. (1994). *Language and Learning for Robots*. CSLI Publications, Stanford, CA.

[DiSimoni1978] DiSimoni, F. (1978). *The Token Test for Children*. DLM Teaching Resources, USA.

[Epictetus1928] Epictetus (1928). *The Encheiridion*, volume 218 of *Loeb Classical Library*. Harvard University Press, Cambridge, MA.

[Fellbaum1998] Fellbaum, C., editor (1998). *Wordnet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

[Gentner and Markman1997] Gentner, D. and Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, 52(1):45–56.

[Gibson1977] Gibson, J. J. (1977). The theory of affordances. In Shaw, R. and Bransford, J., editors, *Perceiving, Acting, and Knowing*. Lawrence Erlbaum, NJ. This is a full INCOLLECTION entry.

[Goldstone and Rogosky2002] Goldstone, R. and Rogosky, B. (2002). Using relations within conceptual systems to translate across conceptual systems. *Cognition*, 84:295–320.

[Gorniak2005] Gorniak, P. (2005). *The Affordance-Based Concept*. PhD thesis, Massachusetts Institute of Technology.

[Grosz and Sidner.1986] Grosz, B. and Sidner., C. (1986). Attentions, intentions and the structure of discourse. *Computational Linguistics*, 12:175–204.

[Guerra-Filho et al.2005] Guerra-Filho, G., Fermuller, C., and Aloimonos, Y. (2005). Discovering a language for human activity. In *Proceedings of the AAAI Workshop on Anticipation in Cognitive Systems*.

[Hanson et al.2005] Hanson, D., Olney, A., Prilliman, S., Mathews, E., Zielke, M., Hammons, D., Fernandez, R., and Stephanou, H. (2005). Upending the uncanny valley. In *Proc. AAAIŠs National Conference, Pittsburgh*.

[Harnad1990] Harnad, S. (1990). The symbol grounding problem. *Physica D*, 42:335–346.

[Hartley and Zisserman2000] Hartley, R. and Zisserman, A. (2000). *Multiple view Geometry in Computer Vision*. Cambridge University Press, Cambridge, UK.

[Hayek1952] Hayek, F. A. (1952). *The Sensory Order*. University of Chicago Press, Chicago, IL.

[Hermer-Vazquez et al.1998] Hermer-Vazquez, L., Spelke, E., and Katsnelson, A. (1998). Sources of flexibility in human cognition: Dual-task studies of space and language. *Cognitive Psychology*, 39:3–36.

[Horvitz and Klein1993] Horvitz, E. and Klein, A. (1993). Utility-based abstraction and categorization. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*.

[Hsiao2003] Hsiao, K. (2003). Early video demo of ripley the robot: beanbags. Video.

[Hsiao et al.2003] Hsiao, K., Mavridis, N., and Roy, D. (2003). Coupling perception and simulation: Steps towards conversational robotics. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*.

[Ishiguro and Minato2005] Ishiguro, H. and Minato, T. (2005). Development of androids for studying on human-robot interaction. In *Proceedings of 36th International Symposium on Robotics, TH3H1*.

[Ishiguro et al.2006] Ishiguro, H., Miyashita, T., Kanda, T., Ono, T., and Imai, M. (2006). Robovie: An interactive humanoid robot. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

[Itti et al.1998] Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259.

[Johnson-Laird1983] Johnson-Laird, P. N. (1983). *Mental Models*. Cambridge University Press, Cambridge, MA.

[Jurafsky and Martin2000] Jurafsky, D. S. and Martin, J. H. (2000). *Speech and Language Processing*. Prentice Hall, Englewood, NJ.

[Juster and Roy2004] Juster, J. and Roy, D. (2004). Elvis: situated speech and gesture understanding for a robotic chandelier. In *ICMI '04: Proceedings of the 6th international conference on Multimodal interfaces*, pages 90–96, New York, NY, USA. ACM Press.

[Kosslyn1985] Kosslyn, S. M. (1985). *Image and Brain: The Resolution of the Imagery Debate*. MIT Press, Cambridge, MA.

[Kuhn1962] Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.

[Laird et al.1987] Laird, J. E., Newell, A., and Rosenbloom, P. S. (1987). SOAR: an architecture for general intelligence. *Artificial Intelligence*, 33(1):1–64. Laird, J.Newell, A.Rosenbloom, P.Newell, A.

[Lamere et al.2003] Lamere, P., Kwok, P., Walker, W., Gouvea, E., Singh, R., Raj, B., and Wolf, P. (2003). Design of the cmu sphinx-4 decoder. In *Proceedings of EUROSPEECH-2003*.

[Langley and Laird2002] Langley, P. and Laird, J. E. (2002). Cognitive architectures: Research issues and challenges. Technical report, Institute for the Study of Learning and Expertise, Palo Alto, CA. This is a full TECHREPORT entry.

[Marr1982] Marr, D. (1982). *Vision*. Freeman and Co., San Francisco, CA.

[Maslow1943] Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, 50:370–396.

[Mavridis2004a] Mavridis, N. (2004a). Inverse kinematics for ripley. Progress Report at http://web.media.mit.edu/ nmav/thesis.htm.

[Mavridis2004b] Mavridis, N. (2004b). Voxeliser/shape recognition/fwd kinematics. Progress Report at http://web.media.mit.edu/ nmav/thesis.htm.

[Mavridis2005a] Mavridis, N. (2005a). Constraints for cross-species intention recognition and feeling / desire communication. Generals Paper at http://web.media.mit.edu/ nmav/thesis.htm.

[Mavridis2005b] Mavridis, N. (2005b). An initial exploration of the problem of communication among substantially different beings. Draft.

[Mavridis2005c] Mavridis, N. (2005c). A near-minimal grounded situation model design for a conversational robot that can pass the token test. Generals Paper at http://web.media.mit.edu/ nmav/thesis.htm.

[Mavridis2005d] Mavridis, N. (2005d). Pathways towards studying human action and the proposal of the human activity language. Generals Paper at http://web.media.mit.edu/ nmav/thesis.htm.

[Mavridis2006a] Mavridis, N. (2006a). Grounding ontologies. Draft at http://web.media.mit.edu/ nmav/thesis.htm.

[Mavridis2006b] Mavridis, N. (2006b). A video demo of voxelizer: voxelized shape models through multiple views for ripley the robot. Video at http://web.media.mit.edu/ nmav/thesis.htm.

[Mavridis and Hsiao] Mavridis, N. and Hsiao, K. Several videos of ripley the robot. at http://web.media.mit.edu/ nmav/thesis.htm.

[Mavridis and Roy2004] Mavridis, N. and Roy, D. (2004). To ask or to sense ? planning to integrate speech and sensorimotor acts. Draft at http://web.media.mit.edu/ nmav/thesis.htm.

[Mavridis and Roy2006a] Mavridis, N. and Roy, D. (2006a). Grounded situation models for robots: Where words and percepts meet. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

[Mavridis and Roy2006b] Mavridis, N. and Roy, D. (2006b). Video accompanying grounded situation models for robots: Where words and percepts meet. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

[McGuire et al.2002] McGuire, P., Fritsch, J., Steil, J. J., Rothling, F., Fink, G. A., Wachsmuth, S., Sagerer, G., and Ritter, H. (2002). Multi-modal human-machine communication for instructing robot grasping tasks. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

[Minsky1988] Minsky, M. (1988). *Society of Mind*. Simon and Schuster, New York.

[Minsky2006] Minsky, M. (2006). *The Emotion Machine*. Simon and Schuster, New York.

[Naryanan1997] Naryanan, S. (1997). *KARMA: Knowledge-based active representations for metaphor and aspect*. PhD thesis, University of California, Berkeley.

[Nicod1970] Nicod, J. (1970). *Geometry and Induction*. University of California Press, Berkeley and Los Angeles, CA.

[Ogden and Richards1989] Ogden, C. K. and Richards, I. A. (1989). *The meaning of meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*. Harcourt, Cambridge, MA, reissue edition.

[Oney2004] Oney, S. (2004). Quantifying gsm performance through the flock of birds sensors. UROP Project Report at http://web.media.mit.edu/ nmav/thesis.htm.

[OpenCV] OpenCV. Open source computer vision library. at http://www.intel.com/research/mrl/research/opencv.

[Picard1997] Picard, R. (1997). *Affective Computing.* The MIT Press, Cambridge, MA.

[Pito1999] Pito, R. (1999). A solution to the next best view problem for automated surface acquisition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21.

[Plato1929] Plato (1929). *Timaeus*, volume 234 of *Loeb Classical Library*. Harvard University Press, Cambridge, MA.

[Premack and Woodruff1978] Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain sciences*, 4:515–526.

[Putnam1973] Putnam, H. (1973). Meaning and reference. *The Journal of Philosophy*, 70(19):699–711.

[PVM] PVM. Parallel virtual machines library. at http://www.netlib.org/pvm3/.

[Rao2002] Rao, S. (2002). *Visual Routines and Attention.* PhD thesis, Massachusetts Institute of Technology.

[Regier and Carlson2001] Regier, T. and Carlson, L. (2001). Grounding spatial language in perception. *Journal of Experimental Psychology*, 130(2):273–298.

[Richards1990] Richards, W. (1990). Mental models workshop: abstracts.

[Richards2005] Richards, W. (2005). Anigrafs. Book in progress at http://people.csail.mit.edu/whit/.

[Roy1999] Roy, D. (1999). *Learning words from sights and sounds: A computational model.* PhD thesis, Massachusetts Institute of Technology.

[Roy2005] Roy, D. (2005). Grounding words in perception and action: Computational insights. *Trends in Cognitive Sciences*, 9(8):389–396.

[Roy et al.2003] Roy, D., Hsiao, K., and Mavridis, N. (2003). Conversational robots: Building blocks for grounding word meaning. In *Proceedings of the NAACL Workshop on Word Meaning*.

[Roy et al.2004] Roy, D., Hsiao, K., and Mavridis, N. (2004). Mental imagery for a conversational robot. *Systems, Man and Cybernetics, Part B, IEEE Transactions on*, 34(3):1374–1383.

[Russell and Norvig2003] Russell, S. J. and Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*. Pearson Education.

[Searle1969] Searle, J. (1969). *Speech Acts*. Cambridge University Press, Cambridge, England.

[Searle1980] Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3:417–424.

[Simons and Rensink2005] Simons, D. and Rensink, R. (2005). Change blindness: Past, present, and future. *Trends in Cognitive Sciences*, 9:16–20.

[Siskind2003] Siskind, J. M. (2003). Reconstructing force-dynamic models from video sequences. *Artif. Intell.*, 151(1-2):91–154.

[Smith1996] Smith, B. C. (1996). *On the origin of objects*. MIT Press.

[Snow1972] Snow, C. (1972). Mothers speech to children learning language. *Child Development*, 43:549–565.

[Sofge et al.2003] Sofge, D., Perzanowski, D., Bugajska, M., Adams, W., and Schultz, A. (2003). An agent-driven human-centric interface for autonomous mobile robots. In *Proceedings of the 7th World Multi-Conference on Systemics, Cybernetics and Informatics SCI'03*.

[Sutton and Barto1998] Sutton, R. and Barto, A. (1998). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.

[Turing1950] Turing, A. M. (1950). Computing machinery and intelligence. *the Journal of the Mind Association*, 59(236):433–460.

[Ullman1996] Ullman, S. (1996). *High Level Vision*, section 9, pages 263–313. MIT Press, Cambridge, MA. This is a full INBOOK entry.

[Zwaan2006] Zwaan, R. (2006). Situation model theory overview. Available online at http://www.nd.edu/ memory/theory.html.

[Zwaan and Randvansky1998] Zwaan, R. A. and Randvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2):162–185.