Situated Concepts and Pre-Linguistic Symbol Use

Dissertation zur Erlangung des Grades
"Doktor der Kognitionswissenschaft"
(PhD in Cognitive Science)
Im Fachbereich Humanwissenschaften
der Universität Osnabrück

Vorgelegt von
Ulaş Türkmen
Schlagvorderstrasse 3
49074 Osnabrück
Geb. 15.02.1980 in Diyarbakır, Türkei

Osnabrück, August 2009

Abstract

In the recent decades, alternative notions regarding the role of symbols in intelligence in natural and artificial systems have attracted significant interest. The main difference of the so-called situated and embodied approaches to cognitive science from the traditional cognitivist position is that symbolic representations are viewed as resources, similar to maps used for navigation or plans for activity, instead of as transparent stand-ins in internal world models. Thus, all symbolic resources have to be interpreted and re-contextualized for use in concrete situations. In this view, one of the primary sources of such symbolic resources is language. Cognitivism views language as a vessel carrying information originally located in the processing mechanisms of the individual agents. Situated approaches, on the other hand, view language both as a communicative mechanism and as a means for the individual agents to enhance and extend their cognitive machinery, by e.g. better utilizing their attentional resources, or modifying their perceptual-motor means. Taking inspiration from these ideas, and building on multi-agent models developed in other fields, the field of language evolution developed models of the emergence of shared resources for communication in a community of agents. In these models, agents with various means of categorization and learning engage in communicative interactions with each other, using shared signs to refer either to pre-given meanings or entities in a situation. In order to avoid falling into the same mentalist pitfalls as cognitivism in the design of these models, such as the stipulation of an inner sphere of meanings for which communicative signs are mere labels, the role of communication should be viewed as one of the social coordination of behavior using physically grounded symbols. To this end, an experimental setup for language games, and a robotic model for agents which engage in such games are presented. The setup allows the agents to utilize shared symbols in the completion of a simple task, with one agent instructing another on which action to undertake. The symbols used by agents in the language games are grounded in the embodied choices presented to them by their environment, and the agents can further use the symbols created in these games for enhancing their own behavioral means. The learning mechanism of the agents is similarity-based, and uses low-level sensory data to avoid the building in of features. Experiments have shown that the establishment of a common vocabulary of labels depends on how well the instructors are trained on the task and the availability of feedback mechanisms for the exchanged labels.

Acknowledgements

First and foremost I would like to thank my advisors Kai-Uwe Kühnberger and Helmar Gust, who have provided an ideal environment at the Institute for Cognitive Science, and valuable comments and advice. I am deeply indebted to Radomir Zugic for his invaluable help in the software department, and his diligent questions on our common subjects; without his contribution, the experimental parts of this work would not have been possible. My colleagues at the Graduate School, Konstantin Todorov, Markus Eronen, Miriam Kyselo and Ilaria Serafini were both valuable friends and great companions in our common quest to a PhD. I would like to thank Carla Umbach and Peter Bosch for the lively discussions in the colloquium. Daniel Weiler, Stefan Timmer helped me with useful discussions on how to get things running, and which directions not to take. Frank Schumacher was always there when I needed serious criticism, for this I am thankful to him. Mario Negrello and Stan James read drafts, and made important corrections. Bernadette Clarke, Hendrik Kettler, Jens Kasper, Martin Schmidt and Torsten Buss made Osnabrück an enjoyable place with their presence. My family back in Turkey, Yusuf, Münevver and Onur Türkmen, gave me incredible support and encouragement, without which I wouldn't have been able to make it till the end. I would also like to thank my housemates Inessa Granowsky, Merle-Marie Schäffer and Ela Warnecke for their friendship and incredible understanding of my neglect of domestic duties. Finally, I would like to thank Josh Homme for the great music that accompanied me through the years working on this dissertation.

yürekte, kitapta ve sokakta yenebilmek yalanı,
anlamak, sevgilim, o, bir müthiş bahtiyarlık,
anlamak gideni ve gelmekte olanı.

Annem ve babama

Contents

Chapter 1

Introduction

A major topic in cognitive science is to what extent intelligent behavior stems from the processing of symbolic representations. How such representations are created, stored, processed and communicated has been a central subject in many fields and research programs; arguably, it has been the center of research in artificial intelligence (AI). The dominant approach in cognitive science regarding this topic, at least until the end of the last decade, has been the representationalist one, which relied on the use of parallels between folk psychological terms and language used to talk about behavior to attribute belief states to intelligent beings, and then build artificial agents which use such states in order to create models of a domain, reason in this domain, and carry out actions according to their goals and intentions. The claim that the disembodied and detached processing of such states amounts to genuine intelligence has been made possible by the availability of technical means to realize mechanising discourse and the theoretical tools that were inherited from fields allowing such discourse. One of these tools was the theoretical background for representational mechanisms, which claimed a primary role for mental representations in accessing the world in which intelligent agents live in. This background served as a viable tool for creating an inner arena of cognition in which such models could be created and mental phenomena could be studied independently of the real world.

The work presented here follows an alternative paradigm which relies on a fundamentally different understanding of meaningful interaction, and derives its sources from research in robotics, artificial life and various philosophical approaches to cognition. This alternative paradigm is known as situated-embodied cognitive science, and it aims to understand intelligent beings as socially and physically situated agents tightly coupled to their environment. In this view, symbol use is associated primarily with communicative phenomena, and methodological parsimony is seen as an important principle to avoid unnecessary intellectualizing of behavioral capabilities. When these views are combined, it can be concluded that the interaction of agents in communicative tasks should become the main topic of study if the dynamics of symbol use has to be studied. This conclusion is also bolstered by research in situated

cognition, in which the interdependence of conceptualization, perception and communication are studied in various social contexts.

The approach presented here is also a synthetic one. A computational model of pre-linguistic symbolic reference will be presented, and the relevant issues will be discussed on this model. Synthetic modelling is one of the distinctive aspects of cognitive science, in that computational implementations are regarded as a measure of realism and a significant means for progress, because the work put into building concrete applications leads to new ideas, predictions and questions. The model presented derives its inspiration from psychological work on categorization, and recent debates in similarity-based methods in psychology.

In this introductory chapter, the issues which will be discussed in more depth will be introduced. The role of representations, and their explanatory use will be discussed, coupled with an overview of theories of the role of symbolic representations and language in cognition. In addition, some of the driving concerns and methodological commitments will be presented. An overview of the following chapters will be given for orientation.

1.1 Cognitive Science and Meaning

There is a clear consensus on the development of the cognitive revolution in psychology and related sciences. The cognitive revolution involved the change of status of theoretical mechanisms which were postulated as a part of theories about the psychological processing carried out in animals and humans. Such mechanisms (or at least singular constructs) were already proposed by behaviorists, but what differentiated cognitive psychology was that the postulated internal mechanisms had "surplus meaning" (Greenwood, 1999). Surplus meaning refers to the ability of theories to generate substantive explanations and lead to further development, due to their metaphoric connections; an example is the planetary model of the atom by Niels Bohr. In the case of the cognitive revolution, the computer metaphor allowed the construction of theories which could claim the existence of internal processing mechanisms, as long as these mechanisms could be realized on computing machinery. According to Fodor (1994), such a cognitive theory can be defined as any theory that postulates representational states that are semantically evaluable and rules, heuristics or schemata governing the operation of such representational states. A quasi-linguistic representational system and a simple theory of correspondence between representations and their references was the most prominent feature of early cognitive science. In this paradigm, representations were processed according to their form in a truth-preserving way that would mirror their content.

As these observations make clear, cognitive science was based on the use of formal languages to describe systems, and the embodiment of the resulting formal systems on computing machinery. The idea that thinking is computation, or more generally the processing of a calculus of mind, is rather old. Hobbes, for example, claimed that "reasoning is but reckoning" (Haugeland, 1981b). Leibniz, who found natural language too vague, wanted to combine

a "Lingua Universalis" with a "Calculus Ratiocinator", which would together reduce reasoning to arithmetic calculation (Davis, 2000). Cognitive science can be seen as the heir to this school of rationalist thinking, taking its modern form in the use of computers as experimental media as well as discursive resources for generating and presenting new ideas. Modern computers present the ability to automatically process formal structures; they are *interpreted automatic formal systems* (Haugeland, 1985, p.47). They can carry out exhaustively specified operations, i.e. algorithms, on information given in terms of formal representations. When these algorithms are specified so that they preserve truth in a certain domain, that is, the specification of the symbols fed to the algorithm concurs with the interpretation of the output symbols, the system can be used for a certain task, as e.g. a chess playing program, or a logic reasoner.

The crux of the discussion surrounding whether this kind of processing, the fundamental mode of traditional AI systems, is sufficient for an exhaustive theory of human intelligence, and whether scientific theories relying on the abstraction of algorithmic processing of symbolic structures are sufficient for understanding the human mind, is the status of these symbols as meaningful units. More precisely, the question that has to be answered is whether the meanings of the representations which comprise the structures processed by computers are parasitic on the interpretations of the humans that design and interpret them, or whether they are, due to the role they serve in the whole system, genuinely meaningful, thereby granting the programs which process them the status of intelligent systems. The traditional approach in cognitive science has been to opt for the second answer, which states that, despite the simplicity of our engineered systems, these symbols are meaningful to the extent that the symbol processing systems approach realistic human behavior.

The duty of the cognitive scientist is thus to analyze human behavior, locate systematicities, decode these into algorithmic procedures, and build systems which embody these procedures, at the same time producing testable hypothesis and further phenomena to be studied. The data generated, and what has been discovered in the design of the algorithms, are a coupling possible only through the combination of analytic and synthetic methods offered by cognitive science. The result of this approach has been models which duplicate the world in an inner arena in order to create formal representations on which the reasoning mechanisms can function. This gesture of internalization, of positing maps and models in the head, led to cognitive science having a certain character, and determined what was treated as a subject worthy of studying, and how the fundamental human capacities were formulated as research questions. Cognitive science, and especially AI concentrated on those tasks which have an abstract character, and whose inputs and outputs can be characterized in terms of formal-representational structures; early examples for such tasks are game playing, text comprehension and automatic translation. Domains which did not fit this mould, such as perceptual and motor tasks, were stated in terms that would extend this framework, delegating to perception the role of creating a description of the environment in symbolic terms, and to motor mechanisms the role of the execution of centrally-created plans. The traditional position which is a result

of this constitutive role of representational mechanisms and symbol processing is called cognitivism.

Efforts to integrate research into various modules and capacities in the cognitivist paradigm into complete agents, especially ones with robotic capabilities, showed that the sum of many specialized modules does not necessarily equal complete intelligence. Decades of research both into the specialized problems of AI and more general issues have demonstrated that the dream of full-fledged AI is more distant than initially presumed. A number of problems have cropped up repeatedly, resisting all hand waving and simplistic solutions. Among these problems are the common sense problem, which concerns the representation of the simplest pieces of knowledge needed for basic reasoning (e.g. "If a man takes a shower, his foot gets wet") and the relevance problem, which concerns how the discovery of which pieces of information are relevant in coping with a certain situation gets computationally intractable as the knowledge base of an agent grows. These problems have led some to speculate that they might be symptoms of a deeper issue, namely the reliance on internal representations, and more generally, the postulation of an inner space of reasoning and representation. What has to be done to advance AI research is therefore to go back to fundamentals, both at the level of perception and action, and at the level of symbolic capabilities for mental reasoning, and look at how people, animals and artificial agents use their resources in concrete situations, reinterpreting what they have previously learned and categorizing situations in particular ways. This approach, called situated cognition, rejects the notion that the symbols used in traditional AI programs are intrinsically meaningful just by fiat of their having been designed for a certain task and conforming to the initial interpretation, and argues that the problems faced by AI are repercussions of the strict reliance of cognitive science and AI on symbolic representations, without sufficient concern about how such representations are created and their meaning maintained by the community which uses them.

It is nevertheless obvious that human beings make use of symbolic representations, and many complex capabilities are a result of the ability to use such representations to coordinate behavior individually and socially, that is, conforming to the standards of a community of agents. The evolutionary and phylogenetic roots of this capacity, as well as the computational capabilities necessary for it, are subjects of intensive research in many different fields. One of these fields is the synthetic modelling of multi-agent communities which engage in communicative behavior, creating a self-organizing dynamics through which a shared lexicon arises. The theoretical background for this work derives from the situated cognition literature, where it is claimed that the primary source of symbols in intelligence is public symbol use for communicative purposes, which culminates in language in humans. In order to understand how a socially constructed vocabulary of symbols can be maintained through communicative inter-agent interaction, symbol use can arise in a society, and thereby endow the individual agents with symbolic capacities for activities other than communication, multi-agent models are constructed, and the emerging dynamics is studied in the light of changing agent-level capabilities.

However, certain tendencies from the mentalist paradigm are still present in the work on dynamics of communication. An example for these tendencies is the locating of original meaning in an entity in the internal processing mechanism, and communication serving the purpose of creating correspondences between labels and these internal meanings. As it will be argued in the next chapter, locating an original meaning in internal representations in the head is a central tenet of the representationalist approach, and leads to many technical difficulties in AI, and philosophical problems in cognitive science in general. The work presented here aims to build on work on the dynamics of communication, at the same time integrating alternative philosophical theories of meaning and symbol use. These theories view the fundamental significance of symbols in their use for embodied action in shared situations, serving agents as resources to organize their behavior in a socially coordinated way. More precisely, the aim of this thesis is to situate the meaning of symbolic communication in concrete task contexts.

1.2 Language and Cognition

Humans are the only living beings which can engage in complex and protracted cognitive activities such as problem solving, planning, intentional deliberation and social exchange of highly structured information, while at the same time dynamically modifying their behavior to suit radically changing situations. They are also the only beings that can use context-free, productive and multi-modal means of communication. This co-occurrence is the most simple reason to think that there is a connection between human intelligence and the linguistic capacity. The extent of this connection has been the subject of much debate, and various positions have established themselves as the basis of different philosophical traditions and scientific research programs. In cognitive scientific terms, this problem takes the form of the question "What is the role of language in cognition?".

Language is one of the areas which lies at the foundation of cognitive science. The cognitivist research program drew its most fundamental tenets from the work of Chomsky and fellow linguists, who claimed an innate language faculty, a language "organ" responsible for the acquisition and processing of language (Chomsky, 1988). This claim for the division of mental capacities into various units and furthermore the assertion of a knowledge of language, instantiated as rules in the human cognitive apparatus, became a fundament for the rest of cognitive science. The role of language inherent in this paradigm is one of language as primarily a communicative tool, serving as a vessel for the structures human beings have in their minds in 'mentalese' and want to transmit to others (cf. Carruthers, 2008). The fundamental units of meaning are therefore the representations which words are overt forms of. In AI, this approach has taken the form of language understanding as encoding and decoding of linguistic signs to internal structures. This position is called the communicative conception of language, since it describes language as a vessel for meaning originally located

in the head (Carruthers and Boucher, 1998).

In the psychology literature, there is considerable proof pointing to the cognitive use of language. It has been shown, for example, that young children working on a task made more self-directed verbalization when the task was more complicated (Berk and Garvin, 1984). Furthermore, the children who made more verbalizations were more successful. An interesting experimental study of the use of linguistic cues involves navigation in a maze. It has been shown that rats use only geometric information when they have to reorient in a closed area, ignoring other cues such as colors or scent (Cheng, 1986). Reorientation refers here to the rats being put in an area, allowed to discover it, and then taken away, spun around and placed back in the area, so that they have to find their way again. In experiments with children and adults using a similar setup, it was found out that children between 18 and 24 months also relied solely on geometric cues. When a toy was hidden behind a corner of a rectangular area, children looked at all corners of a room with equal probability searching for a toy, although one of the walls had a stark color, serving as a cue for the corner to look for. As can be predicted, adults used this cue for reorientation (Hermer and Spelke, 1996). This, on itself, does not prove anything about language use in spatial orientation. However, two experimental studies suggest that linguistic capabilities are crucial in such a reorientation task. The first comes from Hermer-Vazquez et al. (1999), who have shown that when adults, who normally use non-geometric information when they have to reorient, also default to solely using geometric cues when they are engaged in verbal shadowing of continuous speech. Furthermore, this effect was removed when this shadowing was a continuous rhythm, "suggesting that the interference effect [...] did not stem from general limits on working memory or attention but from processes more specific to language" (Hermer-Vazquez et al., 1999, p.3). Also, Shusterman and Spelke (2005) have shown that simply training children to use the words "left" and "right" correctly has a marked impact on their success in the disorientation task. These experimental results point to a significant role of the individual's use of her linguistic capacities outside of a communicative context in spatial orientation. The position on the cognitive status of language and symbolic capacities which derives from such results, and claims a more central role for these capacities, is called the cognitive conception of language (Carruthers and Boucher, 1998).

Another interesting study which demonstrates the role of symbolic capacities in intelligent behavior was carried out by Thompson et al. (1997). In earlier studies, chimpanzees with and without language training were tested in what is called a conceptual matching-to-sample task (Premack, 1986). In this task, the subjects are asked to match the relations between the objects in two sets. When they are presented the sample AA (e.g. a pair of physically identical cups), the subjects have to pick EE (e.g. identical pair of shoes) as analogous to the sample. Since the relationships of two pairs of objects have to be matched to each other, the matching of physical relationships is not enough for successful performance; what is necessary is the matching of two abstract relations. Premack (1986) had shown that chimpanzees with language training performed

better than those without any language training. The research question posed by Thompson et al. (1997) was what aspects of language training were responsible for the improvement. They found out that the best-performing chimpanzees among their subjects were the ones that were trained in labeling of relations and referents with tokens; they had been trained to match numeric displays of objects with Arabic numerals (e.g. 3 for XXX and XXX for 3) and labeling pairs of objects as being either the same or different ("If identity, choose heart" or "If non-identity, choose diagonal"). The authors conclude that "experience with symbolic tokens per se produces a system for universal computation in the chimpanzee, as it also does for the human child, if not other species" (p. 42). Another interesting effect was observed in an experiment which involved two chimpanzees in a task with a reversed reinforcement contingency. The chimpanzees were presented with two bowls of candy. One of the chimpanzees had to make a choice by pointing at one of the bowls, after which he received the other bowl, whereas the bowl it pointed at was given to the other chimpanzee. Even after hundreds of training trials, the chimpanzees were unable to point at the bowl with the lower number of candies; they were not able to inhibit their direct response to the bowl with more candies (Boysen et al., 1996). In a modification to the task, the bowls were replaced with Arabic numerals showing the number of candies in the bowls. When one of the chimpanzees which had earlier been trained with using Arabic numerals as quantity symbols was tested in this modified task, it was immediately able to invoke the optimal food-sharing rules: "As long as Arabic numerals served as stimuli, Sheba consistently selected the smaller numeral and earned the greater number of candies. When candies were reintroduced as stimuli, her performance immediately deteriorated but returned to more optimal levels when Arabic numerals were again substituted for the candy array stimuli" (Boysen et al., 1996, p.77). The results of this experiment show that external symbols can be useful in inhibiting direct response to attractive stimuli, in order to receive delayed but higher reward.

These and similar findings point to a more central role for language in intelligent behavior. However, proposing a theory of the cognitive role of language relying on the classical internalist picture, where the mind already has a language of its own, is futile at best: any functional role attributed to a socially acquired language will be a derivative of this internal language. Due to this reason, recent theories on the cognitive function of language also envision an alternative role for public language. An example is Clark (1998), who counts six different functions language and symbols serve, other than communication.[1] Among these roles are memory augmentation, environmental simplification and attention allocation. Clark (1998) calls his approach the supra-communicative view of language; this theory will be studied more in depth in Section 3.3.

Proposals like that of Clark (1998) show that it is possible to envision an alternative role for language, making it possible to avoid the internalist tendencies of the traditional paradigm, and at the same time laying the groundwork

[1] Another interesting example is Carruthers (1998), who gives language the role of unifying the outputs of different modules in the mind.

for a cognitive scientific theory of symbol use grounded in the social practices in a community. The central aim of the work presented here is to build on this and similar proposals and present a synthetic perspective into the role of primitive communicative capacities in the social organization of embodied behavior in communities of agents. The interaction of two dynamics will be studied in a multi-agent model, with the aim of understanding the conditions necessary for the emergence of a common vocabulary in a community of agents. These dynamics are the bodily dynamics of the individual agents, and the multi-agent dynamics of symbol use and referential behavior. The fundamental question to be addressed is how a community of agents can devise ways to refer to choices they have in the environment for embodied activity, without building in internal representations. Relying on situated approaches to the role of symbols and language, the meaning of symbols in the absence of internal representations will be located in the behavioral choices the agents have acquired in their previous interaction with the environment. The situation in which the agents engage in language games will present distinctions in which the symbols used in communication can be grounded. Once such symbols are acquired by the agents, their behavioral capabilities can be augmented for more complicated interaction, which further transforms the situation as it is available to the agents.

It must be pointed out that the linguistic capacity of humans exhibits various kinds of complexity at different levels, each of which is still a domain which requires its own forms of understanding. Neither the work presented here, nor other contributions in the field of language evolution and dynamics of communication, yet have any radically new ideas to offer when it comes to understanding the whole complexity of language, with its complex syntactic structures and inter-language variability coupled with universal constraints. Nevertheless, it should be possible to look at linguistic communication not solely as transformed mentalese, but as a tool that enhances other cognitive abilities. It is further a good idea to start such an alternative understanding at the most basic level of communication, at the level of single symbols. Therefore, even when there is talk of language in the rest of this work, one should think of very simple settings of primitive symbol use, comparable to the utilization of gestures, albeit with a recognizable form, so that signs can be reused and learned.

1.3 Overview

In the next chapter, an overview of the main criticisms arrayed against cognitivism will be presented. The main focus will be on AI, and the technical problems faced by AI work. It will be argued that these technical problems were symptoms of more deep-seated philosophical issues. The third chapter will focus on the embodied and situated approaches to cognitive science. These approaches, after having appeared as a reaction to cognitivism, have established themselves as independent research paradigms, and offer a range of alternative tools, both theoretical and computational, to study human intelligence. Afterwards, in Chapter 4, recent work in the emergence of communication and the

dynamics of symbolic communication will be given. This is a preliminary to the main contribution discussed in this dissertation, a model of symbolic communication and situated symbol use, presented in Chapter 5. In Chapter 6, a discussion of the advantages and shortcomings of the model will be presented. In the last chapter, a short overview will be given and conclusions will be drawn.

Chapter 2

Representationalism: Cognitive and Philosophical Issues

The situated approaches to cognition have established themselves as a paradigm for cognitive science. They have developed tools and a research vocabulary for all the relevant domains of cognition and scientific research on cognition. In the early years of the search for an alternative to the representationalist method, it was nearly an imperative to start a study in the situated-embodied tradition with a criticism of cognitivism. This is to be expected of a new scientific paradigm, because every new effort and alternative paradigm starts as a reaction and as an anti-stance (Kuhn, 1996). Cognitive science itself is a good example, as it defined itself as an alternative to behaviorism, as a revolutionary option.

It can nevertheless be claimed that cognitivism is a special case as far as paradigms in the human sciences go. This specialty derives from two factors. The first is the role of the received philosophical-theoretical background in the formulation and popularity of cognitivism, in that cognitivism could rely on a rich background of philosophy which had already worked out the fundamentals of computational thinking. The second is the historical and social context in which especially artificial intelligence established itself as not only a scientific method, but as a design for the future. Due to these reasons, and as a necessary way of highlighting the starting points and initial driving forces of research on situated cognition, it is a good idea to remember the most important criticisms directed at cognitivism. In addition, the philosophical and historical background delivers a number of insights, not the least about the structure of the scientific enterprise as a situated and historically shaped human endeavor.

In this chapter, first, a review of the scientific shortcomings of cognitivist approaches to cognition will be given. These shortcomings were the primary reason for the search for an alternative theoretical and experimental framework.

Afterwards, the philosophical background which led to the acceptance of the cognitivist approach will be discussed. Situated and embodied approaches to cognition have derived their philosophical inspiration from a number of sources; the two most important of these, Wittgenstein and the phenomenological tradition, will be shortly discussed. The main features of and recent developments in the situated-embodied program will be outlined in the next chapter.

2.1 Cognitivism in AI and Psychology

Cognitive science is built on the idea of operationalizing the human being, i.e. making it amenable to scientific study, as an information processing system. The forms of scientific explanation which information processing accounts enable are different from those of the other sciences, which discover rule-based regularities. According to Haugeland (1978), cognitive scientific explanation brings together two forms of scientific explanation. The first of these is morphological explanation, where "an ability is explained through appeal to a specified structure and to specified abilities of whatever is so structured" (Haugeland, 1978). An example is the organization of the DNA, where replication is a capability based on the double-helix structure. The second kind of explanation is systematic explanation, where, in addition to a morphological explanation, "organized cooperative interaction" plays an important role. In such an explanation, a system is broken down to its functional components, and their interactions make clear "how it works". The description of the inner workings of e.g. a radio would be an example of such an explanation, but the system does not have to be an engineered one; many biological explanations also rely on a similar idea of systematicity.

In order to formulate an information processing account of a system, it has to be described in terms of an *intentional black box* (IBB): the outputs of the system have to make sense in terms of the inputs and a domain in which the system is assumed to function (Haugeland, 1978). When we talk about a system as, for example, playing chess, we use quasilinguistic descriptions in order to describe what it is doing: the chess playing program gets out its queen, plays aggresively, wants to capture a pawn etc. What now has to be done is to pick the outermost IBB apart, to arrive at functional components which in turn themselves are IBBs. The components which are either stipulated (if an existing system is being analyzed) or created (if a new system is being engineered) in this step are again described in terms of the quasilinguistic description with which the system itself was described in the first place. These component IBBs can themselves be further deconstructed. This process of further decomposition grounds when an element can be physically instantiated on a substrate, such as a logic gate on silicon or a weighted connection on a synapse.[1] The advantages of this kind of scientific explanation compared either to behaviorism, which rejects ascriptions

[1] Dennett (1978) argues for a very similar perspective when he describes the work of cognitive science as defining simpler and simpler homunculi, until these can be instantiated in processing machinery.

of meaningful internal states and intentions, or to introspection are obvious. Cognitive science can now study mental phenomena and thereby speculate on intervening processes and states between input and output. It can also do this without fear of getting lost in speculation on private inner impressions, because any speculation or stipulated process has to be "modelled", i.e. instantiated as an algorithm, or shown to be implementable as such.

As it will be argued later, the history of cognitive science (and AI in particular) is one of the relaxation of the constraints on the forms these IBBs can take, and the status of the quasilinguistic descriptions. The earliest computers, which could process much less information compared to the modern ones, were constrained in the kinds of computations they could carry out in a viable period of time to rule-based symbolic computations. This limitation, which hindered practical work on alternative approaches to computation, and the preoccupation with certain mental tasks as the paragon of human intelligence (such as games, automatic translation, solving physics problems etc.) led to a concentration on certain kinds of computational explanation and the study of these mental tasks in psychology. This early approach to cognitive science has been called cognitivism.

The cognitivist program found its most prominent articulation in the two fields which were among the primary contributors of cognitive science: AI and cognitive psychology. After the second world war, these two fields cooperated closely, establishing AI as a viable method to test theories, and cognitive psychology as a field that studied humans to search for proof of the computational mechanisms developed by AI research.

2.1.1 AI: The paradigm statement of cognitivism

The best-known and most-quoted statement of the cognitivist research program in AI is that by Newell and Simon (1976), which frames the study of intelligence in terms of a physical symbol system (PSS). A PSS has three constituents:

1. A set of entities, called symbols, which are physical patterns that can occur as components of another type of entity called an expression (or symbol structure).

2. A set of processes that operate on these symbol structures and modify, create, destroy, or reproduce them.

3. A world in which these symbols and processes are embedded and with which the system has the relationships of *designation* (dependency of the behavior of the system on an object) and *interpretation* (expressions in the system designate a process, and the system can carry out this process when the expression occurs).

Two things have to be noted about this definition, which is the result of long years of work in AI. First of all, it is a sufficiency claim: a system designed on these principles has the capacity to act intelligently. Second, this is an

operational claim, in that it is a principle for construction, aimed at directing research in a certain direction and defining a paradigm. Touting the importance of the central processing mechanisms, it relegates the property of intelligence to these central mechanisms, making them the focus of cognitive research. A more simple form the PSS hypothesis can be observed in scientific practice has been rather succinctly formulated by Clark (1989):

> *The strong-physical-symbol-system (SPSS) hypothesis*: A virtual machine engaging in the von Neumann-style manipulation of standard symbolic atoms has the direct and necessary and sufficient means for generating intelligent action. (p.12)

In the kind of computation envisioned by cognitivism, the "standard symbolic atoms" are manipulated solely due to their syntactic properties; as Haugeland (1981b) points out, "given an interpreted formal system with true axioms and truth-preserving rules, if you take care of the syntax, the semantics will take care of itself" (p. 23). As it was mentioned above, cognitive science is based on the idea of treating intelligent systems as intentional black boxes, which means that the syntactically structured inputs and outputs to the system are stated in terms of the domain in which the system is assumed to be/supposed to be intelligent, such as pieces of a chess game and chess moves, objects in a blocksworld, words in a text to be translated etc. Since the symbolic atoms have to share the structure of the inputs and outputs in order to function in a system processing them, their meanings are also derivative of the intentional interpretation which accorded the system the status of intelligence and understanding in the first place. That is to say, the atoms in the inner space of computation derive their meaning from the interpretation of the input and output tokens. Such a system which makes it possible to find "a neat mapping between a symbolic (conceptual level) semantic description of the system's behavior and some projectible semantic interpretation of the internally represented objects of its formal computational activity" is called "semantically transparent" by Clark(1989, p.18). It should be noted that the domination of semantically transparent systems in AI research appeared as a natural situation to the researchers, as McDermott (1981) observes:

> Most AI researchers react with amusement to proposals to explain vision in terms of stored images, reducing the physical eye to the mind's eye. But many of the same people notice themselves talking to themselves in English, and conclude that English is very close to the language of thought. (p.153)

As Anderson and Perlis (2002) point out, the definition of a symbol in Newell and Simon (1976), and the future incarnations of the same idea, are in fact too general to establish an empirical theory. A pattern is called a symbol when it can designate or denote; they may designate other symbols, or patterns of sensory stimuli and motor actions (Vera et al., 1993, p.9). Where the definition of a symbol in PSS is too general, a look at the periphery of the definition,

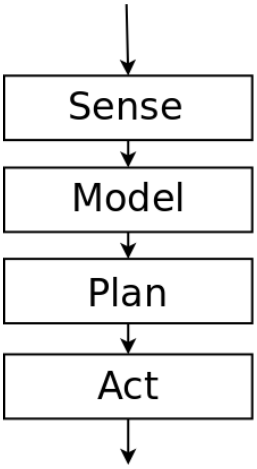

Figure 2.1: The Sense-Model-Plan-Act architecture.

to how the system is supposed to interact with the environment, reveals more: "A physical symbol system interacts with its external environment in two ways: (1) it receives sensory stimuli from the environment that it converts into symbol structures in memory; and (2) it acts upon the environment in ways determined by symbol structures (motor symbols) that it produces" (Vera et al., 1993, p.9). These correspond to the designation and interpretation relationships mentioned in the above definition of a PSS. In these relationships, and the ways symbols are produced and consumed, we see the central idea of cognitivism as it was used as a design principle in AI: the central processing mechanism is fed by the peripheral perception mechanisms, and delivers representations of actions to the peripheral motoric mechanisms. The possibility of such ideal peripheral mechanisms was simply assumed, however, and the correspondence of the internal symbols to the objects in the assumed domain, hence the meanings of the symbols the physical symbol systems processed, was solely due to the interpretations of these systems as functioning in these domains. Brooks (1991) calls the division of intelligent systems into perceptual, motor and central units the SMPA architecture for "Sense-Model-Plan-Act" (See Figure 2.1.1) and claims that the different submodules were also delegated to different research groups and laboratories, which frequently did not have contact. This led to the difficult aspects of perception and motor control to vanish from the radar of mainstream of AI, despite their central importance in cognition.

Scientific work in AI, especially before the 1980s, consisted therefore mainly of two areas: central algorithms, like planning, problem solving, knowledge representation etc. and the peripheral mechanisms, like vision or motor control. The first of these areas, work on central processing mechanisms in the symbol processing paradigm, is called Good Old Fashioned AI (GOFAI) by Haugeland (1985). In work in the framework of GOFAI, the interface between the central

mechanisms and peripheral modules was assumed to be well-defined, and it was taken for granted that the symbols that were processed in the central algorithms had the meanings attributed to them by the researchers. Furthermore, computation was limited, in the domains which were seen as pure AI, to rule- and logic-based methods, although peripheral fields like machine vision employed a much wider range of algorithmic methods.

2.1.2 Discontents with Cognitivist AI

After initial successes and hopeful claims, discontents with the results of cognitivist AI started surfacing. A report commissioned by the RAND corporation, which oversaw and funded most of AI research in the U.S., came to the conclusion that the machine translation efforts underway had failed to make progress. Another report commissioned by the British Science Research Council (the Lighthill report) came to similar conclusions on AI work carried out in Britain, and funding was cut down to a fraction of its earlier amount. Automatic translation was one of the possible applications of AI with a lot of promise, both due to the importance, in the post-Sputnik world of Cold War, of rapidly translating documents from Russian into English, and as an ideal showcase of AI methods, since it involved conversion of textual material to more text. As the authors of both reports pointed out, however, this problem turned out to be much more difficult than initially projected.

Dreyfus (1979) presented the outcomes of these reports in a critical and philosophical context in one of the best-known and earliest criticisms of GOFAI. His criticisms can be summarized under a number of general points. The first of these concerns the stipulated symbolic interface between the central processing mechanism and the peripheral modules. The inputs to an AI system, at least in the systems which accept appropriately formatted symbolic input, are already pre-processed so that the program can carry out symbolic operations on them. Dreyfus (1979) points out that in most contexts of human activity, the production of such a formulation (conceptual perception) is already the most complicated and crucial part. The presentation of a problem therefore already contains a considerable part of its solution. A very similar criticism is made by Chalmers et al. (1992) in their discussion of computational research on analogy. They point out that in most models, the conceptualization on which structural comparisons and derivation of analogical structures are to be carried out is already presented to the system in the form of a structured representation, and the construction of this representation is the fundamental performance exhibited by humans, especially when analogies are concerned. That is to say, even if humans were to be assumed to use such structured representations, the significance of these structures, and their suitability for use in different kinds of reasoning tasks, would be a result of the mechanisms through which they are created in the first place.

Another important problem discussed by Dreyfus (1979) is that of the role of context and background, especially in language comprehension. In order to disambiguate certain utterances, the speaker of a language has to make use of

world knowledge he possesses. An example quoted by Dreyfus (1979) is the sentence "The box is in the pen" (p.217). In order to find out what "pen" refers to, one must know that boxes usually do not fit in pens, hence the pen mentioned here is probably not a writing instrument, but the small fenced area in which a child plays. However, if this sentence was uttered by James Bond to another agent, "pen" might as well be a writing instrument, because special agents tend to possess such instruments. Evidently, the context in which such an utterance is made has to be brought to bear on the interpretation of the sentence. The problem arises once this context is assumed to consist of individual facts itself. If the context is also to be recognized as a collection of individual facts, how can one formulate the further context in which these facts in turn are embedded?

There are two ways of avoiding an infinite regress. The first is to postulate an absolute context, a "background", which serves to conceptualize and disambiguate all situations. The construction of such a database of background knowledge has been undertaken a number of times in the history of AI. One of such projects started with the "naive physics manifesto" by Hayes (1979), who proposed formulating a database of everyday knowledge, especially of the physical world. This database would consist of facts encoded in propositional form, and would serve as background to any processes reasoning about the world. After a decade of work on this project, by Hayes and many others, McDermott (1987b) came to the conclusion that such a project is impossible to carry out, and this was more due to principal reasons than resource limitations.[2]

The second way to solve (or rather understand) this problem is to break out of the cognitivist framework and accept the existence of knowledge which does not consist of a collection of facts, expressed in terms of context-free representations, but is formulated as *know-how*, as a way of being in the world. This relies on an interpretation of phenomenological philosophy, and will be discussed at the end of this chapter.

2.1.2.1 Parallel Distributed Processing

The rebirth of the perceptron approach as Parallel Distributed Processing (PDP, also called connectionism), which brought the ideas of distributed representations and parallel computing into the forefront of cognitive modelling, was the first step towards a proliferation of conceptions of computation and representation in AI and cognitive science, and alternative considerations on the meaning of processes internal to the agent. Renewed enthusiasm in neural networks was a result of work on networks with hidden units, which were proved to be able to represent any arbitrary function (Haykin, 1999, p.208-209), and new algorithms for learning with such networks, especially backpropagation learning.[3] Although there are currently many different uses for neural networks, and con-

[2] Another well-known such project is the Cyc project which set out to do something very similar, albeit without the limitations of first-order logic (Lenat and Feigenbaum, 1991). See Smith (1991) for a thorough criticism of this project and its results.

[3] There are many textbooks on neural networks; for a comprehensive text see e.g. Haykin (1999). For a philosophical discussion of the significance of PDP see Clark (1989).

sequently many different network stuctures, initially, neural networks were used for pattern recognition and classification tasks.

The contrast of neural networks to the classical symbolic systems lies, at the operational level, in their capabilities to recognize incomplete patterns and construct partial stimuli, their abilities to generalize to novel situations, and their tunability to changes in the environment (Hinton et al., 1986). Individual patterns and experiences are not represented locally in single nodes, but in the connection weights of the network; neural networks are thus said to implement "distributed representations". This makes neural networks more robust, and leads to graceful degradation in case of "lesions", i.e. removal of parts of the network or individual connections. Due to these properties, neural networks were initially regarded as the solution to some of the most significant problems of classical symbolic systems, for example by Dreyfus and Dreyfus (1988). The advance of neural network approaches also brought their paradigmatic application areas, pattern matching and micro-cognition, to the forefront of cognitive modelling. As Hofstadter (1982) points out, this difference of focus can be understood as one between time frames in which perception and further cognition take place. Hofstadter (1982) quotes Simon, who claims that "Everything of interest in cognition happens above the 100-millisecond level" (p.632), and argues that exactly the opposite is the case; everything of interest in cognition happens below this threshold.[4]

The distributed nature of representations in neural networks caused discussions on the status of the representations these networks process, and the "meanings" of the weights and network activations produced in response to stimuli. It was argued that, because neural networks did not function on compositionality principles, they would be limited to the peripheral modules responsible for perception and motor control, and never be useful for understanding central processes (Fodor and Pylyshyn, 1988). Connectionists responded to this criticism by arguing for the existence of a subconceptual level of processing:[5]

> The subconceptual level hypothesis: Complete, formal and precise descriptions of the intuitive processor are generally tractable not at the conceptual level, but only at the subconceptual level (Smolensky, 1988, p.105-106).

Despite their many advantages and favorable qualities in comparison to symbolic systems, neural networks have their own shortcomings. Backpropagation learning, the most-used and best-suitable learning method for neural networks, is a supervised learning method: the network has to be presented with the "correct" responses to stimuli, in order to calculate error values and re-adjust

[4]This distinction is parallel to the peripheral mechanisms vs. central mechanisms distinction explained earlier. The central mechanisms, which correspond to internal deliberation, are assumed to function at a longer time scale than those of perceptual mechanisms, which are unintentional and comparatively spontaneous.

[5]It must be mentioned that this very short treatment glosses over an extensive literature in the field of compositionality in network architectures; for an overview see e.g. Hammer and Hitzler (2007).

the weights. Applied to living beings, this leads to the assumption that they have to somehow sense the correct response in case of an error, or during their childhood, they pick everything done by their elders as correct. It is possible to model reinforcement learning with neural networks[6], but this requires the inclusion of a model of the process the system is trying to learn in the network (Sutton, 1991). Another problem with neural networks is that their topology (connection between the different levels, the number of hidden units etc.) has to be carefully calibrated by the designer; in order to avoid this step, neural networks are sometimes evolved using genetic algorithms, which means using neural networks as generators of dynamic behavior.

Neural networks have become a part of the standard toolkit of cognitive science and AI. In the process, the idea of subsymbolic and distributed representations have become a part of the conceptual vocabulary. Neural networks are also widely used in situated AI in order to model dynamic controllers; this will be discussed in the next chapter. Here, it must be pointed out that the use of neural networks solely as disembodied and unsituated pattern classifiers makes the PDP approach vulnerable to many of the same criticisms arrayed at the symbolic approach. As Clark (2003) observes, "both [GOFAI and PDP] agree that rational thoughts and actions involve the use of inner resources to represent salient states of affairs, and the use of transformative operations (keyed to non-semantic features of those internal representations) designed to yield further representations [...] and, ultimately, action". As long as neural networks are used solely to map patterns of input to output nodes, both of which acquire their meanings due to the interpretation of the designer, and not due to a sensory-motor loop comprising a society of agents and the environment, they will be prone to the same problems as the traditional approaches.

2.1.2.2 Representing actions in a changing world: The frame problem

An issue which attracted considerable attention is representing actions and their effects in cognitive models of worlds which are temporally changing, known as the frame problem (Pylyshyn, 1987). The frame problem originally appeared as a technical problem in AI research, and was concerned with the representation of the effects of actions on the system-internal variables (McCarthy and Hayes, 1969). The term was later appropriated by philosophers to refer to various different problems and research topics, removed from the original AI perspective, which caused ample frustration among AI researchers (see e.g. Hayes et al., 1996, for a particularly acerbic example). In deductive approaches to AI, temporal actions are represented through logic sentences which state what is the case in different situations, and what changes through the application of a certain action. However, stating what has changed is not enough; the designer of a system which has to reason about such temporal changes also has to take into account what *does not* change. It must be stated that, technically, the frame

[6] See Section 5.1.2 for a discussion of reinforcement learning.

problem on itself is not that difficult; as McDermott (1987a) points out; "no working AI program has ever been bothered at all by the frame problem" (p. 116). The problem is technically solved through the application of alternative logical representations and methods; see Shanahan (2008) for a comprehensive overview.

The deeper philosophical issue is nevertheless still relevant, as Fodor (2000) has pointed out. An agent, operating on a collection of logical representations, has to know the consequences of its actions in order to decide on a sequence of actions for an end. If these consequences are represented in the form of facts in an internal symbolic database, the computation of these consequences can become intractable as the size of the database grows, even if it has a very fast mechanism of checking for the results of an action and whether they apply in the case given at hand. The reason is that the combinations of all the conditions and outcomes leads to a computational explosion. As Fodor (2000) argues, this more general problem is not endemic to GOFAI systems, but haunts any rational system whose functioning is based on the processing of stored facts: information relevant to a certain flow of reasoning can stem from anywhere in the system, and catastrophically for a modular view which relies on the separation of the central from the peripheral modules, also from perceptual and motor modules. The problem of deciding on what is relevant to the current worries of a system and what is not is called the problem of relevance by Anderson (2003), who points out that this problem is closely related to the status of representations in an embodied agent for which perception and action are not abstracted away. What is relevant in a given situation depends on the capacities and the aims of the agent, and consequently the situation will be perceived in that way; that is, something is perceived and represented as relevant in the first place:

> One moral which might be derived from the problem of relevance is that it doesn't make sense to think about representing at all unless one knows what one is representing *for*. The [Sense-Model-Plan-Act] model envisions the infeasible task of *deriving* relevance from an absolute world model; but if we step back a pace we would realize that no system can ever have such an absolute (or, context-free) world model in the first place.

An interesting aspect to the frame problem – or generally the relevance problem – is that a somehow similar problem on the Internet, which documents are relevant to a certain query based on the occurrence of the terms in the query, is solved to a sufficient degree by search engines such as Google. These engines are technological proof of concept that huge amounts of data, as much as there is available on the Internet, can be stored in such a way that the relevant documents can be fetched in a matter of milliseconds for relatively complex queries. This should not occlude the fact that relevance in this case is defined solely by the co-occurrence of similar words and phrases. The complexity of the queries is fundamentally limited to the presence of certain terms and the absence of others, and more complex queries specifying further relationships in different domains, or the intentions of the person searching the database are rarely taken

into account. Furthermore, as technological devices, these programs need huge amounts of fine tuning to e.g. fend off attempts to manipulate the search engine by so-called "spammers" to direct attention to their own website. This need for continuous calibration demonstrates that the relevance they calculate is a matter of technological consensus, and depends on what counts, in the eye of the beholder, as a relevant fact or not, and is not intrinsic to the system.

2.1.2.3 Symbol grounding

The issue of the symbols in AI systems having a meaning only for the people that build and interpret those systems finally got a name when Harnad (1990) christened it "the symbol grounding problem":

> How is symbol meaning to be grounded in something other than just more meaningless symbols? This is the symbol grounding problem. ... The standard reply of the symbolist is that the meaning of the symbols comes from connecting the symbol system to the world "in the right way." But it seems apparent that the problem of connecting up with the world in the right way is virtually coextensive with the problem of cognition itself.

From a wider perspective, there is a problem with calling the problem of how the symbols used by humans are connected with the world the "symbol grounding problem", especially in the formulation given by Harnad (1990). First of all, Harnad (1990) assumes that there are symbols which do not have to be grounded, once there is a subset which serve as definiens for all the other symbols. Second, it is assumed that this grounding has to be done by the individual, and the symbols are then usable by this individual. As it will be argued later, the grounding of a symbol is much more the job of a community, a fact which would cause this issue to be called e.g. "the symbol sharing problem". Third, grounding symbols involves not only individual symbols, but also a structural consistency which makes these units into symbols in the first place; an example is one symbol having a part-whole relationship to another one. If such relationships are not part of the symbolic capacity, it is difficult to talk about symbol grounding. As Gasser (1993) argues, symbol grounding is actually a problem of structure grounding.

What Harnad (1990) calls the symbol grounding problem has, despite the issue being discussed in a relatively abstract format, some obvious practical aspects. GOFAI systems which are designed to function in a certain domain and have correspondingly special-purpose representational mechanisms and algorithms are generally brittle and only limited to the domain for which they have been constructed. When, in the case of e.g. a dialogue system, the user wanders just a little outside the domain, or in the case of an autonomous system, the environment exhibits irregularities not accounted for earlier, a breakdown is unavoidable. This "wandering" has to be understood in a natural sense; when we are buying a flight ticket, for example, we might start chatting with the clerk, and such small talk would count as a natural part of the whole exchange. In

case we are talking with an automated system, however, each word uttered has to be strictly pertinent to the ticket buying situation. Therefore, a breakdown which happens in the case of the user uttering something unexpected refers not to the system going broke or malfunctioning, but its mechanic/programmatic nature becoming obvious, such as giving nonsensical responses or demanding correctly formatted input. These breakdowns are the moments where the discord between the nature of our mental processes and those of the computational systems becomes obvious. The brittleness of such designed systems brings up the issue that the meaning of their output is parasitic on our interpretation in a certain context and for certain aims.

2.1.2.4 Modern successes of AI

Modern AI research is carried out in many relevant fields, such as machine learning and robotics. This work is usually organized around certain well-defined tasks and competitions. The best-known of these are the Robo-Cup competition, in which simulated or robotic agents compete against each other in football, and the DARPA challenge, which requires navigating different kinds of terrains with mobile vehicles. When the outcomes of these contests are examined, it is obvious that there is considerable progress in such central areas as planning, building world models from noisy sensory data, and motion control. Although these competitions have not been without their exaggerated claims of future human-level capabilities[7], the pragmatic orientation of such research means that comparisons with human capacities are secondary to practical concerns.

It is nevertheless still all too obvious that these successful systems exhibit extreme specialization in one single area, such as navigation or motor control. The competitions have a great role in delineating these areas, and posing clear goals for the researchers. It is still important to carry out research which tries to bridge different domains, but this is not possible without at first simplifying these to understand the essence of the problems, and to clarify the underlying theoretical difficulties.

2.1.3 Cognitive Psychology

Cognitive psychology, which studies the mental processes "by which the sensory input is transformed, reduced, elaborated, stored, recovered and used" (Neisser, 1967, p.4-5), is based on the same set of metaphors and computational ideas as AI. The main difference is that cognitive psychology starts from the opposite direction, and aims to decode the human cognitive system by analyzing it in terms of knowledge structures and information processing mechanisms which

[7] From the official Robo-Cup website at `http://www.robocup.org/overview/22.html`: "By mid-21st century, a team of fully autonomous humanoid robot soccer players shall win the soccer game, comply with the official rule of the FIFA, against the winner of the most recent World Cup."

manipulate stores of such structures. The paradigmatic approach is presenting simple stimuli in carefully controlled experimental settings, analyzing the responses with the assumption that these are the outputs of a computational system, and deducing from these the mechanisms which are responsible for the particular kinds of output received. The principle aim is to start from general processes, taking these apart into progressively simpler and more specialized algorithms, to arrive at a level which can be instantiated on computing machinery (Lachman et al., 1979).

Symbolic representations play a role in cognitive psychology just as important as in AI, due to the level at which the algorithms and knowledge structures are defined. This is called the "knowledge level" (Newell, 1982), and this knowledge is coded in terms of the observer-dependent symbols discussed above. Inherent in this view is the existence of individual pieces of symbolic knowledge in the head, knowledge which can be recalled, reused and applied to novel situations through transformations. In AI methodology, this knowledge is programmed in and extracted from the environment; a major part of AI research is concerned with finding a suitable representational formalism for different domains. In cognitive psychology, a major goal is to find the putative knowledge representations used by humans. As it is the case with AI, the status of this knowledge is similar to that of linguistic representations, and in its ideal form it is assumed to be abstract, and context- and task-independent.

Lave (1988) presents, in the context of research on learning transfer and analogical learning, an elaborate overview of the intricacies of this view of knowledge, and how it affects experimental setups and the interpretation of results from these experiments. She reviews four articles discussing 13 experiments in the field. In all these experiments, the task to be studied is constructed as a "problem solving" task, by the experimenters, and not by the subjects. To put it in a different way, the formulation of the task is already given as a problem, although the subjects, had they faced the same task in real life, could have formulated and coped with it through a different schema. Furthermore, these problems do not serve anything else other than their own solution; solved problems do not serve any other activity or consequence. A striking result from the experiments is how little transfer there is. If knowledge consists of the storage and application of schemas and tools to novel situations, transfer should have been all too easy to observe. In contrast to such a view of knowledge, the successful cases of transfer rely not too seldom on the use of perceptual mnemonics, such as the perception of two scenes, between which transfer was supposed to happen, through the same guiding phrase.

The status of this assumed representational knowledge as context-independent and abstract leads to difficulties in understanding how human beings acquire and use it at all, and how it is bound to the task and context. The coupling of knowledge to context-independent representations leads to a trivialization of context, which in turn serves to equate the lacking experimental context with the various real-life contexts. This, however is a direct result of the view of knowledge as a tool to be applied to situations, without themselves being effected by the situations: "Since situations are not assumed to impinge

on the tool itself, a theory of learning transfer does not require an account of situations, much less of relations among them" (Lave, 1988, p.41). In the light of these points, Lave (1988) comes to the following conclusion regarding work on knowledge transfer: "There are, then, two consistent, well-structured *lacunae* in this work. One concerns the absent social situation, the other a silence about what motivates problem solving and the transfer of knowledge from one setting to another" (p.42). A possible reason for this emphasis on the abstractness of knowledge, and the conceptualization of activities through which knowledge is created and used as problem solving, model construction, and logical inference is that cognitive psychology has taken the scientist as a model of human thinking; as Cohen-Cole (2005) argues, "From the earliest days of cognitive science, studies of human mental processes treated thinking, perception, and language as relying on scientific methods such as hypothesis formation and theory construction" (p.119). These processes are modelled in terms of logical algorithms. The aspects of scientific thinking which were picked to inform psychological modelling were those which stressed the individual's value as a creative thinker and inventor: "[These psychologists] selected only certain aspects of the scientific process to compare to thinking: inference, invention, problem solving, making hypotheses, and model construction" (p.120). How the information was collected in the first place, the role of the scientific community in the filtering and organization of knowledge, and the amount of correction going into theory building at every stage were simply ignored.

The work of Lave (1988) is a good example of approaches to cognition which have a critical perspective on cognitive psychology, but do not wish to take a simplistic view of the human mental apparatus either. Their aim therefore is to enrich the idea of cognition so as to be able to study it in its full richness. The first step in such a project is to criticize the narrow view of cognition taken by cognitivism; in their review of such criticisms, Osbeck et al. (2007) summarize the criticisms arrayed at cognitivism as follows: "Collectively, these charges of mechanism, dualism, passivity, disembodiment, individualism and isolation from context have done much to raise awareness of the shortcomings of cognitivism as a psychological paradigm ... it is the *individualist* and *formalist* aspects that received the most attention from critics and that have inspired the most fervent turns to alternative frameworks and methodologies" (p.249).

2.2 Studying the human mind

Psychology has the peculiar quality that it relies heavily on metaphors. In every decade, the scientific method of studying the human being corresponds to finding parallels with the current state-of-the-art in technology and using these parallels to expound theories.[8] As McReynolds (1980) points out, the recognition of

[8] See e.g. McReynolds (1980) for an interesting account of the effect of the development of mechanical clocks on metaphorical talk of "springs of action" and the use of these mechanical systems as metaphors in psychological theory. This metaphor was then used in various ways to make a black box model of human science comprehensible.

technological artefacts as possible sources of metaphorical objects depends on the cultural receptivity and the existence of certain tendencies. The well-known metaphor directing cognitive science in general and cognitive psychology in particular is that of the computer. The computer metaphor replaced the control system metaphor promptly, although both arose at the same time, roughly after the Second World War (Gardner, 1985).

The computer enabled the formulation of theories about the human being, stated earlier in similar forms but without the support of the computer as a demonstrating instance, with much more clarity and convincing momentum. It served as a means of introducing a certain discourse to the scientific and popular-scientific sphere, and legitimizing this discourse through the sheer technological power of the modern computational hardware. In its crude form, this took the form of the argument of extrapolation of computational power, and the comparison of this power to human computational abilities. Another related form the computer metaphor took is using names of human capacities for functional components of a computer, and then seeing a theory of these exact same parts in these so named components (von Foerster, 1980). This can take on more insidious forms, as can be seen in some modern models of psychological phenomena. Smith (1996) calls such reading back of theories from engineered systems an *inscription error*: The box built in with the expected function is then discovered to be the correct unit when it functions the way it was designed to.

The computer metaphor flourished on fertile soil, based on the ideas which preceded it; as Fodor (1981) points out, "the computer metaphor predates the computer by about three hundred years" (p.140). The enthusiastic reception it received in certain quarters is a result of the philosophical background especially in the analytic tradition. The idea of meaning on which cognitive science traditionally relies is the one based on intentionality, famously attributed to Brentano. His analysis differentiated between physical phenomena (such as kicking a football) which does not exhibit intentionality, and mental phenomena which does (Bechtel, 1988). In other words, "mental phenomena are those that which contain an object intentionally within themselves" (quoted in Dummett, 1994). As Dummett (1994) points out, "the relation of such a physical act to its object is external, that of a mental act to its object internal" (p.31). The central dilemma of the representationalist position is the necessary distinction between the things about which mental processes have to be and the objects in the external world. This distinction leads to a number of serious philosophical questions. The most notorious of these is the matter of what mental processes or representations are about: If they already contain the thing about which they are supposed to be, are they really about the external objects, or the internal stand-ins for them? If a special status is to be accepted for mental representations, it is very difficult to avoid a solipsist position which would delimit psychology and explanation for mental phenomena to an inner world in the head of the individual.

When such a solipsist position is taken (as in Fodor, 1980), there appears the problem of how to explain the content of a representation, a philosophical version

of the symbol grounding problem. Due to the representational system being cut off from the world, or in the case of AI, removed by perceptual mechanisms still not in sight, the content of the representation must be derived from an alternative source. As Haugeland (1990) argues, the solution proposed by the philosophers taking this stance is to point at the role of the symbols in the whole mechanism; in case the whole system makes sense as e.g. a chess player when we attribute to one symbol the role of the representation of the position of a pawn, we can assume that this attribution, and all the other attributions we make for the elements of the system with this interpretation, are correct. However, the human mind is not just a text to be interpreted, but a production system on its own; how is it possible that such production takes place if the representations do not refer to patterns in another domain, or at least, these patterns are available only when the symbols are activated? The answer to this question gives the key to the mutual dependency of the representational point of view and the symbol processing paradigm:

> Such specification is possible in general if internal symbols and internal operations are semantically articulated – that is, systematically composed of atomic symbols and operations, such that the content and validity of the composites is structurally determined by that of their components. The relevant formal level, therefore, is not physical but syntactical, and hence also digital [...] In other words, semantic activity is possible if thinking is an internal computational process conducted, so to speak, in a language of thought. This is why artificial intelligence and cognitive science belong at the first base. (Haugeland, 1990)

The computer metaphor, computational methodology in cognitive science and the rationalist view of the mind have created together what can be called a "perfect storm" by supplanting each other with a philosophical background, metaphors to talk with, and concrete examples to demonstrate the viability of the abstract mechanisms envisioned. As Fodor (1994) succinctly argues, the combination of the symbol processing machine and the replacement of semantics with syntax can be presented through the use of the computers much more forcibly as a viable theory of mental phenomena:

> Within certain famous limits, the semantic relation that holds between two symbols when the proposition expressed by the one is implied by the proposition expressed by the other can be mimicked by the syntactic relations in virtue of which one of the symbols is derivable from the other. We can therefore build machines which have, again within famous limits, the following property: the operations of such a machine consist entirely of transformations of symbols; in the course of performing these operations, the machine is sensitive solely to the syntactic properties of the symbols; and the operations that the machine performs on the symbols are entirely confined to alterations of their shapes. Yet the machine is so devised that it

> will transform one symbol into another if and only if the symbols are transformed stand in certain *semantic* relations; e.g. the relation that the premises bear to the conclusion in a valid argument. Such machines – computers, of course – just *are* environments in which the causal role of a symbol token is made to parallel the inferential role of the proposition that it expresses.

This combination of representationalism and the computer metaphor had two important consequences, which were so deeply ingrained in cognitive science for a long time that they were not even clearly formulated, even though they were pillars of research. The first of these is the individualistic character of the search for the original mental. The individuals are the ones that represent, and all the accounts rely on the power of the individual to keep up such webs and structures of representational powers standing. Hutchins (1996) states the individualistic character of cognitivism rather clearly:

> Having failed to notice that the central metaphor of the physical-symbol-system hypothesis captured the properties of a socio-cultural system rather than those of the individual mind, AI and information-processing psychology proposed some radical conceptual surgery for the modeled human. The brain was removed and replaced with a computer. The surgery was a success. However, there was an unintended side effect: the hands, the eyes, the ears, the nose, the mouth, and the emotions fell away when the brain was replaced by a computer ... the definition of cognition has been unhooked from interaction with the world. (p.363)

The second consequence is the fundamental division of the world into two: the physical world and the mental world in the head of the agent. The role of the central processing apparatus is to function on the inner world and devise plans aimed at reaching goals; the sensory and motor apparatus have the duty of making sure that this inner world confers with the physical one. In GOFAI and early robotics, this distinction is very clearly seen in the form of the internal world model. As Agre (1997) argues, the inner world was not limited to AI: "Concisely put, mentalism provides a simple formula that gives plausible answers to all questions of psychological research: put it in the head" (p.51).

2.3 A wider perspective

The main reason that the two fundamental tendencies of cognitivism, limiting cognition to the individual and establishing a realm parallel to the shared world in the mental, have been taken to be self-evident and integrated to the scientific methodology without scrutiny is that these tendencies are also deeply ingrained in the general way we think about ourselves and the our relationship with the world. This fact derives from the historical context in which the development of modern philosophical traditions were embedded. In the scholastic period,

the fundamentals of knowledge were safe and clear: the Bible, coupled with a Christian interpretation of the Aristotelian philosophical system, was the utmost arbiter of truth and source of fundamental knowledge. As the authority of the Bible started waning with the authority of the Church, and the physical sciences started gaining ground with their mathematical and experimental methods, European philosophers started looking for an alternative grounding of knowledge. The emerging physical sciences presented a different world to the human being, one which it could understand using its own faculties, but one that at the same time was hostile, as a distance between the real and the perceived was introduced through the idealization of objects (frictionless surfaces, perfect spheres), and introduction of fundamental laws which contradicted everyday experience directly (e.g. bodies which maintain their motion indefinitely). Nature could now be understood and controlled, but only through the language of the computable; mathematics was the medium through which nature spoke to us, and whoever wanted to understand it, had to grasp mathematics. Mathematical formulation became the fundament for the natural sciences, the "exact sciences". The lack of a fundamental text for the human was obvious. Descartes, who already supplied a fundamental mode of understanding for the mathematical sciences with his Cartesian geometry, started looking for such a formulation and language for the human being. The senses could be cheated, and the world was but just one complete attempt at fooling the human being. He found this formulation in the clear and distinct ideas, which found their root back in God.

Descartes' search for a new original and founding text established the idealized mental realm, instead of the shared human world, as the new exclusive text to be interpreted and understood. The development of the mathematical sciences contributed to the idealization of the abstracted and idealized objects. The ideal human being came to be seen as one who could command the abstract capabilities at best. Thinking with such rigid abstractions which are context-independent and linguistically loaded came to represent the paragon of human capacities; scientific thinking, which came to be the exemplary case for such thinking, came to represent the ideal to which all human activity should strive.

As it can be seen from this short history, two central aspects to this "received view" are crucial. The first is the primacy of processes internal to the mind; these are where knowledge has to be founded, and the ground on which the ultimate method has to be developed. The second is rational thinking as the ideal of all thinking. Rationality, in its perfect form, becomes computation. The combinations of these two attitudes came to be known as Cartesianism in later philosophy and cognitive science. The centrality of Descartes' thought to the rationalist orthodoxy is explained perfectly by Taylor (1986), who points out that the philosophical project of an epistemological foundation for the sciences in particular and correct thinking in general has to be understood in the context of the assumption of knowledge as correct representation of an independent reality. If the contents of these thoughts are to supply me with correct knowledge, Descartes argues, their congruence with the outer world should be verified through a reliable method; mere coinciding of these two spheres does not suffice. Descartes connects this point to the newly rising scientific method,

to provide a foundation for the modern rigorous search for truth:

> The seeker after science is not directed away from shifting and uncertain opinion toward the order of the unchanging, as with Plato, but rather within, to the contents of his own mind [...] Descartes is the originator of the modern notion that certainty is the child of reflexive clarity, or the examination of our own ideas in abstraction from what they 'represent', which has exercised such a powerful influence on western culture. [...] the ideal of self-given certainty is a strong incentive to construe knowledge in such a way that our thought about the real can be distinguished from its objects and examined on its own (Taylor, 1986).

The primacy of the mental and the status of rational thinking as ideal and different from other kinds of activity, once established by the foundational discussions, constituted what can be called the received view of most of philosophical thinking. This received view has a number of repercussions which occupied future thinking for a long time. One important among these is scepticism. Since minds and the mechanisms they employ to perceive the outer world are so fallible, and since we have no other means of knowing the world, then how can we know for sure that we know the world at all? That is, we are captive in our own minds and cannot get out. On the other extreme end of the disengaged world and mind view lies the constructivist thesis that we *make* the world. Another result of the received view is the mind-body problem; once the real world and the mental world are separated, the distance between the two becomes unbridgeable. As an extension especially of this second problem, in cognitive science, the Cartesian attitude concerning the study of human intelligence and design of artificial agents exhibits itself in the rejection of the central importance of bodily activity and further stress on abstract thinking to the exclusion (or subsumption) of other faculties: "This denial that sensing and acting in the world require thinking, and the concomitant identification of thinking with the higher-order reasoning and abstraction paradigmatically displayed in language use is perhaps the true heart of the Cartesian attitude" (Anderson, 2003, p.93).

2.4 Alternatives to a Cartesian philosophy

As it was pointed out above, the development of AI can be viewed as a replaying of the fundamental dichotomies of philosophy, but with different responses to breakdowns. The history of the problems and dead-ends of AI research provides rich analogies when compared to the history of philosophy, as an interesting observation by Dreyfus (2007) attests:

> As I studied the RAND papers and memos, I found to my surprise that, far from replaying philosophy, the pioneers in CS had learned a lot, directly and indirectly from the philosophers. They had taken Hobbes' claim that reasoning was calculating, Descartes' mental representations, Leibniz's idea of a "universal characteristic" – a set

> of primitives in which all knowledge could be expressed –, Kant's claim that concepts were rules, Frege's formalization of such rules, and Russell's postulation of logical atoms as the building blocks of reality. In short, without realizing it, AI researchers were hard at work turning rationalist philosophy into a research program.

Therefore, it is highly instructive to look at the state of the philosophical discussion regarding the life-world of human beings and their status as knowledge producing and processing beings. The orthodox view of world as an external puzzle and mind as the internal arena to reproduce it has been criticized by a number of philosophers and philosophical traditions. Among these, the most important for cognitive science are the phenomenological tradition starting with Husserl, and finding its prime representative in Heidegger, and the language-criticial philosophy of Wittgenstein. These two philosophers and traditions have also been the most important influences for situated cognitive science, which will be overviewed in the next chapter.[9]

2.4.1 Wittgenstein

One of the most important and philosophically most pregnant reactions to the internalizing discourse is that of Wittgenstein's. Wittgenstein, after attempting a fundamental methodological grounding of philosophy with *Tractatus*, retracts a number of (especially methodological) assumptions he made in the *Tractatus*, and argues for an understanding of the ways we use words when we talk. The language game concept, which he develops to use as toy situations to exhibit the central ideas has been appropriated by researchers in AI and artificial life to study the dynamics of language. Wittgenstein, instead of establishing an elaborate theory of language, carves out the possibilities it offers, and what can be said and what, although it appears to make sense, is in reality nonsense, especially in a philosophical context. The *Tractatus*, Wittgenstein's only work published in his lifetime, aims to establish this through the use of the logical methods introduced by Frege and Russell, and places itself in the same vein of looking at language to diagnose the problems philosophers got entangled in: "Alle Philosophie ist 'Sprachkritik' " (Wittgenstein, 1984, §4.0031).

In contrast to the *Tractatus*, the *Philosophical Investigations* looks at the ways humans *use* language, how this use is connected to daily practice, and the interfaces between everyday language and its use in philosophy[10]. *Philosophical Investigations* maintains the transcendental focus of the *Tractatus*, in that the conditions for possibility of certain phenomena are studied, albeit with more

[9] The discussion here glosses over the many intricacies discussed in the literature, and the discussion stirred by these individual philosophers. The aim here is to show that there are alternative philosophical construals which can go very deep, and not to give a complete overview.

[10] As it was mentioned above, most of the major themes concerning language in the *Philosophical Investigations* are already mentioned in the *Tractatus*, but not fully developed, and not treated as the primary subject matter. See Blume and Demmerling (1998), Chapter 2 for further details.

accent on how we talk about these phenomena, and the ways our own language tricks us (§90); this is what Wittgenstein calls "grammatical investigation". Wittgenstein is concerned with the ways we get enchanted by language itself, and he argues that most philosophical problems stem from misuse of language, and disregard for the conditions in which proper use of linguistic forms makes sense and are warranted.

Philosophical Investigations starts with a criticism of what Wittgenstein sees as the prevalent view of meaning, which he exemplifies with a quote from Augustinus. In this quote, Augustinus claims to remember how he learned to speak: The adults in his environment pointed to objects, whereby they uttered their names, giving him the chance to connect words with their objects[11]. The essence of this view of language and language acquisition is the idea that "Jedes Wort hat eine Bedeutung. Diese Bedeutung ist dem Wort zugeordnet. Sie ist der Gegenstand, für welchen das Wort steht" (Wittgenstein, 2001, §1). In order to criticize this conception, which he calls "Augustine's conception of language" (§4), Wittgenstein brings our intuitions regarding language under scrutiny.

In order to undermine the intuitions we bring to the understanding of language, Wittgenstein introduces the concept of language games in the *Blue Book* (Wittgenstein, 1991):

> Language games are the forms of language with which a child begins to make use of words. The study of language games is the study of primitive forms of language or primitive languages. If we want to study the problems of truth and falsehood, of the agreement and disagreement of propositions with reality, of the nature of assertion, assumption and question, we shall with great advantage look at primitive forms of language in which these forms of thinking appear without the confusing background of highly complicated processes of thought. (§17)

Until and through *Philosophical Investigations*, the idea of language games is developed and extended to the study of invented language game setups and fragments of actual linguistic processes.[12] Wittgenstein uses the idea of *language games* to demonstrate and study linguistic communication; these are prototypical cases of language use, such as giving and obeying a command, describing an object, telling a story, making assumptions about a certain event etc. (§23).

There are two important features of Wittgenstein's concept of language games (Wittgenstein, 2001, §7). The first is that they depict processes where the use of symbols in communication is interconnected with concrete and situated practice. The simplest language game example given by Wittgenstein involves

[11] It must be pointed out that the choice of the text by Augustinus, despite many other texts exhibiting the same spirit on language, is not coincidental; in the preceding paragraphs of the same work quoted by Wittgenstein (the Confessions), Augustinus paints a picture of the human baby already having wishes inside it, and the other people outside, to whom it cannot communicate its desires. This metaphor of inner/outer is also the topic of further chapters of *Philosophical Investigations*; see p.38 of McGinn (1997) for details.

[12] For a review of the idea of language games and the change it has undergone in various works of Wittgenstein, see Baker and Hacker (1984), Section III.

a construction worker and his assistant. The construction worker uses certain tools and materials, which he refers to using certain words, such as "block, pillar, slab, beam" (Wittgenstein, 2001, §2), and his assistant then fetches these objects for him. This language game is also meant to resemble a simple language acquisition scenario, through which a child gets introduced into the general practice of language use. The general concept of language games and the simplest example have been picked up by cognitive science as a useful setup with the use of which to study linguistic communication and symbol use; these subjects will be discussed in Section 4.2.

The second important feature of the concept of language games is the use of the word *game* to describe these situations. As Wittgenstein (2001) points out in §3, the concept of game is very difficult to pinpoint and define, although we are very adept at using it in communication and agreeing on what is a game and what is not. Further, games do not have a common feature set either (§66); they rather share certain features, like the members of a family, but none of these features is common to all games. A set of objects which belong together in such a set through shared similarities none of which are definitive of the set is said to possess "family resemblance" (Wittgenstein, 2001, §66). Such a case of family resemblance is also valid of Augustinus' depiction of how he learned language, the simple case of language use in the first langugage game mentioned by Wittgenstein, and especially the ideal languages which rely on the power of assertions to depict the world: we can count different individual episodes of language acquisition and use, and each one of them is a "game", but there are many more variations to them which would still count as linguistic communication, and we cannot expect to exhaustively arrive at a definition. All we can hope for is to study those individual cases to discover ever more aspects.

This second point is especially relevant for cognitive science, a field that has an extremely multi-faceted phenomenon as its scientific object. At another point in *Philosophical Investigations*, Wittgenstein comes back to the problem of the status of toy situations (such as his language games) in comparison to the authentic human phenomena. He claims that in philosophy very often the ideal case is opposed to the real phenomenon as a preconception to which it has to live up to (§131); language is a great example for this mistake, as the human language has been compared to all kinds of ideal languages and found missing many desirable qualities. The many aspects of human intelligence have a similar status in cognitive science: compared to ideally rational mechanisms, human behavior is occasionally found to be not rational in a strict sense. Furthermore, various aspects of human behavior are seen as individual subjects independent of each other, and a single model of a subject is assumed to completely explain it. Instead of such a tendency of equating the artificially constructed toy scenarios as equals or ideal cases of the original, Wittgenstein argues for using them as "Vergleichsobjekte, die durch Ähnlichkeit und Unähnlichkeit ein Licht in die Verhältnisse unsrer Sprache werfen sollen" (Wittgenstein, 2001, §130).[13]

[13]Clancey (1997) observes the relevance of this attitude for AI:

One of the primary lessons of this first language game relevant to the current purposes is the codependency of what Wittgenstein calls forms of life and languages: "Und eine Sprache vorstellen heisst, sich eine Lebensform vorstellen" (§19). As the most simple language game demonstrates, all kinds of language use take place in a certain setting where other capabilities of the individuals are necessarily employed to take part in a social practice. Furthermore, various things we can say about the people that take part in the game make sense only in the context of the game and in how they use the signs; for example, when the construction worker says "beam" and means that he *wants* a beam, this wanting is not in the word, or in a certain relationship his mental state has to the concept *beam*, but in the language game, and the ways he can employ the signs in this game. Therefore, the use of linguistic symbols have to be seen in the context of the non-linguistic behavior which they help form, and which they derive their primary significance from: together, they create a form of life. [14]

The first language game is used by Wittgenstein to demonstrate the futility of attributing a meaning to the words used in the game and then concentrating on these meanings as the hidden thing that will provide an explanation. One can assume that the assistant, upon hearing the word "slab", forms an image of the physical object in his head, and this image (the concept) would be the meaning of this word. Even if this were so, this would not correspond to the case at hand: that this word is used in such and such a way in the game, and this complete configuration of embodied and social practice constitutes the meaning of the word. Regarding the same point, Wittgenstein asks us to imagine a tribe whose language is limited by the one in this simple language game (§6). The initiation of the children into this language game would involve "ostensive definition", an adult pointing to an object and uttering a name. Even if we were to accept that he could somehow single out the object as the topic and connect the word with his mental image of the object, can it be said that he has learned to understand the word? No, because his understanding the word would require another round of instruction to teach the child to fetch the object when its name is uttered; if this instruction round taught the child a different practice, his understanding of the word would also be different.

The most relevant lesson for cognitive science to be gleaned from Wittgenstein's discussion of language and meaning in *Philosophical Investigations* is the relationship of the public language to an inner language which is more precise, and which serves to ground all meaning. A typical way of analyzing the short imperative sentence "Slab!", which commands the hearer to fetch a slab, is to view it as the short version of "Bring me a slab!"; this longer sentence is more precise, and states what the utterer wants more clearly. Therefore, one is in-

> Emphasizing the similarities between people and computer models, rather than the differences, is an ironic strategy for AI researchers to adopt, given that one of the central accomplishments of AI has been the formalization of *means-ends analysis* as a problem-solving method: Progress in solving a problem can be made by describing the difference between the current state and a goal state and then making a move that attempts to bridge the gap. (p.6)

[14] For a similar notion in AI, see Agre (1985).

clined to say "The meaning of 'Slab!' in the primitive language of the builders is 'Bring me a slab!' ". From here, it is just a small step to the conclusion that the mental representation, the ultimate meaning, of the sentence "Slab!" in the head of the utterer should be a corresponding sentence in mentalese, the translation of "Bring me a slab" into an internal language. The listener, when she tries to understand this command, decodes this sentence to arrive at a similar representation in her own mental language:

> [T]o assert that unconscious or implicit thought of the longer sentence always occurs is already to assume that "Bring me a slab" is a more accurate expression of the "elliptical" command "Slab!": it is to assume that our operations with "Slab!" are mediated by some more basic relation to "Bring me a slab" (Goldfarb, 1983, p.276).

The most important reason to resist this urge to locate a definite meaning in an internal sphere in the description of a (possibly) elliptical sentence is the regress of description. There is no end in sight when one starts to ground sentences in their descriptions/translations in mentalese; a description can go on forever when the terms in each individual explanation are to be described themselves (§87).[15]

Wittgenstein's work on language games provides a strong argument for the role of the embodied and social practice in understanding symbols use; this is the primary reason it has been taken as a suitable analogy for research into the dynamics of communication. However, his arguments against equating our models of things with the real phenomena, and the various aspects of complex phenomena like communication also have to be taken seriously.

2.4.2 Phenomenology

The domination of the scientific innovations and technological advances, especially at the beginning of the twentieth century, led to the crude scientific worldview to become common sense. The basis of this worldview was representationalist and quasi-platonistic: The world consists of objects, which we perceive and construct in our mind, and science is about the essence of these objects, how they are in themselves. These objects, which appear in our experience as the imperfect copies of the ideal bodies they are in the scientific practice, occupy the geometrical space which we also inhabit. Husserl, in his criticism of this worldview, and the rapid appropriation it has seen in also philosophy and psychology, argued that the ideal objects of this worldview, such as frictionless surfaces or perfectly homogeneous solid objects, were abstractions which emerged out of a certain kind of involvement with the world, and certain social practices (Husserl, 1982). In order to bracket out the preconceptions with which human beings see their world and conceptualize their involvement with it, he proposed what he called the phenomenological method. This method

[15] This argument does not have much force in a cognitive scientific context, since, computationally, there is no serious problem with a highly interconnected network.

consisted of accepting the world as given to the human being in a certain way, and disregarding the questions of how it actually is independent of human experience; more precisely, *phenomenological reduction* aimed at a "redirection of thought away from its unreflective and unexamined immersion in experience of the world to the way in which the world manifests itself to us" (Thompson and Zahavi, 2007).

A student and colleague of Husserl's, Heidegger, took Husserl's phenomenological method further, both in the historical and the transcendental sense.[16] In the historical sense, he urged a reevaluation of the history of western philosophy, especially since the works of Descartes, and turning away from the dominant epistemological concerns this tradition has led to. In the transcendental sense, Heidegger argued for a primacy of the question of being in philosophy. Among the things that make his philosophical stance revolutionary, the integrity of these two aspects comes second only to the specific language with which Heidegger tackled these issues. This language is a very peculiar one, relying on the productive aspects of his native German language, and repetition of key phrases to avoid losing the meaning of certain words to a familiarization effect.

Heidegger (2001) characterizes the intentional sphere of the human being as one structured by the human being itself, where this structuring is the basis of all further inquiries, scientific or philosophical. Therefore, in order to answer the question of what is, which can be paraphrased as the question of how the world in which humans live comes into existence as a space of meaningful entities and, simply, objects we can refer to and conceptualize, a foundational ontology has to start from the understanding the human being has of this world even when asking the question: "Die Seinsfrage ist dann aber nichts anderes als die Radikalisierung einer zum Dasein selbst gehörigen wesenhaften Seinstendenz, des vorontologischen Seinsverstëndnisses" (p.15).

Since we are cultural-historical beings, however, the understanding the Dasein has of being is possible against a certain background and carries with it its own development; it is historical. It has been occluded by the turns western culture and philosophy have taken. The most consequential of these turns is the Cartesian turn, which instilled an understanding of the world as something independent from us, and ourselves as individuals trying to make sense of it. The Cartesian turn is a result of the tendency of the Dasein to understand itself in terms of its own world, what Heidegger (2001) calls the "ontologische Rückstrahlung des Weltverständnisses auf die Daseinsauslegung" (p.16). The Cartesian framework has become a background, the shared common sense, when it comes to even asking the questions in the first place. Therefore, the first duty of a philosophy trying to clear the grounds for a proper understanding of human existence is putting itself before the traditional questions, and undertaking a "destruction" of this background. This move of Heidegger's is analog to Wittgenstein's method of getting at the way philosophical questions are for-

[16]The relationship of Heidegger's philosophy to that of Husserl, who was a teacher and mentor of Heidegger, and their personal relationship, is a subject discussed intensely. The details of their philosophical differences will not be discussed here, but the interested can refer to §12 of Tugendhat (1970).

mulated in the first place; the example of how the explanation of a sentence as simple as "Slab!" is taken to be the meaning of this one-word sentence is a typical case.

Heidegger's dismantling of the Cartesian common sense, as Guignon (1983) argues, can be summarized as consisting of two arguments (p.38). The first of these aims at the ontological assumption of the Cartesian tradition, which describes human participation in the world using a subject/object schema. Against this assumption, Heidegger argues that the immersed practice of daily life grounds the contemplative attitude, and comes before it. The second argument aims at the method with which certainty is to be established, and all knowledge is to be grounded, according to Descartes. The method proposed by Descartes is inspired by the scientific method, where a complex is first taken apart, and then understood as an interaction of its parts. Heidegger argues that the scientific method also relies on our everyday existence and certain kinds of social/embodied practices. In order to partake in such a practice, one has to be immersed in it. Therefore, a global description of our practices, and of the knowledge that grounds in these, is impossible.

Heidegger proposes to formulate an answer to this question on the practice of humans, viewing their daily activities as the necessary conditions for an objectivising and representational discourse, especially so when it comes to science. The human being can engage in certain activities, including science, only thanks to its being in the world at all times, as an agent engaged in and realizing a certain form of life. The implications of this view for the philosophical project of grounding knowledge are clear:

> What you get underlying our representation of the world – the kinds of things we formulate, for instance, in declarative sentences – is not further representation but rather a certain grasp of the world that we have as agents in it. This shows the whole epistemological construal of knowledge to be mistaken. It doesn't just consist of inner pictures of outer reality, but grounds in something quite other. And in this "foundation", the crucial move of the epistemological construal – distinguishing states of the subject (our "ideas") from features of the external world – can't be effected. We can draw a neat line between my *picture* of an object and that object, but not between my *dealing* with the object and that object (Taylor, 1986).

Therefore, it can be said that, concerning the discussion on intentionality, Heidegger opposes a strict distinction between physical and mental phenomena on the basis of their aboutness, and wants to look for the grounds of meaning primarily in the embodied and socially situated practices of human beings. This is due to the way human beings exist in the world; we are always in the world, and this *being in* is not only in the sense of physical containment, but in the sense of *caring* about it, and the things presenting themselves as tools to ends, and not meaningless blobs. Through this perception of the situation as meaningful entities, the world becomes the horizon of what we can do, and what kinds of existence are possible.

One concept from the work of Heidegger which has received attention from the AI community is his distinction of tools as being either *zuhanden* or *vorhanden*. During practical engagement, the tools used by humans for a certain purpose recede to the background and become transparent to the user, practically becoming an extension of the human embodied apparatus; in this case they are said to be *zuhanden* (ready-to-hand). In the case of a breakdown, the particular existence and structure of tools become apparent, and the tool serves not as a way of enhancing our activity, but as an impediment; in this case, such a tool is said to be *vorhanden* (present-to-hand) (Heidegger, 2001, p.66). When the everyday activities we engage in with things break down (e.g. when a hammer I use regularly is broken), I become conscious of the situatedness which normally makes my interactions possible:

> [W]hen something we intend to use is discovered to be unusable, or when something we need is missing, or when something we need to get around stands obstinately in the way [...] things are no longer there ready-to-hand, and just in that instance we catch sight of the very situatedness that normally characterizes our existence [...] The situated view (which in other terms Heidegger calls the "hermeneutic situation") is something that qualifies all theoretical knowledge and all third-person scientific accounts. Moreover, being situated is something that in its inconspicuousness tends to escape our attention, but not simply because we overlook it, in the way that we might overlook something in the environment. Rather, it is part of what it means to be situated that the fact of being situated commonly goes unnoticed (Gallagher, 2008).

Rationalizing and deliberation happen primarily in the case of such breakdowns; normally transparent situations become problem-solving settings. The major mistake traditionally made in philosophy is to confuse these problem-solving situations with the situated engagement of ordinary daily life. In such engagement, what can be called the background of the practices is not necessarily represented, as there is no need to represent them; when I am going out the door, I always lock after me. To know that the door is locked and no one can break in, I don't need to know how locks work. However, when this lock gets picked, it becomes of importance to me, and explicit knowledge enters the situation. Furthermore, this background is social in that it is co-consituted by institutions and people (in this case locksmiths, burglars etc.):

> For Heidegger, the Background is primarily socially constituted; it is a network of artifacts and institutions. We avail ourselves of it in virtue of the social practices and institutions in which we participate, and the general tendency or capacity to participate in social practices is a primordial characteristic of Dasein, which I would like to call sociality (Preston, 1993, 61).[17]

[17] For an approach which locates a considerable part of this Background in the biology, specifically evolutionary history of the being, see Wheeler (1995).

As it can be seen, Heidegger responds to the dominant tradition which equates human mental life with the existence of an inner sphere by claiming the primacy of a world in which humans always find themselves, even when taking part in scholarly speculations, and which makes it possible to engage in such activities in the first place. Inherent in such a view of the world is a questioning of the idea of aboutness as intentionality, i.e. as representation; for Heidegger, socially and physically situated practices come first, and these ground any further activities which might involve the use of representations and symbolic means of communication.

2.4.3 A philosophy for the new cognitive science

The philosophies of Wittgenstein and Heidegger not only display the hidden convictions of the rationalist and representationalist traditions on which traditional AI and cognitive science are based, they also offer alternative directions in which to look, and different analogies and comparative instances to base our models on. These are based on the rejection of Brentano's depiction of the mental as the sole intentional phenomena, arguing for an idea of intentionality as directedness, deriving from embodied interaction and cultural embeddedness. As it will be shown in the next chapter, the insights of both Wittgenstein and Heidegger have been incorporated in a new research program for cognition, and have led to radically new insights.

Chapter 3

Embodied and Situated Cognitive Science

The theoretical problems and practical breakdowns presented in the previous chapter led to the rise of an alternative framework in cognitive science starting from the early 1980s. The main focus of this approach was on the embodied and situated character of natural behavior and human intelligence. It is possible to distinguish two strands in this work. One of these is the situated approach, analyzing human intelligence and behavior from a context-bound and social perspective, taking a top-down approach. The other strand is the embodied approach, which aims to develop models of embodied behavior and create bottom-up models of it.

The situated approach has built on the philosophical background explained in Section 2.4. It has been able to build on a considerable body of work in related disciplines in the social sciences. Embodied AI, on the other hand, was born out of a changing perspective on the subject matter of AI, and an engineering insight into the impeding role of representations. Eventually, the two research groups arrived at a common vocabulary and shared research concerns. In the following, first embodied AI, and afterwards situated cognition will be discussed. Afterwards, an extended discussion of the common themes will be presented. In addition, the repercussions of research in these two areas on theories of the role of symbolic communication and representations will be overviewed.

3.1 Embodied AI

The driving forces behind an alternative to the symbolic paradigm in AI and robotics are the realization that the meaning of the symbols used in these disembodied systems was parasitic to the interpretation of the humans that design and use them (discussed in the previous section), and an alternative vision of what the product of work in AI has to be. This alternative vision exhibited itself especially in the decomposition of the agents in the design process. Instead

of relying on the functional level decomposition, corresponding to the SMPA principle mentioned in the previous chapter, a behavior-level (or task-level) decomposition method was adopted. The difference between the two is most pronounced in the work of one of the pioneers of embodied AI, Rodney Brooks. Brooks (1991) advocated going back to the basics of intelligence, namely to bodily behavior and perception. Instead of systems which reasoned about activities and plans, Brooks' robots engaged in relatively simple tasks in the real world. These tasks were simple from a cognitivist perspective, because they did not engage in complex reasoning, but difficult with the existing computational methods, because they involved real-world action and perception in noisy and dynamic environments. The robots and their behavioral modules were designed from the ground up, by developing basic building blocks of behaviors (such as moving without bumping into an object, approaching an object to grasp it etc.), testing these on real robots to bring together the sensory and the motor aspects, and afterwards moving on to more complex behaviors.

The architecture devised to design such control systems was called the subsumption architecture. The central idea of the subsumption architecture was to organize behaviors which all receive sensory input into a hierarchy of levels, with the behaviors higher up the hierarchy either inhibiting or stimulating the simpler ones (Brooks, 1986). This simple architecture allowed the construction of complex behavior from simpler components. Examples for such behavior is navigating a cluttered office area without bumping into obstacles or, on the more complex side, collecting the empty bottles on the desks in an office space (Brooks, 1990). The coupling of the agent through sensory-motor loops to the environment, instead of creating elaborate plans from a world model, was found to be one of the key ingredients for creating agents which can react to their environment in real time. This led to the now well-known principle that "The world is its own best model" (Brooks, 1991). This principle, one of the most concise statements of situatedness, has to be understood as stating a contrast between simple robots in continuous interaction with their environment and the robotic agents which relied heavily on exact world models in AI, especially in robotics.

Most important of Brooks' achievements is his redirecting AI research from isolated symbolic problem solvers to the concrete aspects of sensing and acting. With Brooks' work, one can diagnose a turn to the simple-but-complete approach to agent design of cybernetics, but without the extensive control system vocabulary and the design methodology belonging to it. Simplicity here refers, as explained above, to the focus on tasks which do not involve complex symbol-processing capabilities such as problem solving, but instead are concerned with apparently basic embodied capabilities which are in fact difficult because they have to cope with noisy and dynamic environments. In addition, Brooks presents a clear future direction for AI research by coupling his work to the more epistemic concerns of the situated approaches by presenting his work in the context of a physical grounding hypothesis, which states that "only through a physical grounding can any internal symbolic or other system find a place to bottom out, and give 'meaning' to the processing going on within the system"

(Brooks, 1991, p.15). Although the claim of embodied AI has sometimes been explained or perceived as a rejection of all kinds of representation mechanisms, the real core of the enterprise should be seen as the grounding project, which encompasses understanding all kinds of intelligence, including those concentrated on by cognitivist AI, as stemming from embodied capacities.

The contrast of functional vs. task-level decomposition is an important one, as it rises out of the differing concerns of the two approaches. In cognitivist AI, the aim is to understand individual functional units by building subsystems cohering to the specification; in embodied AI, a complete agent is taken as the fundamental unit of work (Arkin, 1998; Maes, 1990; Pfeifer and Scheier, 2001). A behavior is defined as "a regularity observed in the interaction dynamics between the characteristics and processes of a system and the characteristics and processes of an environment" (Steels, 1994); the principle aim is to compose such behaviors, through a deliberation mechanism, into a control system achieving more complex behavior.

Another area in which the principles of embodied AI have been applied is the study of animal behavior and the design of biologically inspired robots to understand animal intelligence. In the animat approach to AI, complete robotic agents are built in order to model and thereby understand the behavioral capacities of animals, with a special focus on the adaptive aspects (Dean, 1998). As (Wilson, 1991) argues, there are three important principles underlying animat research:

1. Maintaining the realism and whole-ness of the environment so as to avoid special-purpose solutions

2. Maximizing the physicality of the sensory signals, so as to avoid predefined symbolic inputs

3. Employing adaptive mechanisms maximally, to minimize the introduction of new machinery and maximize understanding of adaptation

As these principles make clear, simplicity is one of the driving goals of animat research. This simplicity is in the service of scientific objectivity, as less knowledge built in leads to more objective models. The risk of making an inscription error, that is, reading out what was already built in (cf. Section 2.2), is thus lowered. Furthermore, as the well-known ant parable of Simon (1969) reminds us, complex behavior is not always the result of complex structure, but may well stem from simple behavioral repertoire coupled with a complex environment. In order to understand such complex behavior through synthetic methods, it is important to keep the environment in which the artificial agents are tested as realistic as possible. A good example of the animat approach is that of Webb (1994), who studies the phonotaxis behavior of the cricket. The cricket females locate the males by walking or flying toward the calling song the male produces. The robotic model by Webb (1994) makes use of simple filters and direct connections between the sensory channels and motors. The robot was tested in experimental conditions comparable with those used with crickets, and was

able to demonstrate a number of key effects that occur also in experiments with crickets, such as phonotaxis, recognition of the ideal syllable rate, and choosing one sound source despite the existence of multiple sources. Webb (1994) concludes that "the cricket's response can plausibly be explained by a combination of slow auditory neural response (effective low pass filtering) and temporal summation in motor neuron response (effective high pass filtering)" (p.51). Another example is that of John et al. (2006), who built a robotic model of the visual discrimination performance of chicken. The robot used a simple similarity-based learning mechanism to acquire the correct reaction to the stimuli presented in a Skinner box. The robot was then tested in the same environment with the chicken. Despite the simplicity of the computational mechanism, and the absence of traditional perceptual filtering and representational mechanisms, the robot achieved very similar results to the chicken. It also displayed the same difference in learning speed with stimuli of different dimensions.

3.1.1 Physical coupling instead of representation

One desirable result of task-level decomposition is that the focus of the analysis of a complex, context-dependent and goal-directed behavior is not the control units inside the agent (such as an internal map, a goal database, representational libraries etc.), but the possible couplings with the environment which would lead to the emergence of the desired behavior. An alternative approach that takes this idea seriously and aims to replace the traditional talk about representations and algorithms with the temporal dynamics of continuous systems and agent-environment coupling is the dynamic systems approach (Port and van Gelder, 1995). In this approach, intelligent agents are described in terms of a differential function of the variables of the environmental interaction in which they engage and their internal states (Beer, 1995). An interesting example of a dynamical model of interaction is a model of the A-not-B error well-known in developmental psychology. The A-not-B error is made by 7-12 month old infants who are instructed to reach for an object in one of two opaque containers. The object is placed in one of the containers while the child observes. Once the child has learned reaching the first container to pick the object, in a test trial, the object is placed in the second one, in clear view of the infant. The majority of children will still reach for the first container to pick the object. Instead of the traditional theory of immaturity of object permanence, Thelen et al. (2001) created a model based on a movement planning field with two attractors. The properties of these attractors were modulated according to two external inputs, specifying the visual appearance of the field and the visual cue used to draw the infant's attention, and an internal input from a one-dimensional memory field. The movement planning field, like a surface which has two holes in it towards which the surface bends, is shaped according to these input values, and the depth of the attractors is changed according to the current visual input and memory. Thelen et al. (2001) show that the assumption of an immature goal-directed reaching system is enough to reproduce the A-not-B error with the contextual effects which cannot be accounted for by the traditional approaches.

As it can be seen from this example, the main distinguishing features of the dynamic systems approach are the mechanics of change and adaptation, and the role this mechanics serves in explanation: "Here the explanatory focus is on the structure of the space of possible trajectories and the internal and external forces that shape the particular trajectory that unfolds rather than the physical nature of the underlying mechanisms that instantiate this dynamics" (Beer, 2000). Nevertheless, there is the problem of how such a dynamic system is to be designed, and why a certain kind of task decomposition is to be favored, compared to the others. This problem has been studied in further attempts to create dynamic systems, especially the evolutionary approaches.

3.1.2 Minimally cognitive behavior

One aim common to many embodied agents approaches is the creation of autonomous agents that display "minimally cognitive behavior". As it was already mentioned, the aim of embodied AI is generally to create complete agents, that is, agents which can autonomously function in an environment, and adapt to changes in the environment and in its own states. Instead of dividing such an agent into serial modules (the SMPA architecture), complexity is achieved starting with simple but complete agents(cf. Section 3.1 above). These simple agents are minimally cognitive in that, although their sensory-motor capabilities are limited and the tasks they have to carry out rather toy-like, they are nevertheless perfectly adapted to these simple tasks and serve as starting points for more complex forms. One of the principle reasons for using simple mechanisms and behaviors is to enable as little knowledge as possible to be built into an agent. This is called the machinery parsimony principle by Agre (1995). The machinery parsimony principle allows the analysis of the task dynamics independent of the prejudices of the designer, and leads to more objective models of phenomena.

Genetic algorithms are one of the most popular methods for studying embodied intelligence with minimal prior assumptions. Originally proposed by Holland (1975), genetic algorithms are based on a computational analogy to the biological process of evolution, which consists of the genetic encoding of traits and the transmission of these traits to offspring (heredity), chance mutations in the transmitted genetic material (variation), and the evolutionary selection of the successful individuals from a population for further reproduction (selection). What is encoded in the hereditary material in the computer simulations is generally a neural network (Harvey et al., 2005), but even hardware can be evolved using genetic algorithms (Thompson, 1998). Also, the environment does not have to be simulated; for example, Floreano and Mondada (1996) evolved neural network controllers for a Khepera robot on a physical setup. Neural networks designed through artificial evolution embody dynamic systems, and are generally not assumed to have any similarities to biological systems. In addition, the biological nature of artificial evolution does not have to be stressed; genetic algorithms can also be seen as undirected incremental search in a big search space (Harvey et al., 1997).

An interesting example of work on evolutionary algorithms which displays the most significant features of this kind of work is that of Izquierdo et al. (2008). Learning in biological systems and artificial neural networks is traditionally seen as the modulation of the connection strengths between neurons according to error signals. Izquierdo et al. (2008) argue, based on recent research into natural and artificial learning, that there are other mechanisms occurring over varying timescales which are also responsible for learning. Building on previous work which achieved simple learning without synaptic plasticity (Phattanasri et al., 2007), the authors pick the behavior of a worm as the target of their model. This worm (*C. elegans*) learns at which temperature value it had received ample bacteria as food source, and moves to that temperature when it has to search for food. The structure and behavioral repertoire of the network is relatively simple. The network consists of five neurons, one of which is the output (mouth) neuron. Two inputs specify the temperature and the food situation. The connection strengths between the nodes cannot be changed during the lifetime of one agent; they are constant, and are passed on with occasional mutations to the offspring. The mouth neuron specifies whether the mouth of the worm is open or not. The selection criterion is whether the mouth is opened in the testing phase at the temperatures at which they were kept (with food) in the training phase. The most successful agents achieve 97% correct in testing trials. These agents evolve finite state machines which can handle continuous values, what the authors call "continuous state machines". The individual states in these state machines are like infinite tapes on which the networks oscillate as long as there are no perturbations. Once the environmental conditions change, however, the network changes from one such state to another.

As this example shows, the most important advantage of using genetic algorithms is that they allow minimal cognition with minimal knowledge built in. The evolved dynamics is sometimes very counterintuitive and interesting on its own right.[1] The main problem with evolutionary algorithms, however, is the simplicity of the setups. Such simplicity is necessary due to the computational requirements of artificial evolution simulations on the one hand, and the controllability of the evolutionary process on the other. For the populations to converge on a solution, the fitness function has to be precisely controlled and, if more complex behavior is targeted, progressively changed. The designer has to cope with a different kind of complexity, in that the fitness function has to account for all aspects of the desired outcome. Another problem with evolutionary methods is that they allow only limited application of and comparisons with well-known psychological methods. When the centrality of psychological methods for cognitive science is considered, this aspect proves to be a crucial shortcoming.

[1] In the following chapter, applications of the genetic algorithms to agent communication will be reviewed; multi-agents setups are among the domains in which genetic algorithms have delivered their most interesting results.

3.1.3 Embodiment as scientific principle and explanatory construct

The above explained directions of research laid the foundation of what Beer (2009) calls the SED approach for situated, embodied and dynamical approaches. One of the most important commonalities among these approaches is the significance attributed to intelligent beings being bodies in a concrete environment instead of disembodied processing mechanisms. This property, called embodiment in a scientific context, serves both as a research guideline and an explanatory construct. As a research guideline, as it was mentioned above, embodiment places the scientific value of AI agents on their ability to function in real environments with complex sensory-motor contingencies. Cognition becomes not a matter of transforming complex logical structures, but of coordinating a physical body; this is the methodical counterpart of the recognition that perception and motor control are just as complicated as other aspects of cognition and are different aspects of the same union of capacities.

As an explanatory construct, the concept of embodiment is utilized to describe the various levels at which the coupling of the agent with the environment leads to intelligent behavior. Examples are how a certain kind of tuna fish uses the currents and vortexes in the water to swim in a way that would be impossible with its own muscles, or how baseball players can catch high flying balls by coordinating their running speed so that the ball looks to them not as curving down but appears to follow a straight line (Clark, 1999). A classic synthetic example of the use of embodiment to replace computation is by Schmitz et al. (2001). When a simulated model of a hexapod robot was implemented on a physical robot, it was realized that the interaction with the environment could be used to simplify the computations needed within the agent. Specifically, the change of stress on the joints of the legs of the robot when one foot was raised made a set of sensors and related computations unnecessary. This insight into an important aspect of the behavior of the locust was made possible by the availability of an embodied model.

These are simple examples of the use of body-environment coupling to either engage in acts which would have been impossible otherwise (in the case of the tuna fish), or simplify the computational load radically (in the case of the baseball player). Such cases of coupling serve to demonstrate the role of the body in intelligence in a further and more significant extent, namely as the last context in which all activities are grounded. The human being exists in a physical situation through its body, and this situation is the context which determines the possibilities to act and socially engage with other beings (Gibbs, 2006). Therefore, what has to be studied is not the architecture of the disembodied processing mechanism, but the dynamics which arises out of the coupling of an agent with the situation at many different levels. The first of these levels is the simple sensory-motor level, of which – natural and synthetic – examples were given above.

How the low-level body-environment coupling can lead to the complex capacities of the human beings is one of the most important research projects of

the situated approach. This project is generally called the grounding project, as its aim is to understand the higher-level capacities as a natural extension of the embodied capacities (Brooks, 1990). The core idea of the grounding project is that the concrete situation, understood from the perspective of the human being, and made possible by the bodily existence of the same, is the primary object to be studied, if intelligent and meaningful human existence is to be understood. This is the idea of situatedness, which will be studied in the next section.

The most important problem that has to be solved by the grounding project is moving from the concrete and temporally coupled activities – of which bodily activities are the prime example – to more abstract and temporally extended ones. These activities are those which involve the use of representational resources, applying them in different contexts, learning new ones, and creating new ones if necessary. What differentiates these activities from the first kind is in the first place their being embedded in a wide-ranging social and institutional setting. Clark (1997) points this out on the example of a complex social activity such as ship navigation:

> The mature cognitive competencies which we identify as mind and intellect may thus be more like ship navigation than capacities of the bare biological brain. Ship navigation emerges from the well-orchestrated adaptation of an extended complex system comprising individuals, instruments, and practices. Much of what we commonly identify as our mental capacities may likewise turn out to be properties of the wider, environmentally extended systems of which human brains are just one (important) part. (p. 214)

What makes the social and institutional setting possible in the first place is language; the reasons for this primacy, how language serves to augment cognition and how it can be understood using synthetic models will be discussed in Section 3.3.

In the embodied AI and cognitive science literature, the biological roots of the idea of embodiment are stressed at various levels. Whereas it serves as an inspiration for the behavior-based approaches to build complete autonomous agents which receive raw sensory data and have physical bodies (Maes, 1993), for more biologically-oriented approaches, biological argumentation leads to equating life with cognition. Such work is inspired primarily by Maturana and Varela (1980), who argued that the defining feature of living beings is *autopoiesis*, i.e. the capacity to produce oneself. What is called "the enactive approach" tries to understand cognition by basing on this definition of life, equating life so defined with cognition, and building minimal living systems to compile a library of autopoietic processes, building progressively more cognitive systems from the ground up (Bourgine and Stewart, 2004).

3.2 Situated approaches

In contrast to the embodied approaches in AI, situated approaches to cognition take a top-down view and analyze human behavior in daily situations and try to understand the emergence of meaning and structure. This approach takes its inspiration from the philosophical alternatives to the empiricist/rationalist orthodoxy, discussed in Section 2.4, but at the same tries to maintain a cognitive orientation, in that the focus is on the representational capacities of human beings, or rather communities in their concrete situations.

In one of the earliest studies in the situated paradigm, Agre and Chapman (1990) work in an AI framework to question the prevalent idea of plans and offer an alternative one. The question of structured behavior is a fundamental one for cognitive science. Traditionally, the source of structured behavior is located in the head, as it was pointed in Chapter 2, in the form of plans. Plans, in the traditional view, are serial commands to be executed in order to reach a goal or achieve a certain state of affairs; this is what Agre and Chapman (1990) call the plan-as-program view. These plans are created by a central controller, and then passed on to an executive model which carries out the commands, and in the case of any problems or contingencies, returns control to the central module. The most important problem with such a view of plans is the reliance on a world model. The execution of a plan takes place in the real world (or whatever domain the program is functioning in), whereas the creation of the plan is done using a world model. This leads to all the objects in the environment having to be specified explicitly in the plan passed on to the execution module. Furthermore, the execution module misses any opportunities in the environment, since the recognition of these opportunities requires the capabilities of the central module.[2] As an alternative to the plan-as-program view, Agre and Chapman (1990) propose the plan-as-communication view. Contrasting the way humans use linguistic resources, such as plans, with the execution of a program, Agre and Chapman (1990) propose to model activity as arising out of the interaction of a simple machinery with a complex environment constantly in motion, without internalizing the environment. In essence, their aim is to stay true to the "The world is its own best model" maxim of Brooks (1991), and unite the planning and executing aspects of traditional approaches in a system that *looks* at the world and sees the necessary steps to be taken.

Agre and Chapman (1990) present two examples of agents that use linguistic and representational resources in a task setting. The first of these, Pengi, is a program that plays a computer game called Pengo (Agre and Chapman, 1987). In this game, the aim of the player is to survive an arena populated with hostile entities by manipulating pieces in the environment to attack these entities. The agent built by Agre and Chapman (1987) plays the game using a number of techniques devised as an alternative to the traditional approaches. One of these is the use of deictic representations as an alternative to context-independent representations, which a planner would need to include in the plans it submits

[2] (Chapman, 1985) also documents practical difficulties with the plan-as-program approach, especially regarding intractability and combinatorial explosion.

to the executor in order to identify the correspondence between the objects in the world model and the real world. These deictic representations are attached, as markers in the visual system, to indexically and functionally individuated objects in the game screen. The objects are indexically individuated in that this labelling depends on the perspective of the agent; if an object is not relevant for the current situation, it is not labelled. The objects are functionally individuated in that the content of the label is dependent on what the agent is doing at the moment. There is, for example, a marker for `the-bee-i-am-chasing`, which is overlaid on the hostile bee. This marker then directs the behavior of the agent, and in case this individual bee disappears from view, the marker is removed. As in the work of Brooks, Agre and Chapman (1990) rely on the interaction of a simple behavior network with a complex environment for the emergence of complex behavior[3]. Behavioral arbitration is achieved through a hand-coded network of behavioral rules. These rules and their arbitration is written by the designer as a result of the analysis of the task. Another commonality with behavior-based robotics is the already mentioned absence of an inner world model: "When Pengo needs to know where something is, it doesn't look in a database, it looks at the screen". As it can be seen in the behavior-based AI examples given above and the example of Pengi, these two aspects, tight coupling of the agent with the environment and situatedness, are two sides of the same coin. Together, these two concepts build an alternative explanation for complex structure in behavior, as contrasted to plans in the head, or inner world models maintained through representational relationships with the world.

An extension of the plan-as-communication view is Sonja (Chapman, 1991), which also plays a computer game called Amazon, a very similar game to Pengo. The aim in this program was to study situated instruction use. Situated instructions are those given in the course of a social activity in a common situation, such as playing a cooperative computer game together. Situated instructions are syntactically simple, because both players possess a common knowledge of the game. They have to be obeyed in a temporally fitting fashion, because they are given pertinent to the situation at hand in a certain moment. In their observation of experienced players who played Amazon together, Agre and Chapman (1990) recognized that the players relied to a great extent on their common knowledge of the game, and the possibilities they perceived on the basis of this common knowledge. For example, in a situation where two choices were available (two doors, turning left or right etc.), one player could simply say "No!" to avoid the other continuing with the choice he made, and pick the other one. This is possible because both players see clearly that there are only two choices. Another example is where one player commands the other to turn left. Satisfying this instruction can take many forms, such as waiting until the next left turn, first killing an adversary on the right and then going left, or picking a resource somewhere else before turning left. All of these actions count as turning left, despite their completely different flow, because they get to be interpreted

[3] A similar example which was mentioned in Section 3.1 is the ant which walks on the beach and creates complex patterns not because it intends do so, but because of the structure of the beach (Simon, 1969).

to be turning left based on the perception of the situation. The program built by Chapman (1991), Sonja, uses instructions given by an instructor to make decisions and better manage its activities. In doing this, it uses similar methods to Pengi, such as indexical representations and visual routines. The use of instructions is guided by the contingencies of the situation at hand, in that these instructions are not treated as commands, but as recommendations. Any more important actions are handled first, before an instruction is processed.

In these and similar studies in the situated paradigm, the practices of people in shared situations, and their use of symbolic resources in these situations has been the primary subject matter. The studies mentioned in this chapter, the breakdown of the cognitivist paradigm, and the alternative philosophical views to the tradition have led to the status of symbolic communication and symbol use as they relate to human intelligence and behavior to be reconsidered. What was questioned primarily were theories of language and symbolic communication which view it as a means of encoding and exchanging information. According to this view, as discussed in Chapter 2, the meaning of language derives from the representations in the participants' head, and understanding a linguistic utterance consists of matching it onto an inner representation. In contrast to the picture depicted by this approach which leaves out the environment, it has been argued that conversation takes place in a concrete environment and is dependent on this environment not only as a resource for a limited set of information. Rather, a situation is an infinite frame of reference; "every utterance's situation comprises an indefinite range of possibly relevant features" (Suchman, 1987, p.60). The situation also figures as a repository for many things not mentioned in speech: "Our situated use of language, and consequently language's significance, presupposes and implies an horizon of things that are never actually mentioned" (Suchman, 1987, p.60).[4]

The status of the situation as a repository of possible means of determining the references and meanings of utterances also serves as an explanation of the vagueness of language and symbolic communication in general. Vagueness is traditionally seen as a weakness of human language, and in proposed symbolic communication schemes, definiteness of formulations is praised and seen as an advantage. The vagueness of language is not a weakness; it is a central feature of all mechanisms and resources which require interpretation. The situation is searched for possible meaningful correspondences with an utterance, and this search also gives the situation a new form, and makes it possible for the agents to see new possibilities and chances. This feature of language is shared by

[4] A field primarily occupied with the constitution of social structures, and how these emerge through the daily interactions of humans is ethnomethodology. The ethnomethodological movement was born in part as a reaction to the standard view in the social sciences that the social scientist had to take the existence of an objective world of social facts or received norms for granted, and then try to describe human attitudes and actions as a response; this principle is embodied in Durkheim's maxim that "the objective reality of social facts is sociology's fundamental principle" (Durkheim, 1938). A fundamental starting point of the ethnomethodological method is replacing the conception of human behavior as reacting to an objectively given social world with the assumption that "our everyday social practices render the world publicly available and mutually intelligible" (Suchman, 1987, p.57).

all resources used in a situated fashion by human beings, e.g. plans: "It is precisely because our plans are inherently vague – because we can state our intentions without having to describe the course that our actions will take – that an intentional vocabulary is so useful for our everyday affairs" (Suchman, 1987, p.38).

If such descriptions of the relationship of language, social communicative practices, and perceptual situatedness are not supplemented with computational observations and concepts, it is difficult not to fall back to the traditional method of building in such descriptions into cognitivist computational examples, or interpret existing systems simply with a different vocabulary. In his analysis of situated cognition and research in robotics, Clancey (1997) provides an analysis which shows what kinds of computational methods are necessary for implementing and computationally understanding situatedness, and how cases in which such methods were at work in collaborative situations has been neglected by the traditional approaches. Clancey (1997) picks the case study of a group of people working on the creation of a synthetic paintbrush, documented by Schön (1978). The first synthetic brush designed does not function properly, and the inventors go back to studying a natural brush to find out what is essential to the proper functioning of a brush. At a certain point, one of the inventors recognizes that the paint is stored in the capillaries between the hairs of the brush, which leads him to proclaim that "A paintbrush is a kind of pump" (p.207). With this new way of conceptualizing a paintbrush, the inventors recognize that while painting, the painter vibrates a brush. This movement corresponds, under the interpretation of the paint brush as a pump, to pumping the paint out. This reconceptualization culminates in a general theory of paintbrushes as pumpoids. According to Schön (1978), this episode of collaborative discovery can be seen to proceed through three stages

1. A similarity is conceived in terms of the flowing of a liquid substance.
2. New details are perceived and described.
3. An explicit account of the similarity is articulated as a general theory of pumpoids.

As it is obvious in this example of collaborative thinking and design, there are certain things in this process which do not match a traditional, description-based explanation, in which the inventors would perceive the situation, match this perception with a model, and create descriptions of this model in natural language for communication. The first of these is that the features of paintbrushes are not available on themselves, without a certain way of seeing in which they would come to the foreground: "for example, spaces between the bristles weren't even seen as *things* to be described until the pump metaphor developed" (Clancey, 1997, p.208). Furthermore, the pump metaphor was not fixed by a number of features before it was conceived in the context of paintbrushes, and "[m]ore fundamentally, a pump is reconceived with respect to this example; it is not a general category that is just instantiated or matched against

the situation" (Clancey, 1997, p.208). It is therefore logical to see the reconceptualization of the paintbrush as a pump as a coupled perceptual-conceptual process, where these two different levels are coordinated in a different manner than before once the similarity between the concrete artefact and the abstract description is discovered:

> In perceiving and saying that one thing, the paintbrush, is an example of something else, a pump, we change both how we view paintbrushes and how we view pumps. Thus, the metaphorical conception is generative of both perceptual features and descriptions. Indeed, the painters' experience indicates that the previously articulated features of these objects are incommensurate at first; all they have is a sense of similarity (in the seeing) and a sense of discord (in the saying). The new way of seeing and the tentative way of talking arose together, but the painters don't yet have a descriptive model to explain this relation. (Clancey, 1997, p.209)

Based on this particular observation regarding the role of metaphors in discovery and the coupling of perceptual and conceptual dynamics, and further examples displaying the codependence of perception, conceptualization and description, Clancey (1997) arrives at the following conclusion regarding how meaning is conceived in a situation:

> Not only is meaning contextually determined, but what constitutes *a situation* to the observer -the context- is itself partially constructed within the interpretation process. The meaning of a representation is not inherent, partially because the *representational form* itself is not inherent. Both the perceptual form of the representations and its meaning can arise *together* – not serial, not parallel-independent, but *coupled and mutually constraining.* (p.203)

Seen in the context of cooperative descriptive modelling, as with the inventors working on the paintbrush, collaborative work on understanding a system or creating a new one consists of modifying the conceptualizations and perceptual categorizations available to the group, arriving at new descriptions which were not possible with the initial perceptual means shared by the individuals. In work done on expert systems, the knowledge and performance possessed by the individuals were stripped from both the perceptual contingencies, the co-development of perception and conceptualization, and the concrete situations in which this co-development occured and were formed into what was for the observer knowledge. Such an abstraction was possible because of the Cartesian view, according to which information must flow only in one direction, from the sensory mechanisms to central command, essentially as in the sense-model-plan-act architecture of traditional systems (see Figure 2.1.1). As an alternative to this approach, Clancey (1997) argues for what he calls the transactional perspective: "Our names for things and what they mean, our theories, and our conceptions *develop in our behavior* as we interact with and reperceive what

we and others have previously said and done [...] the processes of looking, perceiving, understanding and describing are arising together and shaping each other" (p.3).

A result of the transactional perspective is a reconsideration of the relationship between representational entities and sensory data. The standard view is that representations stand in for sensory data. The role of representations is to reduce the complexity of sensory information, serve recognition of objects and situations, unite the variances in the data, and filter noise. In this paradigm, information flows in one direction in an intelligent system, with sensory data getting more compressed as it becomes more abstract and context-independent, finally resulting in symbolic representations. Activity is then the result of further processing of information from these representations to form plans, and delivering these plans to an executive module. In the transactional view, the flow of information is in both directions, with the perception of features in sensory data, categorization, and the use of names for these objects arising in parallel processes which constrain and determine each other mutually. Such a view leads to an alternative vision of what kind of computation is necessary for realizing systems similar to human capabilities. A necessary condition for realizing a transactional system is the coordination of the lower- and higher-level processes in action; that is, higher-level abstractions should not serve as rigid forms into which all sensory information has to be fitted somehow, but as means of variously organizing sensory information.

3.2.1 Psychological evidence on situated cognition

A view outside of the symbol processing paradigm found its way into psychology and neuroscience initially through the study of mental imagery and action, which provided evidence for the involvement of circuitry used in perception and motor action in processes traditionally explained with recourse to abstract symbol processing. In experiments studying the manipulation of three dimensional images, Shepard and Metzler (1971) showed that the time it takes to rotate an imaginary object was proportional to the degree of rotation. This and similar results (Kosslyn et al., 1979) showing congruities between perceptual processes and mental procedures led to a response from the proponents of the symbol processing paradigm, whose main argument was that symbol processing could explain the same findings with minor modifications, while at the same time presenting a more general theory of mental processing (Pylyshyn, 1973).

Through further findings, the status of perceptual and motor processes has been strongly established by now. What Barsalou (2008) calls the study of grounded cognition has become a very active field in psychology and neuropsychology, with various results corroborating the significance of sensory and motor capabilities in various task settings. For example, humans find it more difficult to manipulate string drawings in ways incompatible with their joints (Parsons, 1987). Tucker and Ellis (1998) have set up elaborate experiments in which people had to answer yes/no questions on the spatial orientations of kitchen utensils. Their responses were faster when the presented physical situation supported the

direction in which they had to press a button. In further experiments, Symes et al. (2007) demonstrated the existence of "purely physical affordances", that is affordances "solely revealed by the physical structure or arrangement of the object" (p.239). They showed that whether an object presented in a task is shown in a graspable direction had a significant affect on the response latency of the subjects.

The field of language comprehension has been a stronghold of the symbol processing paradigm and one of the traditional subjects of computational models in this vein. As discussed in the preceding chapter, there is reason to doubt that language comprehension relies solely on context-free symbols, due to the nature of certain kinds of utterances which cannot be disambiguated unless experience with the physical environment is not brought in. Nevertheless, it is possible to argue that these aspects are also represented in symbolic form in a central knowledge register, and are fetched when needed. An example for a case in which such an experience is necessary are the following two sentences, adapted from Zwaan and Madden (2005):

1. Jack looked across the room to see where the whisper/explosion came from.

2. Jack looked across the valley to see where the whisper/explosion came from.

Obviously, the word "whisper" is the suitable choice for the first sentence, and "explosion" is suitable for the second. In order to come to this conclusion, much more has to be known then the definitions of the words whisper, explosion, room and valley.

The question of whether this knowledge is solely of propositional character or whether traces of perceptual and motor experiences are also relevant has been studied in psychological experiments. Zwaan et al. (2002) report an experiment in which subjects read a sentence and then were presented an image, and asked whether the object in the image was mentioned in the sentence or not. The images had the object either in a form fitting to the way the object was mentioned in the sentence (match condition) or not (mismatch condition). The subjects responded faster and more accurately for the match condition, where e.g. the picture of an eagle in flight was presented after the sentence "The ranger saw the eagle in the sky". In the mismatch condition for the same sentence the eagle was presented as standing on a rock. For a review of work in this direction, see Zwaan and Madden (2005). As these studies demonstrate, the tools of traditional psychological research have been successfully adapted to study the situated character of human intelligence. One critical aspect is that the social aspect is only derivatively included in these studies, using scenarios including other people or utterances from third parties.

3.3 Situated Representations

From the perspective of situated cognitive science, it has become clear not only that language has a primary role in the organization of human intelligent activity, but also that the converse is true, namely that embodied human activity and the human organized world have an important role in the grounding of language and its efficient use. The human ability to represent is tightly coupled with language and how language is used in meaningful situations. The situatedness of representations used by humans therefore refers to two aspects that bind together these representations: language is grounded in concrete and meaningful situations in which humans take part, and the representations humans use in all other interactions are in turn rooted in language, and are possible only through language. A weaker version of the second part of this thesis could be formulated as follows: There are certain capabilities which allow human beings to use representations, and these capabilities are employed to an equal extent in both symbolic communication and in the use of private symbols and representations. Therefore, it can be said that humans use situated representations also in thought processes that have traditionally been modelled with context-independent representations. In the following, the term *situated representations* will be used to refer to two things:

- Human representations, which are, as explained above, situated due to the nature of human cognition.
- Representations in computer models that are situated in the same sense.

It is, however, difficult to envisage criteria for calling an artificial construct situated, a status derived from living beings. This difficulty is aggravated by the practice especially in AI research, but also frequent in cognitive science, of giving the name of a supposed human capacity to a computational process or structure, and reading back the same name as proof that the two entities are the same thing; an architectural variant of the inscription error. It will be assumed in the rest of this text that situated representations in an artificial context are simply approximations to the human case, and they are principally different from the human one, with the difference between them being the focal point of research.

As it was discussed in the preceding chapter, it has been traditionally assumed that public symbols used in communication are understandable by the communicators by virtue of their having representations in their head onto which these public symbols can be matched. The situated approach, however, argues for a parallel between communicated symbols, which are public, and the symbols intelligent beings can use in their own activities, which can be categorized as private symbol use. If there is to be internal representation use, what is the role of these symbols? What function do they serve for the agent that can use them, not only in communication but also for itself?

There are a number of proposals in the situated cognition literature which have offered an alternative conception of language worth considering here. One

of these proposals has been made by Clark (2005, 2006, 1998), who offers a comprehensive view that links perception, action and attention with language. The fundamental question Clark tries to answer is what the computational role of language-based symbols is, and what kind of an adaptive value they confer on humans and other animals which can learn to use symbols. Taking his sources from situated cognition, Clark (2006) compares his approach to the cognitivist conception of language as the transcoding of an inner language (discussed in Section 1.2), which he dubs the "Pure Translation" view of language (p.370). As an alternative to this view Clark (1998) defends what he calls the "supra-communicative view of language". According to this view, language is not simply an information stream requiring translation, but a tool which can be used for environmental restructuring by the individual, "a species of external artefact whose current adaptive value is partially constituted by its role in re-shaping the kinds of computational space that our biological brains must negotiate in order to solve certain types of problems, or to carry out certain complex projects" (Clark, 1998, p.162).

Symbols, in the form of public constructs of which linguistic symbols are a subset, are used by humans to create what Clark (2005) calls *surrogate situations*: "[W]e may conceive perceptually encountered or recalled symbols and sentences as acting less like inner data structures, replete with slots and apt for genuine inner combinatoric action, and more like cheap ways of adding task-simplifying and attention-reconfiguring structure to the perceptual scene" (p.240). In this sense, symbolic structures are akin to the tools we use to visualize complex relations or constructs, such as diagrams, sketches, or simple plans. These tools are an alternative to the real situations which they come to represent, in that they relax the temporal restrictions on such real situations, and decrease the information load by simplifying them and bringing to the foreground only those aspects which are relevant; the sketch of a part of the city, on which we want to plan a path, has to include only those things which would help us navigate the area (streets, salient buildings etc.), and not each and every detail in the street scene. In fact, simpler surrogates are better cognitive aids, because they omit the distracting details and help us focus on the necessary aspects. An important feature of such surrogates is that in their use we interact with a situation that we have created and can direct, as compared to the dynamic settings in which we can control only our own reactions. Therefore, in the use of surrogates, we have the chance to shape the dynamics of the situation to our own aims and liking. Furthermore, since we can acquire the use of such surrogates, these can be transmitted from generation to generation, causing a cumulative effect, through which existing surrogates lead to more and better surrogates. Words and sentences, on this view, can be thought of as low-level surrogates which become a part of a situation, that leads to a reconfiguring of the perception of the scene based on attention: "Instead of thinking of linguistic encodings as enabling informational integration by acting as a common format for the outputs of multiple modules, we can think of the whole process as one not of translation into a unifying representation, but of *attention-based coordination*" (p.240).

The ways such scaffolding can be achieved are manifold. One is by using words as filters, in that they are attached as discrete labels to associative concepts. This enables computationally enhancing possible operations with these concepts, and the acquisition of more complex concepts and relationships (Clark, 1998). An example is the case of the chimpanzees which could judge second-order relationships after training with symbolic tokens, mentioned in Section 1.2. This example can be thought in the more general context of "following trajectories in representational space, and leading others reliably through certain trajectories" (Clark, 2006, p.372), a necessary capacity for complex problem-solving skills and socially embedded cognition. Looking at human behavior and intelligence from a dynamical and distributed perspective, it can be speculated that language and symbols in general serve to discipline and stabilize dynamic processes of reason and recall. Words and sentences can thus be used to control others', as well as one's own behavioral coordination. Another capability which the use of linguistic symbols could lead to is second-order cognitive dynamics, i.e. capabilities like self-evaluation, self-criticism etc. This kind of thinking is inherently reliant on the existence of language, and it is safe to bet that other animals do not have such capabilities, due to their lack of linguistic capabilities (Clark, 1998, p.177)

The reasons linguistic constructs are highly suitable for such amplification of cognitive capacities stem from their use in communication, according to Clark (1998). In order to function properly in communication, public language has to be shaped into a suitable means of exchanging ideas; this involves re-formulating, inspecting, and criticizing linguistic statements. Such detailed and repeated use and moulding of language leads to a vocabulary whose words can be used in a minimally context-dependent manner. This is a direct result of the need to keep the vocabulary to a manageable size, which in turn leads to the words being reused in different situations. These are properties which would make a representational tool ideal also for the individual needs:

> By 'freezing' our own thoughts in the memorable, context-resistant and modality-transcending format of a sentence we thus create a special kind of mental object – an object which is apt for scrutiny from multiple different cognitive angles, which is not doomed to alter or change every time we are exposed to new inputs or information, and which fixes the ideas at a fairly high level of abstraction from the idiosyncratic details of their proximal origins in sensory input. Such a mental object is, I suggest, ideally suited to figure in the evaluative, critical and tightly focused operations distinction of second-order cognition (p. 178).

Clancey (1997), in his discussion of the transactional perspective, draws a number of conclusions regarding the role of symbols in human cognition, as a result of his transactional approach. These do not amount to a theory or account of the role of language in human thinking, but offer a similar appraisal of the role of symbols in a computational characterization of the human mental apparatus. In this scheme, conceptualization, which is the basis for symbols and

deliberation, takes on a specific role: "conceptualization is, broadly speaking, a means for an agent to coordinate behavior without being bound to reflexes or the history of conditioned learning" (p. 172). In order to achieve this role, conceptualization has to build on the mechanisms that are used by the agent for sensory-motor coordination: "reasoning involves coordinating ideas in the physical, behavioral sense of holding categorizations active and categorizing their relations [...] [A]rchitectures based on stored descriptions have limited ability to interpret, improvise, analogize, or adapt because the understanding of conceptual models in people is based on and consists of the experience of physical coordination" (p.197). The role of language, especially in the common dialogue situations described above, is then one of bringing together different aspects of an experience, by allowing these conceptualizations to be consciously deliberated, and also discussed with others: "In some sense, the verbalizing process *holds active* disparate experiences originally associated by only a superficial perceiving-as-coupling (of image, sound, gesture, odor, etc.), allowing a more abstract conceptualization to be constructed (a theory of pumpoids)." (p.210).

A theoretical proposal regarding how the process of language comprehension can be understood has been made by Barsalou (1999), who first delineates the established view of language comprehension as the creation of a mental model consisting of amodal representations from a text. This model afterwards serves as an *archival memory* against which conclusions can be drawn or questions can be answered. As the discussions in the previous chapters have shown, methodologies which rely on amodal symbols face numerous theoretical and computational difficulties. The alternative to the amodal representations view proposed by Barsalou (1999) takes language comprehension to be preparation for situated action. Before discussing such a function of language, Barsalou (1999) reminds us of the modern biases we have regarding the function of language. Three such biases are of primary importance; first of all, we focus, in our understanding of language comprehension, on the form it is put to use in modern education; traditionally, teachers stand in front of the class and impart knowledge to them using language. In this picture of knowledge transfer, language serves as a container to pack knowledge, which then gets stored in the heads of the students. The second kind of bias comes from the self-understanding of the scientists themselves, who take pride in their ability to store and recall information acquired through language. The third kind of bias concerns the ease of studying language comprehension in the form of the construction of mental models. It is easy to devise experiments where subjects are presented with texts and then asked questions, whereas it is relatively difficult to construct experiments which break out of this paradigm, and bring together action and textual representation, especially in a laboratory environment.

One way of getting rid of these biases is keeping in mind the evolutionary requirements for language; this is exactly what Barsalou (1999) does, who argues that language evolved not to impart abstract knowledge, but to coordinate the actions of groups, such as hunters. Language comprehension would in such a scenario serve indexing, i.e. relating linguistic constructs to real-world entities. Indexing can take place in various time frames, and with different relations to

the situations in which entities are to be indexed. The situation and the entities can be present, as in immediate indexing, where the conversationalists simultaneously view the physical situation under discussion. In displaced indexing on actual experience, conversationalists discuss an absent situation that they viewed earlier. When the topic of the discussion is an experience that is similar but not identical to what the conversatiolanists went through, displaced indexing on similar experience takes place. Finally, displaced indexing on componential experience refers to individuals discussing a situation which they have not experienced as a whole but whose components have been experienced in previous situations. Displaced indexing does not necessarily have to refer to the past, it can also refer to situations in the future:

> The purpose of communication is not simply to archive information about the world and the events that occur in it, the purpose is to coordinate the actions of multiple agents toward achieving goals in specific situations. Thus the fact that language is often about displaced situations does not pose a problem for indexical and situated views of language. To the contrary, the whole point of such language is typically to optimize situated action later as the envisioned situations become immediate.

Barsalou (1998) proposes a theory for modelling such indexicalizing, which relies on the simulation of perceptual experiences. Independent of the precise mechanism through which such connection of the linguistic symbols to entities in situations takes place, the points argued by Barsalou make clear that the most important issue that has to be tackled by situated theories of linguistic communication is the processing of distant situations, either in the future or the past. Another important contribution of the criticisms and alternative model presented by Barsalou (1999, 1998) is that in his work, the situated, social and cognitive aspects are brought together, regardless of the traditional divisions between the various approaches.

One last theory worth mentioning in this context is that of Tomasello (2003), who, basing on extensive experimental work with infants and apes, argues that human babies are distinguished by their ability to engage in activities involving varying degrees of joint attention, such as following an adult's attention, or directing an adult's attention intentionally. There is a very strong temporal correlation between infants' ability to engage in such joint attentional activities and the emergence of language comprehension and interaction (Carpenter et al., 1998b). Tomasello (2003) argues that language, especially in its early stages, should be understood as an activity of sharing and directing attention. It further results from the understanding the children have of the others as intentional agents which can pay attention to the shared entities in the environment. Construed this way, language becomes an act of sharing attentional focus with agents which are assumed to have similar perspectives as the child itself: "The infant's first production of language is nothing other than her emerging ability to express her own communicative intention that other persons join her in attending to something" (Tomasello, 2003, p.50).

Language, which arises on the basis of such capacities, then takes on a cognitive function due to the properties linguistic symbols and structured representations attain in their function as intersubjective tools. These properties derive from the fact that the speaker has to consider not only his or her own perspective, but the myriad ways the others can conceptualize a given situation:

> The intersubjective and perspectival nature of linguistic symbols thus creates a clear break with straightforward perceptual or sensorimotor cognitive representations. It removes them to a very large extent from the perceptual situation at hand, and in ways much more profound than the fact that they can stand for physically absent objects and events (and other simple forms of spatiotemporal displacement). Rather, the intersubjective and perspectival nature of linguistic symbols actually undermines the whole concept of a perceptual situation, by layering on top of it the multitudinous and multifarious perspectives that are communicatively possible for those of us who share a certain set of linguistic symbols. (p. 53)

The issue of how symbol use and linguistic proficiency arise are also discussed by Tomasello (2000); his views on these topics will be presented in the next chapter in the context of evolution of the language capacity.

Embodied AI and cognitive science, in taking proposals like these regarding the role of language in intelligence and a characterization of language as a communal phenomenon instead of being inherent in the knowledge of the individual, have created multi-agent models to study the use of symbolic representations. In the next chapter, these models will be studied, in the light of the lessons learned from situated interpretations of linguistic communication and symbolic representations.

Chapter 4

The Dynamics of Symbolic Communication

The centrality of language for human intelligence, in particular, its role as the basis of symbolic capacities, have been discussed in Chapter 1, and further elaborated in Chapter 2. It has been argued that language, or symbolic communication as a simpler starting point, has to be studied as a tool used to organize embodied behavior of communities, a position which would give the symbolic means used in communication a dual role. Similar observations have led many in AI, artificial life and related fields to regard the emergence and dynamics of language as an ideal starting point to study higher level intellectual capacities, without building in the representational capacities to be explained in the first place. Research into the origins of linguistic communication deriving from this conviction on the centrality of communication is being carried out using artificial agents with comparatively simple processing mechanisms, but embedded in a perceptual context, and using the communicative capabilities in social situations. In this chapter, an overview of work on the emergence of language and symbolic communication will be given. Before that, however, a short discussion on the current theories of the roots of language and communicative symbol use will be presented for orientation.

4.1 Approaches to the emergence and evolution of language

The evolution of the language capacity is a topic very open to speculation and wild theorizing. Prior to the modern advances in biology, linguistics and computational methods, the topic of the origins of language was partly shunned as too speculative; so much so that the presentation of papers on this topic were forbidden at the *Société de Linguistique de Paris* from 1866 on (Christiansen and Kirby, 2003). Modern discussions on how language appeared as a cogni-

tive ability and how it took on the form and the variability we can observe are informed by the advances especially in linguistics, but many other disciplines make substantive contributions. This multidisciplinary background is connected to the relevance of related questions for many fields. One related question is the cognitive role of language. The relation is due to the simple fact that, as a cognitive capacity, language capacity has to have certain benefits for the individual agent for it to survive either through evolution or generational transmission by learning. The nature of mechanisms underlying the linguistic and symbolic capacity in human cognition can also be added to this combination, because specifications for processing mechanisms include background assumptions about the improvements symbolic capacities lead to, and evolutionary paths these mechanisms might have emerged through.

As it has been made clear in Chapter 1, the focus of the work presented is on the role of primitive symbol use in cognition. Therefore, a detailed review of the discussion on the evolution of the whole capacity is not relevant here. Nevertheless, an overview will be given for orientation purposes, and to demonstrate the connection between the positions on the role of language in cognition, how it evolved, and where the meaning of symbols used with communicative intentions stems from. This overview will be limited to the accounts from linguistics and cognitive science, and will not deal with the issues of the evolution of the vocal tract and other biological & physiological aspects.[1]

There are a number of common difficulties evolutionary accounts of language have to face. One of these is that there is no comparable capacity among animals. Most of the other complex capacities or organs of humans have analogous examples in other animals, which presents the possibility of studying them comparatively. Another important difficulty is that our understanding of the language capacity is still limited, which leads to insufficient information on the different components the complete language capacity necessitates. Another problem with the evolution of communication in general is the survival value of communication for the individual, as compared to its value for the community. A warning signal, for example, has value for the members of a community, because it helps them flee earlier in case of danger. It does not, however, confer any obvious advantage on the animal which makes this call. It might even confer a disadvantage to it, because its location is revealed to any possible predators. The most popular explanation for this discrepancy is to look for a non-communicative behavior which might be a precursor of the communicative one, and base an explanation on the specialization of this behavior (cf. Hauser, 1996).

One last difficulty, especially from a Chomskian point of view, is the amount of knowledge of language that has to be built into the "language organ". As it was explained in Chapter 1, the Chomskian approach to the knowledge of language is to posit an "organ" which converts linguistic input from the environment into a "language module". Since this input is limited in scope, and does

[1]For an extensive review of research on evolution of the language capacity see Hauser (1996). For an overview of research on the evolution of speech see Fitch (2000).

not present the whole variation of the language, the language organ has to possess a considerable amount of knowledge in order to create a complete language module. If an evolutionary approach is to be accepted, this knowledge has to be assumed to appear through evolution. The most significant problem with such a conclusion is that, to put it succinctly, half a language is no language: the set of rules which are necessary for the creation of a language module are dependent on each other, and individual rules cannot function on their own. Chomsky has been known to be against an evolutionary explanation for the language capacity due to these reasons, which have led to a conviction that the language capacity is simply too complex for explanation through evolution. Other authors have extended his position to argue against an adaptivist theory of the evolution of language capacity. In his statement of the problem reaching beyond the linguistic domain, but taking it as a starting point, Piattelli-Palmarini (1989) argues for a nonadaptationist account of the language capacity. In biology, one form of evolutionary explanation alternative to adaptation is called exaptation. Exaptation is the shift in use of a behavioral or anatomical trait from the original function for which this trait evolved to another one; the classic example is the birds' feather, which originally evolved for temperature regulation, but later came to serve flight. Piattelli-Palmarini (1989) argues that the language capacity can be explained only through exaptation. The main argument for this conclusion is that the study of innate grammar has disclosed "many instances of specifity and gratuity in the design of all natural human languages, but hardly any instance of traits dictated by generic communicative efficiency, or constraints dictated by the laws of pure logic" (Piattelli-Palmarini, 1989, p.22).

There are nevertheless approaches to the evolution of the language capacity which agree with the innativist position on the existence of innate language capacities, but also argue for an evolutionary basis for these capacities. The best-known of these approaches is that of Pinker and Bloom (1990), who argue for an adaptationist account for the evolution of language capacity. Their argument relies on two main points. The first is the role of natural selection and adaptation in the emergence of complex, apparently engineered systems. Pinker and Bloom (1990) argue that even in the case of exaptation, the trait which serves a new function can evolve to specialize further for the new function; exaptation is just one more mechanism in the service of adaptation and not a replacement for it. Exaptation cannot lead to complex, adapted systems on itself; such systems, such as the eye, can arise only as a result of long and gradual gradient ascent. The second point is the status of language as a complex capacity which has been evolved for the communication of propositional structures over a serial channel. Against the central argument of Piattelli-Palmarini (1989), the authors claim that the innate grammar is primarily a constraint satisfaction device which has to meet certain tradeoffs, and the arbitrariness which one sees in language is a result of this set of tradeoffs.

Interestingly, Chomsky has also reversed his position on this issue and recently argued that language might have evolved through selection. Hauser, Chomsky & Fitch (2002) propose to think of the language capacity as consist-

ing of two components: "faculty of language - broad" (FLB) and "faculty of language - narrow" (FLN). FLB is a combination of an internal computation system (the FLN) and two more organism-internal systems, the sensory-motor and conceptual-intentional systems. FLN, on the other hand, is the abstract linguistic computational system alone, and one of its core properties is recursion, which enables it to take a finite set of elements and yield an infinite array of discrete expressions. The central claim of Hauser et al. (2002) is that the FLN is the only uniquely human component of the language capacity, and an evolutionary explanation for it is the grand scientific problem faced by adaptationists. The FLB, however, is built on similar capacities in other animals; many capacities that are part of the FLB, except for the ones which belong also in the FLN, are based on mechanisms shared with nonhuman animals. The scientific program proposed by Hauser et al. (2002) aims to understand the evolution of both FLB and FLN through comparative studies which delineate these two modules and point at possible central mechanisms which might have emerged through mutations. As an example contribution to such a program, they present evidence from anthropology and comparative psychology to support their claim that the FLB has counterparts among animals. In a direct response to the position presented by Hauser et al. (2002), Pinker and Jackendoff (2005) argue for specializations at all levels in the human cognitive apparatus concerning the language capacity. Their conclusion is that the distinction between the two modules FLN and FLB is not tenable, not only due to the specializations in different parts specific to humans and honed to language production, but also due to the difficulty of specifying a minimal core of language.

Another attempt at a theory of the evolutionary appearance of the language capacity is that of Derek Bickerton. Bickerton (1990) builds on his earlier research into pidgins, simple languages lacking complex grammatical structure and including sentences only a few words long, which arise when two communities without a common language have to communicate. When a pidgin is acquired by children, a refinement in structure and usage is observed, whereby the language becomes a creole. Bickerton (1990) argues that the commonalities between pidgin languages, the language of two-year-old children, and signing chimpanzees points to a primitive form of communication which preceded the full-blown human languages in the evolutionary time line; he calls such a language a "protolanguage". Through the stipulation of such a step between human language and the nonexistence of communication, the task of the development of language is simplified; nevertheless, the step is a considerable one. This step is supposed to have happened through a single macromutation. The requirement for such a big step is the most criticized aspect of the theory of Bickerton (1990) (see e.g. Hauser, 1996). As a response to these criticisms, Bickerton (2003) modified his approach, and argued for the separate evolution of three capacities relevant to language: modality (i.e. speech), symbol use and structure. Whereas symbol use is a cultural phenomenon, modality and structure are results of evolution. Structure, i.e. syntax, has also evolved, but not with a single macromutation, as Bickerton earlier proposed. Instead, syntax is theorized to be a result of a number of simple mechanisms which evolved to serve other

functions, but were then utilized to make communication more successful.[2]

All these approaches have one thing in common: they are based on the Chomskian view that the language capacity is unique, and is served by cognitive machinery that is purpose-specific and has a syntactic core. The Chomskian view of language, as explained in Section 1.2, has laid the foundation of early cognitive science, and is the most widely discussed and cited paradigm for understanding language in the cognitive scientific circles. Such an overwhelming presence can create the impression that the innativist/generative approach is without scientific opposition and alternatives, but this impression is not entirely true, in that there are still many researchers who argue for the acquisition and processing of language through more general-purpose learning mechanisms. One of these people is (Tomasello, 2000), whose approach to the role of language and socially shared symbols was discussed in the previous chapter. In his review of Pinker (1994), one of the most influential and widely-read works arguing for innativist and generative approaches to language, Tomasello (1995) arrays a number of criticisms against what he argues to be the two main features of this approach. The first of these features is the syntactic nature of the stipulated structures responsible for language production. Such syntactic structures do not depend on meaning and their formulation is mainly driven by considerations of mathematical elegance. The second feature is that "all of these universals are described in linguistically specific terms such that it is very difficult to relate them to cognition in other psychological domains [...] Thus, in this view, noun and verb have nothing to do with concepts of object and action, but are defined solely in terms of their syntactic distributions" (p. 136).

One of the central tenets of the innativist position is, as the name implies, the availability of innate syntactical capacities, and the crucial role of these capacities in the acquisition of language. It is assumed that the acquisition of language is to be traced back to the availability of rule-based grammar to children, which is essentially equivalent to the grammar possessed by the adults in terms of its categories, such as subject or verb. Tomasello (2000) presents copious experimental evidence which undermines this assumption, called the continuity assumption (Pinker, 1984). Proof for the continuity assumption usually takes the form of negative arguments, as in the poverty of the stimulus argument: The linguistic input received by the child during the acquisition phase is too limited, and it is impossible for a child to arrive, without innate knowledge, at the grown-up language capacity, on the basis of this input. The essence of this kind of argument for innate grammar knowledge is summarized by Tomasello (2000) as "You can't get there from here" (p. 235). The evidence presented by Tomasello (2000) contradicting the continuity assumption is based on observational and experimental studies. In the observational studies, young children of up to 3 years of age were followed in their daily life and their utterances were recorded. An analysis of data collected this way across a number of languages has shown two things. First of all, there is no evidence of abstract syntactic

[2]See Jackendoff (1999) for a preliminary attempt at listing possible steps in the evolution of syntactic capacity in incremental steps as proposed by Bickerton (2003).

categories in the linguistic behavior of children of this age; for example, when young children learn a verb, they rarely ever use it in a different tense or person from what they initially memorize, and their productions revolve around concrete items and structures. Second, each of these concrete structures undergoes its own development; when a child learns "you kiss" in addition to "I kiss", this new structure is not carried over to "you think" as an extension of "I think". This pattern persists until approximately the third year of life.

Experimental evidence supports such observations even to a further extent. For example, in experiments in which artificial words were used as tracers, it was found out that young children did not generalize intransitive application of verbs they hear from adults ("This is called meeking") to equivalent transitive forms ("She's meeking the car") until three years of age. They also treated nouns and verbs differently, freely combining nouns, but hardly combining the verbs in different ways than they heard (Tomasello et al., 1997). Traditionally, such results get explained away with claims of performance limitations; however, in these controlled experiments, it was shown that the children who could not produce transitive forms of the verbs they heard could use new nouns productively (ruling out reluctance to use new words), performed conservatively when tested for comprehension of transitive forms of the same verbs (ruling out production factors), and used transitive forms if that was the form in which they first heard the verbs get used by adults. Tomasello (2000) argues that these findings, coupled with the absence of positive experimental support for the continuity assumption, undermine the idea that children possess a complete syntactic competence. According to Tomasello (2000), the specific linguistic development path of children growing up in English-speaking environments can be explained in terms of the Verb Island Hypothesis, which claims that "children's early language is organized and structured totally around individual verbs and other predicative terms [...] the 2-year-old child's syntactic competence is comprised totally of verb-specific constructions with open nominal slots" (p.214).

The alternative proposed by Tomasello (2000) to the innativist position is based on cognitive-functional linguistics (see also Tomasello, 1998). This approach to linguistic theory argues that the acquisition of language is achieved through the use of general-purpose cognitive mechanisms, and not an innate language organ exclusively for language. The capacities of the adult, which are a result of such an acquisition process, are also theorized to be of a different character than the set of grammatical rules and transformational capabilities assumed by the generative approach. Instead of an elegant mathematical core, mastery of various linguistic symbols and constructional schemas is believed to lead to adult linguistic competence. Highly canonical aspects result from the hierarchical organization of these schemas, but otherwise, there is no distinction between the core of linguistic competence and peripheries, or a mathematically complete competence and performance which deviates from this completeness:

> A plausible way to think of mature linguistic competence, then, is as a structured inventory of constructions, some of which are similar to many others and so reside in a more core-like center, and others

> of which connect to very few other constructions (and in different ways) and so reside more towards the periphery. The proposal would thus be that the child initially learns individual, item-based linguistic constructions (e.g. verb island constructions), and if there are patterns to be discerned among these different item-based constructions in adult usage, she could then make abstractions and create inheritance hierarchies of constructions. (Tomasello, 2000)

The ability to operate with such symbols and schemas is derived from general cognitive abilities which might take on special characteristics in the domain of linguistic communication. The verb islands discussed above are a developmentally early example of linguistic schemas. As Tomasello (2000) points out, the early generativist arguments for the impossibility of learning language without special-purpose mechanisms are directed against outdated learning concepts from the 1950s, like simple associative learning. Modern cognitive-functional linguistics, on the other hand, can point to much more complicated general-purpose mechanisms, such as analogy making and structural mapping.

As it was argued at the beginning of this chapter, theories on the evolution of the language capacity are usually extensions of the theories on the processing of language. In the case of Tomasello, since he believes that the language capacity is a result of various general-purpose cognitive mechanisms, it is to be expected that he argues for continuity among the animals (as he does in Tomasello, 1995). Despite such continuity, there should be a crucial difference which sets humans apart from other animals, the "small difference that made a big difference" (Tomasello et al., 2005, p.690). According to Tomasello (2000), this difference lies in what he calls *cultural learning*, which is a kind of imitation learning. One sort of imitation learning, what Tomasello (2000) calls mimicking, involves solely the repetition of what the adult says or does, with little or no understanding of the particular aim behind the action or utterance. Cultural learning, on the other hand, requires an understanding of the purpose or function of the behavior to be imitated. Experimental evidence for such a capacity in young children includes 18-month-old infants reproducing intentional action they saw an adult attempting to perform, even if that action was not carried through to completion (Meltzoff, 1995), or 16-month-old infants reproducing the intentional, goal-directed actions of an adult, but not the accidental ones (Carpenter et al., 1998a). One other well-known example of the imitation of intentional action was studied in an experiment by Gergely et al. (2002), where 14-month-old infants were shown a man touching his head to the top of a box to turn on a light. In half of the cases the adult had his hands occupied, so that he could not touch the box with them, whereas in the other half they were free. When it was their turn, the infants who saw the hands-free demonstration bent over to turn on the light with their heads more often than the others. These children apparently believed that if the adult touched the box with his head although his hands were free, there must have been a reason for it. The relevance of this capacity to imitate intentional action for language acquisition derives from the ability of children to recognize, when an utterance such as

"Look! A clown!" is made by an adult, his intention to point her attention to a particular object. Such an intention is called a communicative intention by Tomasello (2000), who further argues that "understanding a communicative intention means understanding precisely how another person intends to manipulate your attention" (p.238). The acquisition of language takes place when the child, through the learning mechanisms mentioned above, matches linguistic structures to the roles they play in the communicative intentions of adults. According to Tomasello (2000), the capacity of young children to learn imitatively is "the initial ontogenetic expression of the human organism's biological adaptation for culture" (p. 240).[3]

Since the publication of Tomasello (2000), further experiments with primates led Tomasello to change his position and propose a different capability as the distinguishing mark of humans. An example for the experiments which provided new evidence is one carried out by Call et al. (2004), in which a human gave an ape food through a hole in a plexiglass wall, but sometimes brought out a piece of food and either refused to give it to the apes or attempted to give it to the ape but was unsuccessful. The apes gestured more and left the area earlier when the human was unwilling than when he was unable, in which case they sometimes waited patiently. This and similar experiments show that "apes understand that other have goals and behave toward them persistently, and that this is governed by what they perceive" (Tomasello et al., 2005, p.685). This is exactly what was supposed to set human beings apart according to Tomasello (2000); therefore, something else must be the small difference that makes a big difference. Tomasello et al. (2005) proposes this difference to be the capacity for shared intentionality. According to Tomasello et al. (2005), children go through three stages of social development in their first year of life:

1. Understanding animate action as goal-directed

2. Understanding the pursuit of goals by actors, and the fact that human beings monitor their actions so that they can recognize when they have succeeded, and keep on trying until they succeed

3. Understanding how human beings evaluate a situation to decide on actions to achieve their goals, and choose plans to come closer to these goals

As it was mentioned above, these capabilities pave the way for cultural learning, i.e. reading adults' intentions interpreting their actions in terms of these intentions. What sets humans apart is the capability of shared intentionality, which follows these steps and builds upon them. Shared intentionality is different from earlier forms of individual intentionality in that it involves participating in collaborative activities including shared goals and socially coordinated action plans. The agents have to coordinate not only their own goals, but also the common goals shared with the others, and the states of the others and how their attention has to be manipulated in order to align them with the common

[3]See Tomasello (1999) for an extended discussion of this view point.

goal and the plan devised to achieve this goal: "Overall, then, collaborative activities require both an alignment of self with other in order to form the shared goal, and also a differentiation of self from other in order to understand and coordinate the differing but complementary roles in the joint intention" (p.681). Infants start engaging in such collaborative activities at around 14 months of age. An example is presented by Ross and Lollis (1987), in whose experiments children of this age played games together with adults. When the adults stopped participating, the children prompted them to re-engage, and sometimes carried out their duties for them. Language is one such form of collaborative activity, in the sense that the symbols have both the aspect of the speaker and the listener, so that the child has to learn to play both roles. More significant is the further fact that the joint goal in a communication context is to reorient the listener's intentions and attention so that they align with those of the speaker.

As it can be seen clearly from this short overview, according to Tomasello (2000, 2005), linguistic capabilities are a result of the social and cultural abilities of humans, and not a detached cognitive module. Therefore, in order to arrive at an evolutionary understanding of the language capacity, we should first understand how these abilities have arisen:

> [S]aying that only humans have language is like saying that only humans build skyscrapers, when the fact is that only humans (among primates) build freestanding shelters at all. Language is not basic; it is derived. It rests on the same underlying cognitive and social skills that lead infants to point to things and show things to other people declaratively and informatively, in a way that other primates do not do, and that lead them to engage in collaborative and joint attentional activities with others of a kind that are also unique among primates (Tomasello et al., 2005, p.690).

The methodological repercussion of the position presented by Tomasello, and other positions based on general-purpose mechanisms for language acquisition and processing in general, would be concentrating on alternative learning mechanisms (such as analogy making and structural mapping), language use and acquisition in shared contexts, and social interaction in common tasks. From a synthetic perspective, such a research program would present two difficulties. The first of these is the significantly different learning mechanism which has to be built. Research into analogies and structural mapping is relatively young, and there are still unresolved problems, especially if a complete agent perspective from perception to action is taken (as argued in Section 2.1.2). The second difficulty involves modelling of cooperative activities. Robotic modelling of cooperation is still in its infancy, and the use of shared representations has an understanding of cooperation as its precondition. Once there is sufficient progress in these two areas, however, the direction pointed out by Tomasello presents a viable alternative to the innativist position that is taken to be the default one on language acquisition in many areas.

4.2 Computational models of the dynamics of communication

As it was explained in the previous chapter, embodied AI shares an understanding of the concept of situatedness with the philosophical research into the embodied nature of human intelligence. Nevertheless, there is still a deep rift between the conceptual apparatus of the two areas. The main reason for this rift is the inadequacy of AI models to create representations without falling into the pit of creating a disembodied inner arena of symbols, and the inadequacy of theoretical studies to deliver an embodied model of symbol use and linguistic behavior. Language evolution is at the middle of this rift, trying to reconcile the embodied intelligence of embodied AI models with the situated descriptions of human intelligence and linguistic abilities. However, it is at the same time vulnerable to the problems that have plagued theories that try to give an account of symbol use. These problems fall into the two general categories:

- A radical empiricism which posits the non-existence of any universal categories or rules, and
- a radical innativism, which relies on a set of innate and immutable rules and recognition mechanisms for feeding those rules[4]

Linguistic behavior has many dynamics which can be studied. Due to the complicated nature of linguistic behavior, it is possible to focus on a subset of these dynamics, and study their interaction and how they affect certain further parameters. In the plethora of recent studies on the origins of language and symbolic communication, different subsets of these dynamics have been selected for various purposes. Among these dynamics, the following are the most important:

- Multi-agent dynamics of referential communication
- Agent-internal dynamics of organizing categories and labels, and learning
- Generational transmission of linguistic knowledge
- Genetic transmission of successful traits
- Embodied and situated dynamics

Although the umbrella term under which research on the computational modelling of communicative behavior is usually grouped is “evolution of language” or “language evolution”, in many cases, the research does not have much to do with artificial evolution but, as it was argued, with the effect of one or more of the above cited dynamics on communicative behavior. It might therefore be appropriate to talk of research into the dynamics of communication,

[4]Smith (1996) calls this difficulty “Scylla of naive realism and the Charybdis of pure constructivism” (p.3).

especially in the context of the work presented here. In their comprehensive review of the work done until 2003 in this field, Wagner et al. (2003) organize the various efforts along two dimensions: whether the agents are situated or not, and whether the tokens used in communication are structured or not. In this classification, the agents' situatedness depends on whether the simulation runs in an environment in which they can interact with other objects and they have outputs which affect the environment, leading to communication about these objects and their manipulation. As Wagner et al. (2003) argue, there is a general progress from nonsituated agents that are simple encoding and decoding units to more situated ones. The nonsituated, encoding/decoding agents match communicative symbols to internal representations as an implementation of the process of interpretation. As research in this area progresses and the agents get more situated, a move away from encoder/decoder games can be seen, towards a more situated view of symbol use. This progress can be seen as a move away from the internalist picture of meanings of symbols as mental representations, towards a more use-based theory of meaning. Meaning here refers generally to the process or entity by virtue of which a public symbol (i.e. one used by many agents at the same time) can be used in communication and in individual activity in a correct manner. The main reason for the coupling between the situated models and meaning-as-use is that, in the case of unsituated models, there is nothing to which the symbols can refer, thereby necessitating an agent-internal correspondent which serves as the meaning to be transmitted. The next logical step in this progress from unsituated models to situated ones is placing the agents in a task setting, thereby enabling the use of symbols to reach a goal; this step would enable further explorations of use-based theories of meaning. In this section, the progression of work in the area of dynamics of communication from encoding/decoding agents to situated ones will be discussed on representative examples from the literature.

4.2.1 Multi-agent dynamics

In an early example of a synthetic study of the dynamics of communication, Hurford (1989) studies the role of the Saussurean sign in the emergence of linguistic communication. The Saussurean sign is a "bidirectional mapping between a phonological form and some representation of a concept" (p.187). Hurford (1989) takes this to be a fundamental component of the Language Acquisition Device proposed by Chomsky, and aims to present a multi-agent model proving the evolutionary viability of the Saussurean sign. This sign is assumed to correspond to a mental representation in the head. The background of the model is sociobiology, which is the study, through quantitative means, of the advantages of innate dispositions, called strategies, in social evolutionary settings. Instead of using genetic algorithms, Hurford (1989) pits artificial agents with different combinations of received signal to concept and concept to uttered signal. The Saussurean agents utter the same signal they hear for the same concept, whereas imitators imitate the transmission and reception coupling of the other agents, and calculators base their transmission behavior on the reception behavior of

the others, and vice versa. The individuals most successful in their communication with the other agents are picked as parents for the next generation of agents; that is, successful communication is assumed to increase the fitness of the individuals. However, no genetic methods, such as random mutations or crossover are used. The main problems with this work is that the agents are not engaged in any kind of practical or embodied activity in an environment, so neither the inner representations nor the uttered signs are situated. Furthermore, it is taken for granted that the agents already have mental representations with which the signals can be coupled; these representations serve as the "meaning" of any symbols used in communication. These simplistic assumptions hinder deeper insights into the role of symbolic communication for societies and the embodied grounding of symbols.

A pioneering study in the emergence of a lexicon is that of Hutchins and Hazlehurst (1995). Pointing out the preoccupation in cognitive science and related fields with individual minds and what goes on inside them, they state their aim as "to put the social and the cognitive in equal theoretical footing by taking a *community of minds* as our unit of analysis". They also commit to the view of language as a shared resource, instead of a mechanism possessed by individual agents, and symbols as "shared denotational resources" which arise in the interaction among the members of a community. Their model is based on a number of theoretical assumptions from work on distributed cognition reported by Hutchins (1996). The most important among these are the following:

- *The cultural grounding of intelligence assumption:* No social mind can become appropriately organized except via interaction with the products of the organization of other minds, and the shared physical environment.
- *The shallow symbols assumption:* The nature of mental representations cannot simply be assumed, they must be explained.
- *The no developmental magic assumption:* The processes that account for the normal operation of the cognitive system should also account for its development through time.

The computational model consists of agents embodying auto-associator networks. The inputs to these networks consist of images in the form of 6×6 pixel grids. The hidden units of the networks act as feature encoders when they are trained to replicate the input pattern, and these hidden units are then made "public" to produce a term in a lexicon. This involves taking the output of these units as utterances. The agent which acts as the listener in a communication round than carries out back propagation not only on the output of the network, but also on the values of these hidden/public units. Through episodes of such conversations, agents arrive at a shared lexicon, which is in practice "a consensus on a set of distinctions".

The recognition of the social aspect of cognition and the stress on research on the social and public nature of symbols is refreshing, compared to the either mentalist or completely rejecting attitude towards representations from the time

of Hutchins and Hazlehurst (1995). This work is all the more interesting due to its cognitive orientation, stressing the possibilities of computational theorizing by taking the community of agents and not the individual as the unit of explanation. On the other hand, activity-as-categorization and the simplicity of the input stimuli are the limiting factors of this model. The symbols neither refer to entities embedded in a common environment, nor are they used in a task context.

4.2.1.1 Steels' language games

Among the researchers in the field of dynamics of communication, Steels is one of the most prolific and probably the best-known. He carried out a number of intertwined studies to model various aspects of symbolic and linguistic communication. His starting point is a view of language as a self-organizing dynamic system, which is then studied in the interaction of a population of agents which engage in language games. These games vary in their communicative extent, how the referent of a word is revealed to the agents, what the units of communication are, and the motor and perceptual apparatus of the agents. Also, the studied phenomena vary considerably, from the emergence of a lexicon to simple grammatic processes.

Steels (1997b) pits three different views of language development against each other: the Chomskian view of linguistic competence through possession of the language organ, language as an adaptive dynamic system, and genetic assimilation, which is a synthesis of the first two positions. He then goes on to make a very convincing case for the dynamic systems approach to modeling communication. The status of linguistic knowledge, traditionally considered, is in the heads of the agents, in the form of internalized rules and word-concept mappings. In the work of Steels, linguistic knowledge is treated as being inherent in the communicative conventions of a community, instead of being encoded in genes and organs in the brain. This knowledge is preserved through learning, a process which also leads to language change, as in the *no developmental magic assumption* of Hutchins and Hazlehurst (1995). Also due to the plethora of selectionist criteria on the communicative dynamics (such as maximizing communicative success, minimizing cognitive processing and memory load, compatibility with the limitations of the sensori-motor apparatus), the various idiosyncrasies of a communicative system at any point in time can be analyzed as a result of the playing out of these factors. Steels (1997b) also argues that the position taken on the origins of symbolic communication has an impact on the nature of linguistic theory:

> The genetic approach championed by the Chomsky school rejects functionalism and sees the language faculty as a formal mechanism whose nature is largely arbitrary. The self-organization/adaptation approach is based on the view that language is a device for communication and representation and therefore sees the specific form of language as the result of balancing physiological and functional

> constraints. Whereas the genetic approach resonates therefore the strongest with the generative, formalistic tradition in linguistics, the adaptive approach is more in line with cognitive and functional grammar.

Another important aspect of Steels' work is the rejection of the mentalist separation of meanings (in the head) and utterances (out in the world). According to Steels (1998), these two things are interdependent: "Language and meaning coevolve. Language is not a mere complex system of labels for concepts and conceptual structures which already exist prior to language, but, rather, the complexification of language contributes to the ability to form richer conceptualizations which then in turn cause language itself to become more complex" (p.385). This co-dependence of meaning creation and linguistic competence is parallel to the view of multi-agent communication serving as the basis of advanced capacities for intelligence: "We assume that language has played a key role in the formation of a symbolic layer in human intelligence and therefore focus on experiments in which the origin of language could take place" (Steels, 1996b). An important precondition for such a foundational role for linguistic symbols is that the language games are a means for creating a grounded system of symbols. The symbols used in the language games should refer to concepts grounded in the sensory-motor interactions of the agents: "Agents start with no prior designer-supplied ontology nor lexicon. A shared ontology and lexicon must emerge from scratch in a self-organized process" (Steels and Kaplan, 1999b).

The general structure of a language game setup as used by Steels in his models involves agents with certain capacities taking part in interactions with each other. Generally, the flow of this interaction is as follows (Steels, 2003): Two agents are selected randomly from a population. They engage in a language game encompassing either objects in their common situation or a part of the common language. Each agent has a certain cognitive architecture and memory mechanisms, which are used to store lexicons or grammars. The outcome of the game is signaled to the agents, or they somehow sense it. This outcome is then used to update the memory structures of the agents, so that the further success of the game increases the more it is played. Therefore, there are four necessary components for specifying a language game:

- An interaction protocol specifying the flow of interaction and the channels used for communication
- A definition of the architecture of the agents
- An environment in which the agents perceive and interact with the objects
- A measure for determining the success of the language games

In the following, first, the earliest examples of the language game setup will be explained to demonstrate the basic idea of the setup. Afterwards, examples of the subjects to which this setup has been applied will be given.

The earliest work done by Steels is in terms of what he calls "the adaptive constructive approach to lexicon formation and meaning creation" (Steels, 1997b)[5]. Before the agents can engage in communicative interactions, they need capabilities to conceptualize their environment, so that they can attach labels to these concepts. This conceptualization capability is acquired through agents engaging in a preliminary *discrimination game*, which involves agents learning to discriminate different objects in their environment (Steels, 1997a). These objects are defined by their features on various channels. Agents build trees which divide these continuous channels into subdomains and thus map them onto categories. Combinations of these trees are then selected in order to pick an object out of a set. This set is defined as the context. In case the existing set of trees is not sufficient to pick the designated object out of the context, one of the channels is re-partitioned, i.e. a new category is created, to enable the discrimination of the object. This way, agents develop all the necessary discriminations necessary to distinguish between a set of objects[6].

Once the agents can discriminate the objects through trees defined on perceptual channels, they engage in various kinds of *language games* (Steels, 1996a). There are two kinds of language games: the naming game and the guessing game. In the naming game, two agents are randomly picked. One of these agents acts as the speaker, and the other acts as the hearer. The speaker picks an object from the environment, and "points" to it (i.e. it is made available also to the other agent), so that the other agent also knows which topic has been selected. Afterwards, the speaker looks for a word for the picked topic by matching the concept for that object to a word in the lexicon. If such a word does not exist, the speaker might create a new word (with a certain probability) and transmit it to the other agent; otherwise the game fails. If the speaker has a word or creates one to match the feature set distinguishing the topic, "the hearer decodes the resulting expression using his lexicon and the game succeeds if the distinctive feature set decoded by the hearer matches with the expected distinctive feature set" (Steels (1998), p.390; see also Steels et al. (2002) for further details). In the second kind of language game, the guessing game, only verbal communication takes place. Two agents try to agree on referring to a topic with the same label, but the perceptual means of how to refer to this topic is decided individually. The guessing game relies on a certain vocabulary already being available (Steels, 1999), and in the absence of a reliable mechanism of pointing, language formation does not take place (Steels and Kaplan, 1998).

The language games setup has been applied in a number of different experimental environments with various kinds of objects and sensory-motor capabilities. The experimental environment which has been the most thoroughly studied is the "talking heads" scenario, which involved perceptually situated agents engaging in language games (Steels et al., 2002). In the talking heads experiments, the agents are pan-tilt cameras with two degrees of freedom, con-

[5]Steels mentions this definition particularly in the context of the work of Hutchins and Hazlehurst (1995)

[6]See Steels (1996c) for technical details of how these channels are partitioned and new categories are created.

trolled by computers. The environment consists of whiteboards with various geometrical figures pasted on them. The agents first engage in discrimination games, as explained above, creating branches of the various sensory channels, such as horizontal and vertical location or color, in order to discriminate these figures from each other. These concepts serve as the meanings which have to be communicated to the other agents in language games. After the generation of these concepts, language games take place. One agent picks a topic and either finds or creates a label which matches this topic. This label is passed on to another agent, which tries to guess which topic was picked. In case this language game fails, the topic is pointed out by the speaker to the listener, which updates its internal structures.

In the course of the language games, incoherence in the vocabulary of agents can appear due to a number of reasons. The two most important reasons for incoherence are synonymy (two words having the same meaning) and polysemy (one word having multiple similar meanings). Synonymy is resolved through weighted connections between words and meanings. Meaning here refers to the category created by partitioning the sensory channels, and then used to pick out different objects. When a language game succeeds, the weight of the connection between the word used and the meaning communicated is strengthened, while the weights between the other words and the meaning are decreased, both by the speaker and the listener. In case a game fails, the weight of the aforementioned connection is decreased. With the use of this weighting mechanism, words that serve successful communication are used progressively more frequently, and one meaning is communicated with only one word after a certain number of games (Steels and Kaplan, 1999a). Polysemy occurs when two different meanings can be used to pick the same object; for example, a red object on the left of the board can be distinguished using either its vertical position or the value of the color channel for this object. Agents can use the same word to refer to such different meanings, and the language games will succeed as long as these meanings serve to distinguish the same object. Polysemy gets resolved when a situation enforces disambiguation, that is, when these two meanings are not sufficient to distinguish an object (Steels and Kaplan, 1999b). In the example with the red object, this would correspond to the existence of two red objects in the scene, or of another object with the same vertical position. Another effect that was observed is the preference for words and meanings which were stable in the sense that they did not rely on subtle differences, such as small differences in color caused by lighting conditions. Instead, the meanings and corresponding words which prevailed relied on distinctions that were robust across different environments (Steels et al., 2002).

Through the use of these mechanisms, the emergence of a robust vocabulary could be observed in a typical experiment. After 2000 agents created around 800 words, the vocabulary was reduced to 100 words, with 8 core words referring to color and direction. The communities were also robust, in that they could create new words when new objects were introduced. Furthermore, new agents could also be introduced into a community, where they learned the already existing vocabulary (Steels et al., 2002).

An early robotic language game involved robotic agents in an area populated by various objects, such as charging stations and other agents (Steels and Vogt, 1997). The agents roam the area, and occasionally switch from exploration mode to speaker mode. Once they are in speaker mode, they search for a possible hearer. A naming game very similar to the above explained scheme then takes place. The agents identify the possible topics in the environment by turning around 360 degrees. The objects are picked by identifying the points where the various sensors (modulated light, visible light and infrared sensors) cross each other; this is possible due to the simple reason that sensors are placed symmetrically on the left and right. The speaker adds itself as a possible topic to the objects thus identified. Afterwards, the speaker picks one topic from this set and "points" at this topic by turning towards it. The hearer can identify the topic by sensing how much the speaker has turned; this is achieved through the use of infrared ray emitters on each robot. With the fixing of the topic for the naming game, the agents try to come up with sets of distinctive features to distinguish the topic. As in the talking heads experiments, the distinctions are achieved by splitting the sensory channels into regions. In addition to the production of new distinctions in order to categorize objects in different sensory contexts, a selectionist procedure removes those categories (i.e. sets of distinctions) which have turned out not to be useful. The rest of the naming game proceeds as explained above, with the similar final result of a shared vocabulary.

Another robotic application of the language game framework is presented in Steels and Kaplan (2001). The robotic agent in this case is the AIBO robot from Sony, and the task is communicating names for physical objects. Since the experiment involved a robot, a mechanism for fixing attention was necessary. This was established by the experimenter showing an object to the robot (holding it in front of its camera) and uttering a word (in this case simply "Look") which caused the robot to fixate on the object. Words from the instructor are processed with a speech-to-text mechanism. The game played is called the classification game; the difference from the guessing game is the availability of a single object. After the attention of the robot is fixed on the object, the speaker (a human) utters the name of the object. In case the agent has already acquired the names of the objects, a variant of the game, where the speaker asks the name of the object, is played. The main difference of this model from the others is the machine learning method employed. Instead of the partitioning of sensory channels, an instance-based algorithm which functions with complete images from the camera is used. The images are transformed into histograms, which are then compared to each other when a categorization is to be made. Words are matched to individual images, and the image with the highest similarity to the image to be classified is considered the winner. The word attached to the winning image then serves as the category label. Steels and Kaplan (2001) point out that this learning algorithm has two advantages. The first advantage is that it supports incremental learning, without a clear distinction between a learning and a testing phase. The second advantage is the speed of learning, without initial erratic behavior. With this alternative learning mechanism, the

robotic agent was able to reach a correct discrimination rate of approximately 80%.

According to Steels (2000), there are three common dynamics at work in these and similar language game experiments.[7] The first of these is reinforcement learning, where the agents receive feedback on the success of their interactions, i.e. whether the categorization they attempted succeeded or not. Reinforcement on itself is not enough to explain how a population arrives at common symbols; for this, self-organization at the system level is necessary. Self-organization is achieved by the agents' use of the successful symbols more often, thereby leading to coherence. The third and final dynamic, which is responsible for the preference of certain symbols, is selectionism. In the case of the talking heads experiment, the categories that are selected, along with the symbols which are initially attached to these, are those that can make robust distinctions in different environments.[8] These different mechanisms are responsible for a population's settling on common communicative tools, be they sound systems, labels or, as it will be explained shortly, grammatic forms.

When it comes to the coordination of the public carriers of meaning (labels, symbols etc.), the language games setup enables focusing on the role of social interaction in constraining the classifications made which can then serve as the "meanings" of the labels used in communication. Steels (2008) argues that the meaning of a communicative symbol is a category deduced by an agent through a categorization mechanism from sensory data; the existence of such a mechanism is also his standard for calling a symbol "grounded":

> In some cases, there is a method that constrains the use of a symbol for the objects with which it is associated. The method could, for example, be a classifier – a perceptual/pattern recognition process that operates over sensorimotor data to decide whether the object 'fits' with the concept. If such an effective method is available, then we call the symbol grounded (p. 223).

The symbols used in communication refer to objects in the environment, thereby making them situated. Nevertheless, this concept-object connection, established by way of agent-internal concepts, also constitutes the limit of their situatedness: Since the symbols are not situated in a task context, the interpretation of a communicative symbol consists only of establishing a connection between a symbol and a perceptual category. The perception of and referring to the different objects is the only aim of the interaction of the agents, because the language games constitute the sole task in which communication takes place. This limitation makes the language games setup, as applied by Steels, indistinguishable from the traditional position on categorization and symbol use discussed in Chapter 2. If this setup can be extended by designing agents which apply the symbols

[7] For an example where vowel systems are studied see de Boer (2000). For another example in the same direction where sound systems of language are studied see Oudeyer (2005). For a language game in which descriptions of situations are exchanged see Steels and Baillie (2003).

[8] Also see Steels (2006) for a discussion of constraints on the scientific viability of work in modelling language evolution.

used in communication in embodied behavior within task situations, one more step can be taken in moving away from the internalist picture of mentalism, and towards achieving symbols that are situated in the sense discussed in Section 3.2.

The language games setup has been used in a variety of further studies that include various other mechanisms. In one such study, Kwisthout et al. (208) focus on the role of joint attention in language acquisition and lexicon formation. The authors cite Tomasello (1999), according to whom there are three developmental stages of joint attention. The first is called "checking attention", and refers to the ability of an infant to verify that an adult with whom she is playing is also paying attention to the same object. The second is "following attention"; children acquire this ability when they can follow the gaze or pointing gesture of an adult in order to pay attention to the same object. At this stage, it is also possible to talk about the child perceiving the adult as an intentional agent. The third kind of joint attention is "directing attention". At this stage, children attempt to direct the attention of adults to entities in the environment through communicative acts. In the computational model by Kwisthout et al. (208), these different mechanisms of joint attention are modelled as the various processes with which the topic of a language game is changed. The objects in this particular language game consist of three-dimensional vectors, each dimension with four possible values. Each one of these values is a feature, and each feature corresponds to a meaning. The lexicon of each agent consists of an association matrix, matching forms (i.e. labels) with meanings. After picking an object as topic from a set of four possible objects, the hearer produces an utterance which matches one of the features of this topic, creating a new label in case none is available. The hearer updates its association matrix, adding one to the connection weights between the label received and the features in the context, after the feature set is adjusted according to the attention mechanisms used. In the check attention mechanism, the context consists of only the features of the topic picked by the speaker. In the follow attention mechanism, the hearer selects a random object, different from the one initially picked, which has the feature initially picked as the meaning to be transmitted. This object is communicated to the hearer, and the hearer fixes the context as the features of this second object. Finally, in the direct attention mechanism, the hearer selects an object which has the feature it interpreted the label as referring to. The speaker signals to the hearer whether this object contains the target feature or not. If it does, the new context is the feature set of the object picked by the hearer; otherwise, it is all the other features in the context. The different joint attention mechanisms are applied in this order in different combinations. Without the checking attention mechanism, perfect communicative accuracy was never reached. Compared to the combinations where it was absent, directing attention led to significant improvements. Even more significant reduction time for reaching perfect communicative accuracy was observed for following attention. According to the authors, the results show that, for lexicon acquisition, following attention is more important than directing attention.

An interesting variation on the idea of language games comes from Galantucci (2005), who created experimental setups in which humans could engage in communication, but without any available pre-established linguistic or symbolic means. Two subjects share the virtual environment of a computer game in which they can succeed only if they cooperate. Cooperation is not possible without communication. The channels of communication are controlled, in that the players do not know each other, they are separated geographically, and the only means of communication available to them is through a graphic interface which prevents the use of common symbols such as letters and drawings. This is achieved through the distortion of the drawings made by the subjects on a magnetic board; the vertical movements on the board are transmitted to the other user, but the horizontal movements are filtered out, and replaced by a constant horizontal progress as on a seismograph. Such manipulation of the drawings aimed to achieve three things: reproducing in the visual domain fundamental properties of spoken communication, preventing the use of common graphic symbols and pictorial representations, and presenting the subjects with a novel signal with minimal familiarity, but with many varying dimensions (such as amplitude, frequency, thickness, presence vs. absence etc.) which can be used for communication. The game played by the agents involved navigating a space with four rooms ordered in a two by two grid; the subjects each started in a random room, and had one move per round. The aim was to end up in the same room with this one move. Each room also had a conspicuous sign in the middle which signified this particular room.

In experiments with 10 pairs of participants, 9 achieved perfect performance and a consistent communication system. Two methods to establish a common sign vocabulary was observed. The first of these was *learning by using*: when a subject received a sign from the other one, and the game subsequently succeeded, he proceeded to use the same sign when he wanted to signal making the same move as in the first situation. The second method for establishing communication was similar to ostensive definition; when the agents succeeded in picking the same room, one of the subjects moved its agent continuously against the sign in the middle of the room and at the same time drew the sign which he wanted to further use to signify this room. The pairs who relied on this procedure went through the rooms at the beginning of the game and established signs for each room. These pairs were also the ones which established a vocabulary the fastest.

4.2.1.2 Modelling the emergence and development of syntax

The methods of multi-agent simulation employed primarily in modelling especially the emergence of a lexicon in a population have also been applied to the study of syntactic structure. As it was mentioned at the beginning of this chapter, the move from purely referential symbols to syntactic structures is one of the defining moments in the evolution of language. The emergence of syntax cannot be modelled by the simple use of labels to refer to different objects; some extra cognitive machinery has to be assumed. This is reflected in the models of

the emergence of syntax.

The iterated learning model (ILM) of Kirby and Hurford (2002) is one of the most interesting models of the evolution of grammar through transmission from one generation to the next. The agents in this model possess a grammar, represented in the form of rules for transformation of strings of characters. This internal grammar corresponds to the I-language in the Chomskian distinction of I-language (what the users represent) versus E-language (what exists as utterances). For a language to persist from one generation to the next, it has to be mapped from the I-language to the E-language (through use), and then from the E-language to the I-language (through acquisition). The aim of the work of Kirby and Hurford (2002) is to see to what extent certain properties of the syntax of human language (e.g. compositionality) can be explained by glossogenetic evolution (that is, evolution in a historical timescale, as languages undergo in culture), instead of arguments based on natural selection.

In one version of the ILM, agents are given sets of meanings to represent to each other in a series of interactions (Kirby, 2002). These meanings are represented as statements in predicate logic. A speaker transforms a meaning given to it into an utterance by applying the grammar rules in its internal representation of language. The listener tries to decode this utterance on the basis of the grammar it possesses; if this is not possible, it gets the utterance-meaning pair in order to learn from it. The speaker learns by inventing strings; in case it cannot represent a certain meaning with its own grammar, it tries to find the closest meaning that it can produce, and deduce from this a new string by creating random letter sequences for the mismatching parts. The listener, on the other hand, learns through an induction algorithm which progressively creates more general grammatical transformation rules, removing the simpler rules in the process. The main criterion for the learning algorithm is "that the learner can always parse the data heard, but may also generalize to unseen examples if the generalization is justified by the data" (Kirby and Hurford, 2002). The selectionist pressure in this model comes from the deletion of the rules which have become unnecessary and are subsumed by the newer, more general rules. In the experiments, a speaker engages in a number of such communicative acts with a listener, after which it gets removed from the population, and the listener becomes the new speaker.

The results point to a two-stage evolution of a possible language. In the first stage, a protolanguage emerges, which has words for some meanings, but not compositionality. Strings of characters are simply matched to different meanings, and there is no systematicity. Further into the simulation, grammars start to emerge which make use of systematicity and general rules. The reason for this progressive systematization is that languages (in the sense of a set of grammatical rules, i.e. I-language) are favored to the extent that they enable more general rules. This tendency is not built into the agents; it rather stems from the simple fact that rules which express more meanings are presented more often to the listeners. Furthermore, in a language which has to express a big number of meanings, more general rules are more viable, since they generalize to meanings not experienced in earlier generations.

The model of Kirby (2002) gives insights into the process through which languages converge on a set of grammatical conventions, and offers strong arguments in favor of language change as an explanatory mechanism for the most characteristic features of human language. However, a number of assumptions make this model less then ideal. The first of these is the assumption of internal structural processing capacities which are independent of language. Another problematic aspect of this work is that listeners have access to the meanings the speakers intend to communicate. This assumes not only that there is a one-to-one matching between internal meaning and linguistic meaning, but also that this meaning can be transmitted clearly to others.

Another approach to the evolution of grammar is that of Steels (2002), who starts from the idea "that grammar learning is driven by the desire to communicate as effectively as possible" (p.251). The setup is based on work by Steels and Baillie (2003), where agents engage in language games involving descriptions of scenes. A scene is presented to the agents in the form of a video (such as a ball rolling towards another ball). This video is then analyzed by the agents to derive a description of the scene. The speaker conceptualizes parts of such a description, verbalizes this conceptualization, and transmits this sentence to the hearer. The hearer then decodes this sentence and checks whether it suits any of the scenes it has in its memory. Steels (2002) studies particularly case grammar, and more particularly the signalling of semantic role with case markers. Semantic roles are deduced from the descriptions of scenes through a form of analogical reasoning. The scene descriptions are stored in predicate logic formulas. When deducing sentences from these descriptions, distinctive predicates are selected for the objects, as in the talking head experiments. It is assumed, relying on the talking heads experiments, that agents have already come up with names for such predicates, such as "red" or "ball". The parsing and production of sentences is achieved using a form of unification grammar.[9] The main feature of this process is that when an agent produces a sentence, it first tries itself to parse it, thereby detecting any ambiguities. In experiments without case grammar, these ambiguities are not removed, and communication between the agents can be achieved only when they are at the same time looking at the same scene. This is due to the ability of the hearer to resolve the ambiguity from the scene itself. For successful communication in the absence of a common context, the speaker adds a marker to signal a certain semantic role. When it encounters an unknown marker in the interpretation of a sentence, the speaker assumes that there is enough information in the sentence to resolve the ambiguities. In experiments with various scenes, agents agree on a grammar after 200 games. This number is much higher (700) and much more markers are produced (28 instead of 11) when the agents are not allowed to generalize their grammar through analogy.

Another attempt to bring together the language games framework with syntactic structures is Vogt (2005b). Here, the ILM and the talking heads setup

[9] See Steels (2004) for more details on the processes of creating sentences and deducing descriptions from sentences.

are joined together, but the interaction between the agents is not limited to teachers and students. As in the talking heads setup, geometrical shapes are used, and perceptual processes classify them according to four values on the perceptual channels (three for color, one for shape). The classification process is different from the original talking heads setup; instead of classification trees, a 1-nearest-neighbor algorithm is used. The agents have rule-based grammars which can either create combinations out of a feature vector, or map a word onto such a vector. New rules are created and existing ones are generalized as in the ILM (Kirby, 2002).[10] The consecutive populations consist of 3 adults and 3 children, and after 6000 language games, the adults were removed. The children then became adults, and new children were introduced into the population. The probability with which the speaker is an adult and the listener is a child was varied in the experiments, in order to test for the emergence of compositionality in different conditions. It was found out that compositionality remained high in the language when the majority of the speakers were children. The author draws the conclusion that "the more vertical the transmission of the language, the less likely compositional structures in languages tend to remain if no transmission bottleneck is imposed" (Vogt, 2005b, p.337).

4.2.2 Evolving communicative behavior

The emergence of communication is a hotly disputed topic also in the field of artificial life. These models generally use genetic algorithms to understand the conditions under which communicative behavior emerges and leads to populations with selective advantages.

One of the earliest examples of such work is that of Werner and Dyer (1992), who want to model communication behavior inspired by animal communication as a first step towards understanding language. They stress that communication should not take place for its own sake, but should serve a purpose in a wider behavioral context. Their experimental model consist of a population of males and females on a toroidal grid, and the wider behavioral context is mating. These agents have genomes which encode connection weights and biases of a recurrent neural network. The males are blind and cannot see where the females are but they can move towards the females in order to mate with them. The females, on the other hand, can see, but they cannot move. Instead, they can produce sounds in order to direct the males towards themselves. Based on this difference in sensory and motor capabilities, the output of female networks were interpreted as sound signals and coupled to the input of the males, whereas the output of the male networks were interpreted as movements on the grid, and their position was changed according to this output. In turn, the position and the orientation of the male was given to the female network as input. The genetic algorithm of Werner and Dyer (1992) differs from traditional genetic algorithms in that the population is not subjected to a selection process arbitrarily at a point in time: when a male finds a female, they are mated, and they and their

[10] For the specific details see Vogt (2005a).

offspring are placed on different points on the arena afterwards. This enables the co-presence of different generations on the grid, which is a suitable condition for common signs to emerge.

In the experiments, first, the males which simply stood in place were eliminated through evolution. Afterwards, the males which simply went forward were selected, and the ones that simply circled or took too many turns were extinct. After around 50,000 time steps, females evolved to give signals to males to turn to face them when they were in the same column. In a control population in which the females could not give out signals the males would find a female in an average of 100 moves. In the signalling population, this value was 40 moves.

MacLennan and Burghardt (1993) study the emergence and evolution of communication in a very similar manner to Werner and Dyer (1992), in that they base communication on the sharing of a certain piece of information which is available to some agents but not available to others. MacLennan and Burghardt (1993) also argue that this information has to have significance to the agents. However, the environments in which the agents operate and their communication mechanisms are even simpler: both the environment and the communication signal are single variables which can take on certain values. The agents are finite state machines, which are encoded in strings in order to apply genetic algorithms. In each step, all agents in a population either emit a signal according to their current state, or take an action based on the state of another agent, which the authors call cooperating behavior. As it can be expected, communicative behavior increases the fitness of the population. However, the simplicity of the environment and behavioral repertoire makes the model of MacLennan and Burghardt (1993) simply a case study in analogies in artificial evolution.

Parisi (1997) compares the artificial life application of neural networks to their use in other and earlier connectionist approaches, especially in the modelling of linguistic phenomena. He gives examples from work on evolving neural network controllers for miniature robots. One of the examples is the individual use of labels to improve categorization behavior and thus fitness for an environment. When a robotic agent which functions in an environment populated by objects with varying significance (i.e. danger and food) receives, in addition to the sensory input from the objects, labels which correspond to the significance of this object, the performance of the agent increases, which means that language can function as "an aid to categorization by stabilizing the class of objects that must be put in the same category and responded in the same way" (Parisi, 1997, p.129). In order for symbolic communication to emerge, these signals have to come from another agent in the vicinity; if this is the case, the signals received from other agents can further be used by individual agents. Why agents should bother to send such signals identifying categories is a further problem addressed. Parisi et al. (1995) have shown that altruistic behavior (in the form of signalling the presence of food) evolves when agents are equipped with an altruism gene which determines how much of the food they eat and how much they give to offspring which has the same altruism gene as themselves.

The emergence of a simple language in a population has been modelled by Cangelosi and Parisi (1998). This study is a clear example of the general

tendency in artificial life studies of symbolic communication of creating agents in an environment with different kinds of objects, different actions that the agents can carry out on them, and different kinds of outcomes in terms of effects on the survival chances of the agent. In the case of Cangelosi and Parisi (1998), the environment is a grid of 20×20, and it is populated by mushrooms of two kinds, edible or poisonous. These kinds of mushrooms are differentiated through their visual properties. Stepping on a cell with a mushroom leads either to acquiring more energy, if case it is edible, or losing energy, in case it is poisonous. The agents, which are feed-forward neural networks with a hidden layer, receive as input the location of the nearest mushroom. The visual properties of the mushroom are also given as input in case it is in one of the eight cells neighboring the agent's current location; otherwise, these input values are set to zero. The agents also have input nodes for signals from other agents. A genetic algorithm picks the fittest agents, decided on the basis of their energy levels, and creates offsprings from these. In one population, the agents do not have communicative abilities; in another, communication is modelled with two agents being placed into the same environment, where one undergoes the same experimental procedure as normal, but also receives a signal from another agent which is supplied with, in addition to the location of the closest mushroom, its visual properties, no matter how far it is. In a comparison of two evolved populations, one with communication and one without, the population without communication had 150 units of energy on average at the end of the experiment, while the population with energy had 250 units of energy on average. Cangelosi (2001) argues that "this simulation shows that a population of simple artificial organisms living in a simple environment can evolve an efficient language with an informative function to help the individuals interact with their environment".

This work has been extended to study certain aspects of linguistic and symbolic communication. The grounding of symbols and the differentiation of different kinds of words have been studied by Cangelosi (2004). A simulated agent with an arm with two degrees of freedom was given a perceptual input of a 5×5 grid. On this grid, either horizontal lines (category A) or vertical lines (category B) with a length of three cells were presented. The agents were feed-forward neural networks with an input layer consisting of 4 units for proprioception, 25 units for visual input, and 4 units for language input. The output was 4 units interpreted as the extension of the two joints for moving the arm. In the first part of the simulation, agents were selected for pushing B stimuli away from and pulling A stimuli towards the agent. Once there is a population of agents which can carry out this task, this population is further evolved for being able to interpret input to the language units. The inputs to the language unit consist of two words of two units each. There are two verbs corresponding to "push" and "pull", and two nouns corresponding to A and B. When a rule is given which conflicts with the default one for which the agents were evolved in the first stage, this new rule overrides the default rule.

In order to compare their performances and grounding of linguistic symbols in behavior, three populations were evolved. The first of these populations had no linguistic experience (no-language), the second was evolved as explained

above (late-language), and a third one was evolved with linguistic input from the beginning (early-language). In a test phase without linguistic input, the performance of the late-language population was the best, with 16.6 successful tasks in 20. Second was the no-language population, with 15.8 successful, and the early-language population was last with 14.4 successful (Cangelosi and Parisi, 2001). According to Cangelosi and Parisi (2001), this result proves that when language is allowed to build on already existing cognitive capacities, it leads to not only better application of language to behavior, but also to improved performance in nonlinguistic behavior. Further analysis of the performance of the late-language group shows better performance in epochs in which only noun words were given as input once language selection is started, but better performance in verb and verb plus noun epochs around 100 generations later. This is explained with the consistency of the no-language task with that in the noun tasks. In order to succeed in the verb tasks, the agents have to ignore some of the previous information. Verb tasks improve once noun performance stabilizes and can serve as a basis for applying nouns.[11]

A similar study aimed to understand the phenomenon of symbolic theft. Categories can be acquired, according to Cangelosi and Harnad (2001), in two ways. The first of these is "sensorimotor toil", which refers to the acquisition of categories "through laborious, real-time trial and error, guided by corrective feedback from the consequences of sorting things correctly or incorrectly". The second way to acquire categories is through symbolic theft, which refers to picking the distinction through an encounter with someone who already has it. Symbolic theft is the primary means of social category learning, where categorical distinctions are spread in a community. An important feature of this process is that it relies on the existence of categories which are already earned through toil; otherwise, these symbols are prone to the symbol grounding problem. Symbolic theft is also the main reason for the adaptive advantage of language, since it allows speakers to acquire knowledge without the risks or costs of direct trial and error experience.

The experimental setup for modelling these processes is very similar to the one explained above, a grid populated with mushrooms. The main differences are that the mushrooms now have two kinds of visual features which can occur together (resulting in mushrooms of type A, B, and AB), and there are three actions which can be carried out on the mushrooms. The actions are eating, marking and returning; a combination of these actions has to be carried out on each different kind of mushroom. The agents also make vocalizations when they approach a mushroom, specifying what they are going to do. What the agents undertake with the mushrooms and which signals are emitted are learned through supervised learning. In addition to supervised learning, a genetic algorithm selects the fittest agents and reproduces them. In order to compare the difference of theft compared to toil, two populations are evolved. The first stage of the experiment is the same for both populations: they all learn to eat mush-

[11] In Marocco et al. (2002), a similar task environment is explored, but with a more advanced robotic task involving an arm with six degrees of freedom.

rooms with feature A and mark mushrooms with feature B, through toil. In the second stage, the agents in the toil population learn to return to mushrooms of type AB through toil, whereas the agents in the theft group learn to return to the AB mushrooms on the basis of hearing the vocalization of the mushrooms' names. Statistical analysis of the number of AB mushrooms collected revealed that the theft population was more adaptive than the toil population. In order to make a direct comparison, members from the two populations were mixed and evolved together. In less than 10 generations, the thieves took over the whole population.

As a critical perspective on the work done till then, Di Paolo (1998) argues that biologically motivated studies of communication have taken for granted what they are supposed to explain, and that the definitions of communication on which they rely as transfer of information already assumes a certain kind of explanation. Information transfer here refers to a signal reducing the receiver's uncertainty about the world, e.g. by specifying a fact not directly visible in the environment, or not known by the receiver. In biology, as well as in the study of evolution of communication, information transfer has been assumed to be beneficial in a selective context, and the form communication has to take place has been derived from this assumption: "By inappropriately deriving operational features from functional conclusions, many researchers have assumed that a necessary condition for communication is that not all relevant aspects of the environment are known equally to all the participants" (Di Paolo, 1998, p.290). In order to counter this tendency, Di Paolo (1998) employs autopoietic theory to define communication as any kind of behavioral coordination, without the necessity of information transfer, or difference in knowledge of the environment between the agents.

In order to test this formulation, a game of symbolic communication is devised, in which simulated agents are placed on a toroidal grid with different kinds of food sources scattered. All agents are born with a certain level of energy, and the level of this energy decides whether they can reproduce (for high energy levels) or are removed from the population (for energy levels under a certain threshold). Energy can be acquired from the food sources, with different kinds of sources necessitating different kinds of actions. Each action has an effective component, which decides how much energy is extracted from a food source, and an external manifestation, which does not have an effect on the amount of energy extracted. In one time unit, a game is played, in which two agents and a food source from their vicinity are selected. One of these agents carries out an action. The other agent perceives the external manifestation of this action, but not what kind of food source the action was carried out on. If both agents carry out the correct action, the total amount of energy in the food source is distributed equally among them. Otherwise, both agents receive less energy. The agents are encoded as stateless machines, simply encoding the response to different kinds of foods and manifestations. When the simulation was run, it was observed that the agents formed clusters, due to the agents in sparse regions dying out. The population of agents is observed to rely on behavioral coordination in order to maximize energy intake from the food sources. The

agents organize themselves in clusters, with these clusters dynamically moving, splitting and emerging. The formation and movement of these clusters are due to the dual dynamics of genetic inheritance and food sources getting depleted.

Further research on the evolution of communicative channels has been carried out by Quinn (2001), who has proposed working on evolving communicative behavior out of other simple behaviors. The emergence of communicative behavior is a difficult topic because the signal and the response have to come to being together; each, on its own, is meaningless and does not proffer any advantage. One possible way to explain the evolution of communication is by postulating an originally non-communicative role for the signal or the response. In the examples seen above, and in general in the genetic models of communication, the communication channels serve solely one purpose, and their having any non-communicative function is categorically excluded. The alternative proposed by Quinn (2001) takes the form of simple Khepera robots, equipped with short-range infrared sensors with a certain range and two motor-driven wheels, coupled in a task of mutual coordination. The agents, controlled by neural networks, are placed in each other's vicinity in a number of different constellations, in which the distance between the agents (always smaller than the sensory range) and their orientation varies. The fitness function used in the genetic algorithm responsible for evolution is constructed to pick the couples which succeed in travelling a certain distance together without bumping into each other, staying in each other's sensory range at the same time. The same fitness value is calculated for the agents in a single run, so that there is no competition between them. The only way the agents can communicate for coordination is by using their IR sensors and motors, and this has to be established without information on the direction the other agent is facing, since the agents are cylindrical. After 2000 generations, 27 of the initial 30 populations produce successful solutions to the task. In all these populations, one of the agents take on the role of the leader, and the other follows. In one successful population, the final agents start rotating counterclockwise when the round starts. Once an agent senses another agent in front of it, it stops and starts going back and forth. The other agent, also having placed its counterpart in front of itself by rotating, then takes on the role of the leader and starts backing up. The first agent follows it, with the two agents now moving together in one direction. Analysis of the networks of these agents show that the back-and-forth movement of the first agent signals to the second agent that it is ready to run as a follower. Further analysis of the evolution of the agents also proves that this signaling behavior was based on the foregoing evolution of obstacle avoidance (i.e. agents going back and forth when they came too close to the other one) and one of the agents backing up because the other was not able to do so.

An interesting recent addition to the evolutionary models of communication which eschew dedicated channels is that of Williams et al. (2008). They call the interpretation of communication as the use of signals to transmit information the IT view, short for information transmission. This is contrasted with the view of communication as encompassing all kinds of socially coordinated behavior, a stance they call the CB (coordinated behavior) view. The

IT view has been occupied primarily with referential communication, i.e. the transmission of signals which refer to objects or situations independent of the agents. The CB view, on the other hand, has focused on other kinds of tasks, such as the leader/follower signalling behavior in the work of Quinn (2001). The aim of Williams et al. (2008) is to study referential communication from the CB point of view. To this end, artificial agents, controlled by continuous time neural networks, are evolved using genetic algorithms on a 2-dimensional circular environment. Instead of coevolution, a pair of agents, one sender and one receiver, are encoded on the same genome. Both agents have sensors with a certain range which tell them where the other agent is. The sender also can sense the position of a certain target location on the circle. The task is to have the sender direct the receiver to this target location through behavioral coordination. Referential communication emerges when the sender is confined to a certain region on the circle which does not include the target point. In the absence of these conditions, the sender simply goes to the point where the target is and places itself a certain distance from the target. When the target can take on a limited number of values (four in the experiments reported), the sender exhibits different kinds of easily distinguishable behaviors, such as moving to the top or bottom of the allowed arc, in order to signal the location of the target. In a further experiment, ten target locations were used in order to force the agents to develop strategies that can be generalized. Under this second condition, the sender-received evolve to cross their paths twice at the beginning of the experiment. The time gap between these two crossings is controlled by the sender, because the receiver moves with a constant speed. The sender uses this time gap to signal the distance of the target point to the receiver, thereby enabling systematic communication of a continuum of values. According to the authors, "this ability to indicate a continuum of locations is analogous to the various deictic indicators used in human communication, such as finger pointing and eye gaze" (Williams et al., 2008, p.708).

Grim et al. (2002) have used both genetic algorithms and neural networks to study how meaning "happens" in a population. Their approach is based on a certain philosophical position regarding meaning on the one hand, and a critical perspective on most models of communication on the other. The philosophical viewpoint relies on the ideas of Wittgenstein, presented in Section 2.4.1. Grim et al. (2002) point out that relational theories of meaning, where either an internal or external object provides the meaning of a word or signal through a relationship of correspondence, are widespread in cognitive science and AI. They further argue that

> [A] grasp of meaning will come not by looking for the right kind of object but by attention to the coordinated interaction of agents in a community. In practical terms, the measure of communication will be functional coordination alone, rather than an attempt to find matches between internal representations and referential matches (p.46).

Grim et al. (2002) contrast the philosophical discussion on meaning with the

discussion on the status of life more than a century ago. Once life was considered a component of all living beings, an inner quality which living bodies had but dead bodies lacked. This conception of life was later replaced by a functional understanding in terms of evolving communities. Such an understanding also allows for treating life as a matter of degree instead of as an either-or property. The development of the biological concept of life can point to the way out of the representationalist view:

> The proper way to understand meaning may be on the analogy of our current understanding of life: not as an all-or-nothing relation tying word to thing or idea, but as a complex continuum of properties characteristic of coordinated behavior within a community – a community of communicators – developing over time.

This is also the reason Grim et al. (2004), who want to measure communicative success and thereby the extent of meaningful interaction by the behavioral coordination of a community of agents, talk about making meaning "happen" instead of providing a locale for it, or modelling its transfer. Coordinated behavior here refers to a community of agents adopting similar communicative and non-communicative behavioral strategies without appropriating each other's internal machinery:

> In the model above, there is no guarantee that the internal workings of behaviorally identical strategies in two individuals are themselves identical. There are in principle nondenumerably many neural configurations which may show the same behavioral strategy. In training to match a neighboring 'perfect communicator', a neural net may not only fail to match the absolute values of its neighbor's weights, but may not even match its over-all structure of relative weight balances. What arises in a community is a pattern of coordinated behavior, but in evolving from an initially randomized array of neural nets that coordinated behavior need not be built on any uniform understructure in the nets themselves (Grim et al., 2002).

The first and most general of the criticisms of Grim et al. (2002) concerning models of language evolution targets the common practice of rewarding both parties in a communicative exchange in case of success. The authors argue that this leads to an artificial environment in which truthfulness simply leads to fitness, which is not a realistic assumption about natural communication. Instead of such a scheme, an individualistic reward structure is employed in their own model. Another criticism, which also relates to the mutual benefit assumption, regards the spatial character of the models. In many models, as explained above, the communicating agents are not embedded in any spatial grid which provides a constraint and a geography to spread at the same time. This problems is also addressed in their model.

The alternative model proposed by Grim et al. (2002) consists of agents controlled by neural networks in a toroidal array. These agents can hide or open

their mouths, and they can give out two different signals, but otherwise they are stationary. The array is also inhabited by sources of food and predators, which travel through it. If an agent is not hiding when a predator visits its cell, it loses a point of energy. If it opens its mouth when a food source is visiting the cell, it receives one point energy. The emission of a signal costs 0.05 energy points. Such a signal can be perceived only by the neighboring eight cells. With these sensory and motor capacities, only two optimally communicative strategies are possible. The first of these is giving out signal 1 when a food source visits the cell, and opening mouth when the same signal comes from another cell. The other part of this same strategy is giving out signal 2 when a predator appears, and hiding when the same signal is received. The other optimal strategy is the exact same, but with the signal switched, i.e. signal 2 for food source and signal 1 for predators. The agents go through a "century" (100 rounds) of simulation, after which each one of them looks at its immediate neighborhood to pick the agent with the highest energy. If this agent has more energy then the agent itself, it is picked as an instructor in a supervised learning task. The learning task is not carried out until the student is identical to the instructor, however: in the experiments, it was found that such rigid copying led to suboptimal strategies dominating from the beginning. Backpropagation-based supervised learning is carried out only for a limited number of random input-output values. Over the course of 300 centuries, nearly the whole population learns to employ one of the perfect communication strategies. Bordering communities are formed over the grid in which one of the two strategies dominate.[12]

The approach presented by Grim et al. (2002) is very interesting in that they take their inspiration directly from alternative theories of linguistic meaning, with a clear eye for the pitfalls of building in simplistic assumptions, such as communication as the transfer of correspondences between signals and objects, or the measure of successful communication being the equality of internal structures among agents. Grim et al. (2001) also argue that there is not only one way of thinking about genetic algorithms, namely as approximations to biological evolution; they can also be thought of as the transfer of "memes", cultural strategies that are partially transmitted and recombined. This focus on the process of behavioral coordination and the spread of different strategies instead of the kinds of learning method used allows them to study the emergence of signal-based communication as a "general process facilitated by the environment pressures of a spatialized environment" (Grim et al., 2002, p.66). However, it must be noted that the physical capabilities of their agents are rather limited, with two signals and two on-off behaviors, without embodied coordination. The idea of behavioral coordination is a very important contribution to models of the emergence of communication, and it offers a lot of potential to be extended with more complex behavioral and spatial dynamics.

[12] Grim et al. (2001) implements pretty much the model but with genetic algorithms.

4.3 Situated representations in language evolution

Building on the discussion in the preceding chapters, especially the discussion of situated representations in Section 3.3, and the overview of the theories and models of the dynamics and emergence of symbolic communication, a possible path to understanding human symbol use through multi-agent models clears up. This path goes through the communicative use of gestures and utterances, grounding these in the shared situation, and the further use of these same gestures for the purposes of the individual. If such a study is to offer any insights into the role of language in human intelligence, the possible pitfalls of the cognitivist approach have to be avoided. Some of these pitfalls can be observed in the work on dynamics of communication, as the overview given above shows. The most important and most indicative of these is the use of system-internal representations as the meanings of communicative symbols, and attaching labels to these representations to create units for communication. The general form this tendency takes in experiments on the dynamics of communication is that the agents are presented with objects which they have to identify in order to communicate about them; their engagement with these objects is limited to identification, with signals created on the basis of this identification (cf. Türkmen, 2007).

In order to model human symbol use in individual activities and communicative situations, situated representations should be seen as a general guideline. As a short reminder, the distinguishing properties of situated representations are that

- their creation is coupled to the perception of their possible use (in communication or embodied behavior) in a situation (cf. the discussion of the brush example on p. 53),
- their use is not privileged, in that they serve as resources that have to be interpreted both in private use and communication (cf. the discussion of plan use on p. 50).

These properties of situated representations entail two primary features which have to be integrated into current research, especially into the language games setup. The first of these is the continuity and simultaneity of *acting* and *uttering*, i.e. a coordination of linguistic acts of symbolic character with robotic tasks which involve goal-directed behavior. Coordinating linguistic behavior with goal-directed behavior is impossible in a paradigm which sees linguistic symbols as mere labels which get attached to categorizations carried out by a separate module. An alternative conception of the relationship between symbolic communication and perceptuo-motor coordination is required. Such a conception should further provide a perspective for the grounding of symbols in embodied interaction, as well as offering an explanation for the special role of language in human cognition; this can be achieved only if the experimental setup is designed to enable embodied behavior modified through the use

of communicative symbols. The second feature that has to be integrated into research on dynamics of cognition is the reevaluation of individual situations to find possibilities of applying situated representations, these possibilities presenting themselves as behavioral strategies that can be achieved through the use of situated representations. Such continuous evaluation of the situation corresponds to re-contextualizing representations in each situation, achieving context-independence through goal-directed application of situated representations instead of through filter-based abstraction of perceptual data.

Therefore, what is expected of a computational study of symbolic communication can be summarized under the following rubrics:

Embodied grounding of meaning The model should build on the embodied capacities of the agents. As in the work of Hutchins and Hazlehurst (1995), the formation of the lexicon should be a process of agreeing on shared distinctions, but these distinctions should be of embodied character. As presented in the work of Steels above, it is possible to let agents create discriminations and base the communicative capabilities on these distinctions. One possible path to follow is to have these distinctions arise out of the sensory-motor coordination, instead of basing them solely on perception.

Social lexicon formation The work presented above on the formation of lexicons in a community shows that it is possible to model this phenomenon through language games in shared situations. Symbols which the agents can use individually should be created in such language games, and shared among the agents for the establishment of common means to refer to entities or choices.

Social coordination of behavior instead of information transfer The grounding of shared distinctions in the embodied possibilities in the environment, coupled with social lexicon formation, would also pave the way for an understanding and implementation of communication not as transfer of information but, as in the work of Grim et al. (2002), social coordination of behavior. The embodied capacities of the agents can be used to create activity-oriented scenarios for symbol use, and the creation of a lexicon can thus be modelled as being grounded in the common embodied capabilities of the agents.

No internal representations before shared symbols Ideally, any representations, symbols or labels should be created in a communicative context. The symbols used for communication should also not simply refer to an inner representation, or computational structure. Possible ways representations could be used this way have been presented in the preceding chapter.

Categorization in a task context As it can be seen in this overview, the task context is very important because it determines whether the meanings are internal or not. As explained in the preceding chapter, situated concepts are instantiated in a task context, because they serve the task at hand. Therefore,

in order to model situated representations, the communicative contexts should be task-oriented and not necessarily object-oriented. That is, the task should not be solely the identification of an object, but a certain interaction with the situation.

In the next chapter, a study on the dynamics of communication will be presented which aims to implement these principles in a computational model. It should once more be mentioned here that the human language capacity is very complicated and comprises many distinct abilities. As mentioned above, it is possible to start a study of the linguistic capacity by concentrating on an initial set of capabilities, concerning the use of individual symbols, their formulation, storage, and comprehension. In the work presented here, the focus will be on these capabilities due to two reasons. The first of these is that the symbolic capabilities make the syntactic ones *possible* in the first place, and the symbolic capacities have not been properly understood yet, as argued in Chapter 2. The second reason has to do with the nature of symbolic communication, the scientific mist that surrounds it, and how understanding symbolic communication would change our understanding of communication and language comprehension in the first place; that is, any insights achieved by understanding situated representations will tell us more about what we have to look for in a theory of further components of the linguistic capacity.

Chapter 5

Similarity-based categorization and language games

The embodied and situated approaches have established the significance of aspects of intelligence which have been neglected both in the cognitive scientific and the more general philosophical context, aspects such as embodiment, social situatedness and the interdependence of sensing, acting and representing. It has nevertheless been a problematic endeavor to bring together the symbolic and productive aspects of human intelligence and the low-level capabilities.

In Chapter 4, it was argued that studies on the origins of language and the dynamics of symbolic communication constitute a candidate approach for understanding the interplay of the embodied, social and representational facets of human intelligence. Furthermore, five guiding principles for further research in this area were presented. In this chapter, an approach to the dynamics of communication will be presented, with a computational model and results from experiments with this model.

5.1 Exemplar-based Learning and Categorization

Exemplar-based learning is a kind of lazy learning which relies on the storing of complete sensory data and avoids the creation of structures and abstractions which would aid faster processing of data in decision tasks. Lazy learning, also called local memory-based learning, refers to the family of algorithms which "defer processing of the dataset until they receive request for information (e.g. prediction or local modeling)" (Bontempi et al., 1999). In exemplar-based learning, episodes of experience are stored in memory in as much detail as possible, to be later retrieved based on a cue stimulus. This means that the relevance of

the content of an exemplar is decided when new stimuli are encountered. This is one of the strengths of exemplar-based learning: the data does not have to be abstracted based on design considerations and knowledge of the designer.

One possible objection to memory-based learning methods is that they violate the parsimony principle mentioned in Section 3.1.2, in that they necessitate excessive information processing, and increasing processing load as more experience is gathered. Computationally heavy processing machinery, such as low-level similarity comparison of a large set of raw data, can be seen as antithetical to a research program which stresses simplicity of proposed mechanisms. Against this argument, it should be kept in mind that the computational load of an algorithm is dependent on the nature of the concrete devices on which these algorithms are executed. For example, parallel processing with a large number of nodes is inefficient on hardware with a serial architecture, such as the von Neumann architecture of modern computers, but not so on a parallel computer with a large number of relatively simple processors. On such parallel machinery, the comparison of large amounts of perceptual data, as it will be proposed with the lazy learning algorithm explained in this chapter, is not as counterintuitive as against the background of the von Neumann architecture.

The two most important aspects of an exemplar-based model are how the sensory data is represented and the similarity and distance measures used when comparing exemplars which store sensory data. The two most suitable perspectives for the discussion of these two aspects in a cognitive scientific setting are the psychological and machine learning perspectives. It must be pointed out that the machine learning approach presented here is a synthetic application of ideas from the psychology literature; therefore, the focus is not on expounding every detail of the exemplar approaches, but presenting the major ideas, and the experimental proof which certain ideas have received. The application developed later extends these ideas with practical considerations in mind for the construction of a program capable of behavioral categorization.

5.1.1 Psychological approaches

In psychology, exemplar-based models have been a significant alternative to other theories of categorization at the latest since the context theory of classification by Medin and Schaffer (1978).[1]

[1]The difference between concepts and categories in cognitive science and psychology is a very fluid one. Theories of categories and concepts have very similar agendas and face the same problems. It can be claimed that concepts have a mentalistic flavor (e.g. when Margolis and Laurence, 1999a, define concepts as subpropositional mental representations), whereas categories are generally seen as things out there, and categorization behavior is the more important problem, rather than categories themselves. Smith and Medin (1981) view categorization as one of the applications of concepts, and in their opinion, "concepts are essentially pattern-recognition devices" (p. 8). The following approach will be taken here: concepts are the names of the things "in the head", which enable differential behavior. Categories, on the other hand, are ad hoc in that they refer to the situations that are treated differently by an agent. Therefore, categorization as a behavioral capacity takes precedence to categories (or concepts) as a scientific subject.

Traditional theories of categorization view it as the application of concepts, abstract mental representations that consist of definitions, to entities. According to what is called the *classical theory of concepts*, "most concepts encode necessary and sufficient conditions for their own application" (Margolis and Laurence, 1999a, p.9). These component conditions are encoded as a set of representations, and specify rules for the inclusion of an object under this concept. One argument against the classical view of categorization has already been mentioned in Section 2.4.1, in the discussion on Wittgenstein. Many categories and concepts used by humans do not have such a strict set of common features, and are instead defined by family resemblances, i.e. commonalities which are not necessary conditions, but are highly indicative in case they occur together. In addition to this and similar theoretical arguments against the classical theory, there are numerous experimental results which contradict it.[2] One of the earliest alternative theories to the classical view was the prototype theory of Rosch and Mervis (1975), according to which category membership is decided on the basis of similarity to a prototypical member of a category[3]. The prototype theory of categories aimed to bring together two sets of experimental results which contradicted the traditional, rule-based approaches. The first of these was the absence of a strict set of features which determine category membership. Category members, as judged by humans, exhibit family resemblances instead of a common set of features. The second set of results concerns the existence of typical category members which have features judged to be characteristic of a category, but not necessary for its definition; it appears that humans use unnecessary features in making category judgments.

The exemplar theory of Medin and Schaffer (1978) was developed to apply the then recent insights in discrimination learning to categorization. The principal aim was to account for experimental results which point to prototype effects without recourse to abstractions, such as feature-based rules for category inclusion or prototypes.[4] The most important idea propagated by the context theory, in applying discrimination learning approaches to categorization, was the possibility of basing classification on exemplar level information, i.e. information retrieved from the exemplars in memory, based on the stimulus (as a retrieval cue) in a trial. Exemplar level information here is defined in contrast to category level information, such as rules deduced from category samples. In order to classify a stimulus, "it is assumed that the evidence favoring a category

[2]See Smith and Medin (1981), Chapter 3 for an overview of experimental evidence against the classical theory. See also Lakoff (1990) for an impassioned critique of the classical theory from a more linguistic perspective.

[3]One can object that when the similarity of objects is defined on the basis of salient features, this similarity can also be stated in terms of rules, which would render an alternative similarity-based approach practically equivalent to the classical one. It is not necessary to claim, however, that similarity ground all categories, but only the ones which are perceptually more primitive. In this case, the distinction of rule- vs. feature-based concepts would still be intact. This topic will be discussed in Section 5.1.1.1.

[4]It is possible that a prototype for a category is among the examples seen by a subject in an experiment, but in case such a prototype was not presented, the derivation of a prototype has to be assumed for categorization. This gives such a prototype the status of an abstraction, because it is not readily available in stimulus data but has to be constructed by the subject.

j response to probe i is equal to the sum of similarities of probe i to the stored j exemplars divided by the sum of the similarities of probe i to all stored exemplars" (Medin and Schaffer, 1978, p.211). In the case of the context theory, and the other psychological theories adopting an exemplar-based approach, the exemplars do not refer to individual experiences and the sensory data caused by these experiences, but to the different stimuli which a subject is confronted with multiple times as the member of a category.

The idea of attributing probabilities to categorization decisions on the basis of similarities to category exemplars can be derived from extensive work on identification, especially that of Shepard (1957), who studied experiments in which distinguishing responses were attached to different stimuli; for example, response R_j (pressing one of a number of buttons) serves to distinguish stimulus S_j (one of a number of different visual figures). The central concern of the model constructed by Shepard (1957), which is called the similarity choice model, was "not with the learning process per se but with the pattern of generalizations exhibited at any one given stage of learning" (p.325). More concretely, the aim of the theory was to find out the nature of the probability of the ith stimulus of the N stimuli $S_1, S_2 \ldots, S_N$ leading to the jth response of the N responses $R_1, R_2, \ldots, R_N$ in an identification experiment. This probability is denoted with $P(R_j|S_i)$. In order to arrive at a formulation of this probability, Shepard (1957) made a number of simplifying assumptions. The first of these assumptions is analyzing the performance of the experimental subjects not in terms of the stimulus-response ($S-R$) associations they make, but in terms of primarily the stimulus-stimulus ($S-S$) confusions which cause them to make errors. That is, it is not the association between stimuli and responses that are confused, but the different stimuli among each other: "[T]he subjects know the connection between any stimulus and its assigned response but still make errors owing to a certain inability to identify the stimulus" (Shepard, 1957, p.328). A direct result of this assumption is that at any given time in the learning process, the confusion probability of a certain stimulus with any other stimulus should be nonnegative, and all such probabilities add up to one. More concretely, if one represents the conditional probability that S_i, when presented, will be taken to be S_j, with P_{ij}^S, the following condition holds:

$$\sum_j P_{ij}^S = 1, \quad P_{ij}^S \geq 0 \tag{5.1}$$

It is possible to calculate, using the confusion probabilities matrix, which depicts the chances of confusion of various different stimuli with each other, and some simple linear algebra, the confusion probabilities for an individual subject or the whole experimental group, based on data obtained in identification experiments. However, there are still too many parameters which have to be decided in these calculations compared to the number of different stimuli ($N(N-1)$ parameters for N stimuli) in order to arrive at a complete description of the performance of subjects. Shepard (1957) argues that one simple assumption would serve to decrease the number of these parameters dramatically, and thus the number of

experiments that have to be carried out in order to derive the confusion matrix for a set of stimuli. This is the assumption that there is a relation between P_{ij}^S and P_{ji}^S of the $S-S$ matrix. One intuitive way of formulating such a relation is taking the confusion probability of two stimuli to be a function of their similarity, so that P_{ij}^S and P_{ji}^S decrease or increase together as the similarity of S_i and S_j are made respectively smaller or larger.

Shepard (1957) further proposed to study the confusability of two stimuli not primarily in terms of their similarity, but their distance in a psychological space, later converting this distance into similarity through a function f. Intuitively, there are three constraints on the function f; it should be positive, it should evaluate to 1 for identical stimuli, and it should be monotonically decreasing, so that with increasing distance between stimuli, the similarity value tends to zero. The most important merit of such an approach is that it allows basing a theory of categorization on sound mathematical fundamentals by relying on the tools and metaphors of geometry. It must be pointed out that the geometric approach to similarity and categorization is one of the various dominant paradigms in psychology on this subject[5]. One common feature of geometric models is that, as a theoretical result of the intuitive geometric constraints on the similarity-distance relationship of stimuli, they rely on the notion of a psychological space as a metric space. A space S is called a metric space if it has a distance function $d(a, b)$ which satisfies the following conditions for all points a, b, c in this space (Gärdenfors, 2000, p.18):

- Minimality: $d(a, b) \geq 0$ and $d(a, b) = 0$ only if $a = b$
- Symmetry: $d(a, b) = d(b, a)$
- Triangle inequality: $d(a, b) + d(b, c) \geq d(a, c)$

Through this assumption, it is possible to apply our intuitions from geometry to an understanding of concepts and categories. This point is strongly supported by recent theories such as that of Gärdenfors (2000), who develops a comprehensive approach to categorization by defining categories as convex regions in psychological spaces. As it will be argued in a subsequent discussion of similarity in psychology, for a theory working on the fundamental aspects of categorization, such as perceptual categories and behavioral categorization, the various complicating aspects of similarity judgements (e.g. structural similarity, familiarity effects) are topics which appear when additional complicating factors such as language, structural factors, social embedding and prolonged experience in different contexts come into play. It would be wrong to expect a model of categorization to explain all these effects, especially from an approach, such as the one presented here, which argues that multi-agent dynamics, and not solely the agent-internal processes, are responsible for the various complexities of conceptual and symbolic behavior.

[5]Other approaches are featural models, such as the Contrast Model by Tversky (1977), and structure mapping models, such as that of (Markman and Gentner, 1993). See Goldstone (1999) for an overview.

Adopting a distance measure in the psychological space, the confusion probability of two stimuli S_i and S_j (assuming S_i has been presented first) can be expressed as follows in terms of their distance in the psychological space, D_{ij}^S[6], a constant d_i associated with S_i, and the distance-to-similarity function f:

$$P_{ij}^S = d_i \cdot f(D_{ij}^S) \tag{5.2}$$

The similarity function f translates distance in the psychological space to perceived similarity of two stimuli, resulting in 1 for the same stimuli and 0 for completely different stimuli ($D_{ij}^S \to \infty$); widely used similarity functions and their properties are discussed on p. 103. Equation 5.2 can be coupled with Equation 5.1 as follows, replacing j with h as the generic index in order to avoid confusion:

$$\sum_h P_{ih}^S = d_i \cdot \sum_h f(D_{ih}^S) = 1 \tag{5.3}$$

Solving Equations 5.2 and 5.3 for d_i, we get the following:

$$d_i = \frac{1}{\sum_h f(D_{ih}^S)} = \frac{P_{ij}^S}{f(D_{ij}^S)} \tag{5.4}$$

Solving Equation 5.4 for P_{ij}^S, we arrive at an expression for the confusion probability (or for the case $i = j$, identification probability) of the stimuli S_i and S_j:

$$P_{ij}^S = \frac{f(D_{ij}^S)}{\sum_h f(D_{ih}^S)} \tag{5.5}$$

There are two further questions which need to be answered, or two missing components which have to be supplied by such an approach. The first of these is the well-known problem of the asymmetry of similarity. Human subjects vary their judgement of various familiar stimuli depending on the order in which these stimuli are presented; for example, North Korea is judged by subjects to be more like China than China is judged to be similar to North Korea (Tversky, 1977). In order to account for this effect, Shepard (1957) proposes to integrate a weight W_j^S with each stimulus S_j, such that if S_i is presented, the probability of perceiving S_j is proportional to W_j^S. The most significant advantage of using such weights is that the metric assumption does not have to be abandoned, as the symmetry condition of a metric space does not get violated. Equation 5.5 can then be modified as follows to reflect the biases in various similarity decisions:

$$P_{ij}^S = \frac{W_j^S f(D_{ij}^S)}{\sum_h W_h^S f(D_{ih}^S)} \tag{5.6}$$

The second question concerns the nature of the distance function and the

[6] As discussed above, a psychological space is a metric space, i.e. a geometric space with a metric.

distance-to-similarity function f.[7] The subject of the relationship between the distance of psychological stimuli and their similarity based on this distance has attracted a considerable amount of research, with the results pointing to different functions for two fundamentally different kinds of stimuli (Nosofsky, 1985). The distance measures are extensions of the Minkowski r-metric, which gives the distance of two points x_i and x_j in an N-dimensional space S according to the following formula:

$$D_{ij}^S = [\sum_{k=1}^{N} |x_{ik} - x_{jk}|^r]^{1/r} \tag{5.7}$$

In this equation, x_{ik} refers to the value of the point x_i in the kth dimension. For $r = 1$, this metric gives the so-called city-block metric, whereas for $r = 2$ it is the Euclidean distance. In experimental analysis, the Euclidean metric was found to best match stimuli with integral dimensions, whereas city-block distance is best suitable for separable stimuli (see e.g. Monahan and Lockhead, 1977, for a direct comparison of different distance measures using the two kinds of stimuli). In the psychological literature, the concept of integral dimensions refers to those which are co-dependent on each other, in that the specification of one dimension requires the specification of the other(s). An example for integral dimensions is brightness and color saturation of a stimulus; these two dimensions, although they can have different values, are perceived as one feature. Separable dimensions are those that can exist independent of each other (Garner, 1974); a mnemonic for finding out separable dimensions is whether one can refer, using natural language, to the different components, without the need for a specialist's knowledge. An example for compound stimuli is circles of different size with spikes pointing from the center in different directions; the size of the circle and the direction of the spike are compound dimensions for such stimuli (used e.g. in the experiments of Shepard, 1964). The distance-to-similarity function best matching these distance functions are exponentials of two kinds. One of these functions is the simple exponential decay:

$$f(D_{ij}^S) = e^{-D_{ij}^S} \tag{5.8}$$

The second function is Gaussian:

$$f(D_{ij}^S) = e^{-(D_{ij}^S)^2} \tag{5.9}$$

[7] One important topic which will not be discussed here is how the dimensions on which the distance function is supposed to operate on are to be determined. With simple stimuli, these dimensions are not so difficult do identify, because they correspond to the relatively few parts or aspects of the stimuli. As the stimuli become more complex, however, a systematic way of determining possible dimensions of the stimulus set is required. The best-known such method is multi-dimensional scaling (MDS), which is a statistical technique for finding coordinates for N objects in a D-dimensional space, once a $N \times N$ matrix of similarities between these objects has been determined by similarity ratings, with $D < N$ (Goldstone and Kersten, 2003). Due to the synthetic approach taken in the work presented here, such an analysis of data for determining dimensions is not necessary; the procession in a machine learning application is not from experimental data to unknown dimensions, but from dimensions already determined by the given raw data to categories.

In various experiments, the Gaussian function has been found to be the most suitable for the Euclidean metric, and exponential decay for city-block metric (Shepard, 1964). The congruence of experimental results on this issue is to such an extent that Shepard (1987) calls the two groups of integral stimuli, Euclidean metric and the Gaussian distance-to-similarity function on the one hand and the contrasting separable stimuli, city-block metric and exponential decay distance-to-similarity function on the other, the "first universal law" of psychological science.

In order to transform the confusion probability function (Equation 5.5) of the similarity choice model into a categorization function, one simple modification which comes to mind is to sum over the members of a category instead of taking similarity to a single exemplar. With this transformation, which is called the *mapping hypothesis* by Nosofsky (1986), the probability that stimulus S_i is classified in category C_J of m different categories ($1 \leq J \leq m$) is given by the following equation (Nosofsky, 1986, p.40):

$$P(R_J|S_i) = \frac{b_J \sum_{j \in C_J} f(D_{ij}^S)}{\sum_{K=1}^{m} (b_K \sum_{k \in C_K} f(D_{jk}^S))} \tag{5.10}$$

Here, $f(D_{ij}^S)$ refers, as above, to the similarity of stimuli S_i and S_k based on distance D_{ij}^S in stimulus space. Capital indexes refer to categories (as in C_J, C_K and R_K), and lowercase indexes refer to individual stimuli (as in S_i, D_{ij}^S and D_{jk}^S). The reaction R_J distinguishes not the individual stimulus S_j, but the category C_J, with $J \in \{1, 2, \ldots, m\}$. The bias parameters b_J are also attributed to the categories and not to the individual exemplars. Equation 5.10 forms the basis for an extension of the context model, called the generalized context model (GCM) proposed by Nosofsky (1986). One important difference of the GCM is that it aims to replicate the effect of selective attention to different dimensions by integrating further bias parameters in the Minkowski metric which is used to calculate the distance D_{ij}^S of the stimuli S_j and S_i:

$$D_{ij}^S = c[\sum_{k=1}^{N} w_k |x_{ik} - x_{jk}|^r]^{1/r} \tag{5.11}$$

The conditions on the parameters are that $0 \leq c < \infty, 0 \leq w_k \leq 1$ and $\sum w_k = 1$. The bias parameters w_k are included to model context effects in similarity judgements. Through these parameters, the stimulus space is warped or extended in the dimension of the parameter which regulates it. For example, when the subjects are paying attention to the color dimension of the experimental stimuli (because it is more salient, the subjects have been primed through other stimuli etc.), the bias parameter for the color dimension is increased, causing the stimuli which would have been judged as similar due to similarities in the other dimensions to drift apart. Since the bias parameters add up to one, attention to one dimension happens at the expense of other dimensions. The scale parameter c serves to reflect overall discriminability in the psychological

space; it increases, for example, with increases in stimulus exposure duration, or as subjects gain more experience with the stimuli. As an improvement over the context model, the generalized context model allows continuous values for stimuli. Also, once a multidimensional scaling solution is derived for the stimuli, their similarity is less ad hoc compared to the context model. Evaluating the experimental data, Nosofsky (1986) claims that "one could make excellent predictions of categorization performance given knowledge of performance in an identification paradigm" (p.53). The significance of the GCM is that it further establishes the viability of using individual experiences for categorizations, with strong theoretical support for the idea of a psychological space and the role of similarity in categorization.[8]

As the aforementioned quote from Shepard (1957) pointed out, the similarity choice model was not concerned with how learning can take place, but the performance of individuals at any given stage. The context model and GCM, on the other hand, offer a glimpse at how categories can be learned, because they model categorization on the basis of experience with exemplars. Since these exemplars act as access points to the categories, their instantiation and processing in category decisions can be the basis for learning categories. One example of an attempt to extend the exemplar approaches of the context model and the GCM is that of Kruschke (1992), who presented an exemplar-based neural network model of categorization, called ALCOVE.[9] This model combined exemplar-based representation with error-driven learning. The neural network in the ALCOVE model is a feed-forward network with three layers. An input layer with N nodes encodes the stimuli, with one node for each of N dimensions. One hidden layer includes nodes which represent the training exemplars. A new node is placed in the hidden layer with every new training stimulus. The activation function of these hidden nodes is virtually identical to the similarity function of GCM. The distance of the jth hidden node with dimension values $(h_{j1}, h_{j2}, \ldots, h_{jN})$ from the input vector $\alpha^{in} = (\alpha_1^{in}, \alpha_2^{in}, \ldots, \alpha_N^{in})$ is computed as follows (Kruschke, 1992, p.23)[10]:

$$d(h_j, \alpha^{in}) = [\sum_i \alpha_i |h_{ji} - \alpha_i^{in}|^r)]^{q/r} \tag{5.12}$$

As in GCM, attention to different dimensions is modelled through the weights α_i. In this equation, q and r specify the kind of distance metric (city-block versus vs. Euclidean) and similarity gradient (exponential decay vs. Gaussian). The stimuli in the experiments with which the model is compared are separable, leading to $r = q = 1$. The activation value a_j^{hid} of the jth hidden node is the

[8] Exemplar models which are based on featural representation of compound stimuli can, under certain limiting assumptions, be shown to be formally identical to or a superset of many other categorization approaches in psychology; see Nosofsky (1990) for a thorough comparison, especially the figure on p. 413.

[9] The GCM was an influential model which lead to many similar models with modifications; see e.g. Aha and Goldstone (1992) for another extension of the GCM.

[10] As the following explanation and equations demonstrate, the ALCOVE model has a close affinity to what are called radial basis function networks in neural network research (Moody and Darken, 1989).

inverse exponential of this distance with an additional scaling parameter called the specificity of the node:

$$a_j^{hid} = e^{-c \cdot d(h_j, \alpha^{in})} \tag{5.13}$$

The most important difference of ALCOVE in comparison to the GCM is in the activation function of the output nodes. Activation of the output nodes is a sum of the activation of the hidden nodes, with the sum controlled by a weight between each hidden node and output node (Kruschke, 1992, p.24):

$$a_k^{out} = \sum_j w_{kj} a_j^{hid} \tag{5.14}$$

Learning in ALCOVE is achieved by changing the weights w_{kj}, along with the attention weights α_i, according to an error-driven learning rule. In experimental comparisons, ALCOVE has proved to model human data as well as GCM and the context model, and also explained some other phenomena not accounted for by the other models, such as base-rate neglect and transfer performance (Nosofsky et al., 1992). As it can be seen from this short overview, the learning model of ALCOVE introduces more parameters to a model already with a parameter for each dimension, and further parameters for each exemplar. The main reason for such a heavy reliance on parametrization is that the exemplars are taken to refer individual stimuli, which means that the subjects are assumed to first recognize these, and then make a category judgement. The fundamental aim of the similarity-based approach to categorization and identification, avoiding category-level knowledge and relying only on information from individual episodes of experience, is thus not completely fulfilled when exemplars require identification of stimuli. An alternative way of implementing learning is to include each episode of experience with a stimulus as a separate exemplar, thereby avoiding the necessity for recognizing a set of sensory data depicting a certain stimulus. Such an implementation can rely on the theory of categorization based on geometric metaphors as depicted in this section, and at the same time learn through data accumulation by embodied interaction and low-level similarity in a psychological space. A learning mechanism based on these ideas will be presented after a discussion of the role of similarity in theories of categorization.

5.1.1.1 Similarity

As it can be seen from this short overview of exemplar-based models in the psychology of categorization, such models rely on storage of sensory data and similarity for explaining categorization performance. The use of similarity to ground theories of categorization has been criticized as both being too general and too specific (Goldstone, 1994). Similarity is too general in the sense that it is context-dependent, and the respect in which two things are similar has to be specified in order to use it as an explanatory construct: this practically makes the idea of similarity a placeholder (Goodman, 1972). On the other hand, it is

not general enough, because categories depend on other kinds of knowledge such as theories of the domain (Rips, 1989). However, as Goldstone (1994) argues, it is not possible to expect similarity to ground all kinds of categories, but only the most basic ones:

> Neither *similarity* nor *category* is a unitary construct – there are variations of each that are importantly different. Similarity cannot ground all category types. Still, the class of categories for which overall similarity provides a partial account are an important class because of their wide inductive potential (p.151).

Furthermore, there are certain kinds of similarities whose robustness have been proven in psychophysical experiments. These similarities, in experiments designed to override them through category membership rules that contradict them, have proven to interfere with the categorization task (Allen and Brooks, 1991). [11]

In cognitive psychology and AI, there is a common distinction between rule- and similarity-based processing (Smith and Sloman, 1994). This distinction can be seen, as Hahn and Chater (1998) argue, as one that relies on the kinds of representations used in processing: similarity relies on partial matching of information which has not been abstracted, whereas rule-based processing involves the strict matching of highly abstracted information.[12] This abstraction process corresponds, as it has been argued in Chapter 2, to the extraction of features from a linguistic description of the stimuli, and any proposed learning algorithms are assumed to function on these features that correspond to words in a description. In the above mentioned series of exemplar-based psychological models of categorization (Kruschke, 1992; Nosofsky, 1986; Medin and Schaffer, 1978), these features are the dimensions of shape, size and color. These dimensions are clearly distinguished from each other, and processes of similarity are assumed to function on these separated features. As it was argued in Chapter 2, this gives a privileged position to the descriptions humans use, disregards the distinction between the description of a process and the generative procedure that underlies such a process and, most important of all, delegates perception to a secondary status by assuming a kind of pure perception which delivers features to a central processing mechanism. This criticism points to an important advantage of the use of machine learning techniques which rely on raw data and low-level similarity mechanisms, instead of pre-processing to extract linguistically significant features: These methods lead to more parsimonious models, in the sense that the intuitions of the designer are not built in. Furthermore, in the context of the research presented here, the aim is to understand how symbolic

[11] See Medin et al. (1993) for a theory of similarity which picks structure as the foundation of similarity judgements. The criticisms made in the following paragraph on this page are relevant also in the case of structure: without understanding the processes through which structure is perceived by human beings, assuming their importance in cognitive processes simply gives precedence to linguistic descriptions.

[12] For a recent article questioning the rules versus similarity distinction, and arguing for commonalities rather than differences in processing, see Pothos (2005).

representations arise in the first place, and how a group of agents come to create a lexicon with which such a description can be made. Therefore, it would be futile to start with such given features.

Further support for both similarity-based and exemplar-based learning comes form detailed studies of the visual performance of hens. Werner and Rehkämper (1999) used integral stimuli in a visual discrimination task in order to test whether chicken responded to the relevant dimensions of the stimuli or remembered complete stimuli, including additional cues which were uncorrelated with reinforcement. A number of effects pointed to the second of these two possibilities. The first of these results was obtained when the chickens, which were trained with a set of 9 pairs of stimuli, were switched from this set in which they were trained to another set of similar stimuli which varied in the same dimensions, without changing the reinforced dimension. After this switch, the performance of the hens, which had reached 90% correct in the first trial, dropped below chance level. If the chickens were abstracting away the relevant dimension, such a drop in performance would not have been observed. In order to further test whether chicken were abstracting features, a bigger stimulus set was presented, and then test trials carried out with a different set of stimuli. In this experiment, it was found out that the size of the training set did not effect the performance in the final testing trial, the opposite of what would be expected if features were extracted by chicken. In a last experiment, the chickens were trained with the whole stimulus set for a high number of trials, and the dependency of their performance on sets of features on various dimensions was analyzed statistically. It was found out that "The discrimination behavior of every hen, even the better learners, depended on cues that were not correlated with reinforcement. Additionally, there was a considerable bias towards the use of relational information between elements of the whole stimulus array" (Werner and Rehkämper, 1999, p.35).

In a further extension of this study, Werner and Rehkämper (2001) picked the 7 best performing hens and tested them in more experiments with the same stimuli. One particular aim of the study was to compare the predictions of feature, exemplar and prototype-based theories with the actual performance of the chicken. One important outcome of the experiments was that, as in the first group of experiments, instead of the abstraction of a single feature being the cause of learning, the animals appeared to respond to various combinations of these features, despite only one such feature predicting reinforcement. Due to the presence of this effect with each hen in each experiment, the authors argue that this fact points to not the result of a side effect, but a direct consequence of the differential retrieval of learned exemplars from memory. This conclusion is further supported by statistical analysis of performance data, and comparison of the results of this analysis to the predictions of various theories of categorization. A principle component analysis (PCA) of the pecking rates of the chicken revealed that the exemplar-based approach was the one that fit the data the best: "Even though the categories to be discriminated by the hens were well-defined, the best fit to the data was not the feature-based or elemental approach, although it is often employed in the explanation of simpler discrim-

ination tasks [...] Exemplar theory was shown to account better for the data of this multidimensional categorization task." (Werner and Rehkämper, 2001, p.45).

5.1.2 Machine learning aspects

In the context of robotic applications of machine learning methods which process noisy and low-level sensory data, the problem of machine learning is traditionally stated as one of function approximation; given data, assume that there is a function which maps it onto the desired values, and try to compute this function. In the statistical machine learning literature, two main families of methods can be distinguished to solve this problem. One family is that of model-based learning, which matches the variables in the input space, through coefficients, to a polynomial function approximating the output function. The best-known method of learning used in this family of methods is the least squares method, where the coefficients in the polynomial function are picked to minimize the *residual sum of squares*, a measure of the error of the approximation. The second family of learning methods is that of nearest neighbors methods. These methods are characterized by their reliance on a metric, similar to the ones used in psychological research on categorization, that decides which sample points in the training data are the closest to the test point, and combining the values of these points with their distance to evaluate the test point.[13]

The main difference between model-based learning and nearest neighbor methods concerns the assumptions they make about the data to be learned from. The use of models requires deep assumptions to be made about the data, since the representation of the data has to be adapted to the model. On the other hand, these methods generally lead to models that perform well with test data independent from the learning data. Nearest neighbor methods, on the other hand, do not necessitate any assumptions about the data, except for a similarity/distance measure. They also do not make any assumptions about the shape of the boundaries between categories. As Hastie et al. (2001) point out, linear models "have low variance and potentially high bias", whereas nearest neighbor procedures have "high variance and low bias" (p.16).

As it is apparent from the preceding review of psychological literature on exemplar-based learning, there is considerable support for a theory of categorization which does not rely, at least for a theory of basic categories, on abstracted information and strict feature-matching. The computational properties of nearest neighbor methods provide further reason to prefer such an approach. The main advantage of an exemplar-based learning mechanism is that it does not rely on abstracting mechanisms and features. The representation of sensory data can thus be left to the specifics of the sensory apparatus, as long as comparison of different sets of such data is still possible. A further reason to prefer a lazy-learning approach such as exemplar-based learning is

[13] See Chapter 2 of Hastie et al. (2001) for an overview. It should also be mentioned that these two machine learning methods are simply the most basic cases, serving as the starting point of further research; both methods have been extended and combined with each other.

that the sensory-motor data in the memory is reevaluated with each episode. What was called the transactional perspective in Section 3.2 involves the coordination of different levels of cognitive processing in the current task context; such recoordination, argued to be indispensable for properly understanding human intelligence, is possible only if previous experiences are brought to bear on the current situation, without having to use the same context-independent abstractions as before. From this perspective, what is traditionally treated as a limitation can be seen as a possibility to implement an alternative vision of computational modelling.

One further advantage of exemplar-based learning is that it is especially suitable for studying reinforcement learning. Reinforcement learning is a statement of the machine learning problem in terms of maximizing a reinforcement signal. An external signal, which specifies a desirable behavior or property, is given to a learning system as a reinforcement signal, whereby this system tries to increase the value of this signal by modifying behavior (Sutton and Barto, 1998). Reinforcement learning can be understood in contrast to supervised learning, in which the agent doing the learning is provided with the error between its response and the ideal response. Reinforcement learning is obviously more realistic compared to such ideal conditions; living beings rarely receive perfect feedback on their actions, if they ever do. Another advantage of reinforcement learning is that it gives an agent the chance to discover an environment, devising its own way to increase received reinforcement; in this sense, reinforcement learning is akin to genetic algorithms. It must be pointed out that reinforcement learning is a particular setup for the machine learning problem, and does not confine the solution to any particular method.

A diagram of the reinforcement learning problem can be seen in Figure 5.1. An important feature of the reinforcement learning problem, which is common with most of the other methods and algorithms in machine learning, is the relationship of the execution of behavior and the input from and output to the environment: The perception of the state is one time step in the behavioral flow, the deliberation of the next move is the second, and the execution of a behavior is the third.

Kernel methods

One other thing to mention about exemplar-based learning is the affinity of this methodology to a recently developed set of methods in machine learning, so-called kernel methods. Kernel methods refer to a family of machine learning methods which use a similarity kernel to map input data into a higher dimension in order to find out regularities which cannot be elegantly expressed and efficiently calculated in a lower dimension. Kernel methods are well-studied in the statistics and machine learning literature, and certain kinds of kernels have been found out to possess favorable features for studying categorization in psychology (Schölkopf and Smola, 2001). The most important result regarding exemplar-based learning algorithms is that in order to find the optimal solution to a machine learning problem in a higher dimension, it is not necessary to trans-

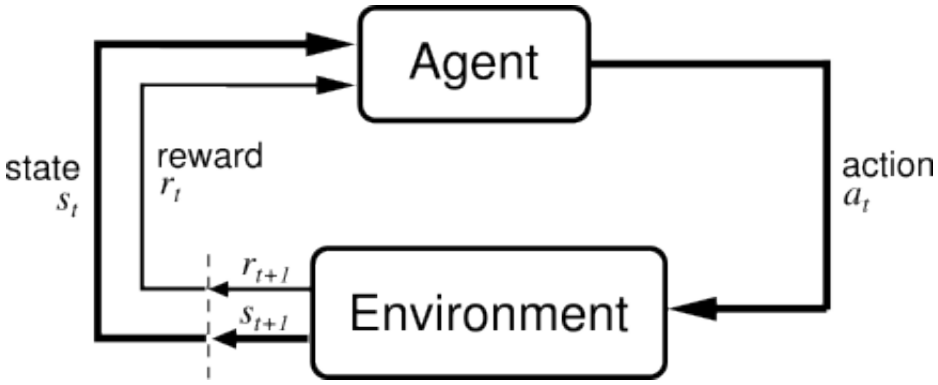

Figure 5.1: The reinforcement learning problem. Adapted from Sutton and Barto (1998).

form the data into this higher dimension, a computationally costly operation. The similarity of exemplars can serve as the dot product in a higher-dimensional vector space, allowing an optimal solution to be formulated as a linear combination of kernel functions centered on the exemplars (Jäkel et al., 2007).[14]

5.2 Categorization and similarity

In order to study the grounding of communicative symbols in behavioral categories, what is needed first is an environment in which such categories can be acquired, and afterwards perceived and acted upon in order to engage in communication. Since the language games framework, discussed in Section 4.2.1.1, has been accepted as a suitable framework to study lexicon formation, this environment must present choices to the agents about which they can communicate. These constraints have led to the construction of a simulated experimental environment, depicted in Figure 5.2. An environment frequently utilized in psychological experiments with animals, the so-called Y-maze, has been adopted with the modification that the two different choices are color-coded as red and blue, in order to facilitate visual discrimination. The Y-maze environment presents a simple setup for physical interaction with two choices for the agent. The most important advantage of such a simple maze is that making a choice corresponds not simply to outputting a symbol or a label denoting a choice, but engaging in a certain kind of bodily activity. The agent is a simulated robot which is able to move freely in this environment; it is modelled after a Khepera robot, and can control its movement by setting the motor speeds of the two wheels on its

[14]For an overview and mathematical explanation of the many commonalities between frequently used methods in the psychology literature, such neural networks and exemplar-based methods and kernel methods see (Jäkel et al., 2008). Jäkel et al. (2008) also establish the special properties of the Euclidean distance measure, or similarity kernel.

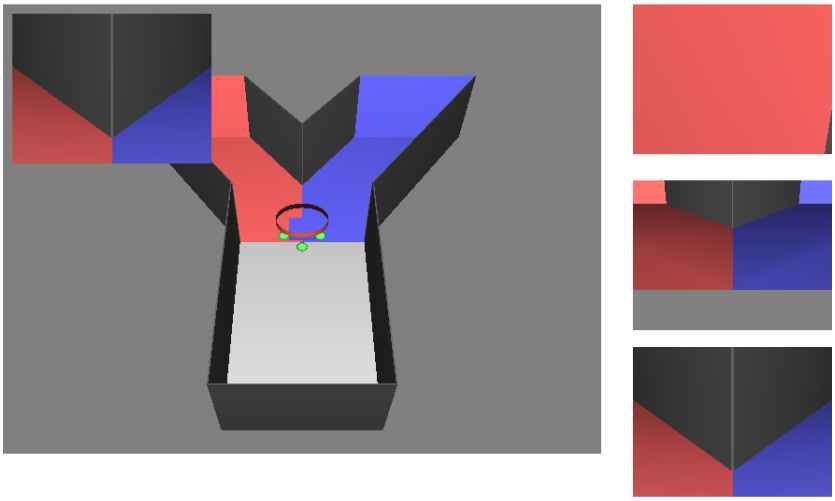

Figure 5.2: The simulation environment. On the left hand side is the environment as seen on the computer screen. The small window on the top left corner shows the image from the simulated camera on the the agent. On the right hand side are three sample images taken by this camera.

left and right sides. Perceptual data in the form of images is obtained through a simulated camera placed on top of the simulated robot, similar to the visual turret extension of the Khepera robot.[15]

As the preceding discussion should have made clear, exemplar-based approaches allow the development of categorization mechanisms which can operate on low-level sensory data, as long as this data can be represented in a psychological space with a distance measure. This property makes the exemplar approach particularly suitable for an application which aims to model behavioral categories without abstractions and extraction of features. In the psychological applications of the exemplar-based approach, an exemplar referred to an identifiable stimulus; carried over to a machine learning application, this aspect becomes a limitation. Also, deriving the identity of objects from sensory data and using the representations resulting from such a process would also go against the situated AI approach, because this would correspond to building a world model from perceptual data. A better alternative is to use the raw data from each episode of experience for learning, without distinguishing which object or case the agent faces. The exemplars, whose purpose is to represent such an episode in memory, thus have to include the following elements:

[15]The code for the simulation environment and the control algorithm can be downloaded from the following URL: `http://www.cogsci.uos.de/~utuerkme/disscode.zip`.

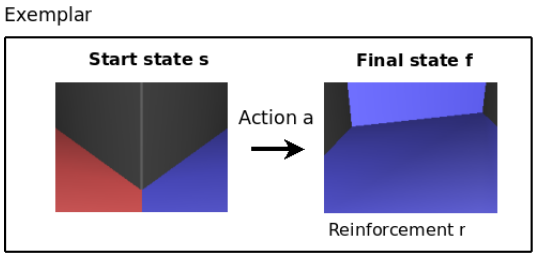

Figure 5.3: The contents of an exemplar

1. The start sensory state s
2. The action $a \in A$ undertaken by the agent
3. The final sensory state f
4. The reinforcement r

For an application which functions in the environment explained earlier, visual data from a simulated camera is an ideal representation of the situation the agent finds itself in. Therefore, the start sensory state s and the final sensory state f consist of image data from the simulated camera. It has to be mentioned here that the sensory states in the exemplars (simply referred to as states form here on) do not constitute the contents of an agent's memory in the traditional sense. The mechanism explained here is intended to implement situated representations, which then consitute the contents of memory used for communication, planing and other similar high-level capabilities. Furthermore, the task at hand can be accomplished using only the visual data, but the inclusion of solely visual data should not be taken as a claim that sensory data form a single source, or only external sensory data, is sufficient for all kinds of embodied tasks; this topic is discussed further in Section 6.3.

One speed value for the left motor and another for the right make up an action a. The set of all behaviors A includes a number of different such actions. For operation in the experimental environment explained here, four behaviors were sufficient. These behaviors are going forward with two different speeds (the motor values $[10, 10]$ and $[8, 8]$), turning left (motor values $[1.5, -1.5]$) and turning right (motor values $[-1.5, 1.5]$). The execution of a behavior is the setting of the speeds of the two wheels to the values given by the behavior, waiting for 1.2 seconds, and setting both motor speeds back to 0. The contents of an exemplar $e = (e_s, e_a, e_r, e_f)$ are depicted in Figure 5.3, with two sample images from the camera. The aim of the agents in the experiment is to collide with one of the red or blue walls, depending on the the experimental stage which is being carried out (as explained below), and avoid the other walls. The reinforcement is determined according to which wall the robot collided with, as detected by the simulation environment.

The image data is represented as RGB images of size 42 pixels width by 32 pixels height. Each image pixel takes three bytes, one byte corresponding to one channel of red, green or blue. The distance of two pixels p_i, p_j, with values $p_{i,c}, p_{j,c}$ in channel $c \in \{R, G, B\}$ is calculated as follows:

$$d(p_i, p_j) = \sqrt{\sum_{c \in \{R,G,B\}} (p_{i,c} - p_{j,c})^2} \tag{5.15}$$

Before the calculation of this distance, the three different channel values of RGB are averaged (i.e. $R = R/(R + G + B)$) in order to compensate for brightness differences. One can thus say that it is the ratios of the different channels being compared when pixels are compared, and not the individual byte values. On the basis of the pixel distance, the distance of two images s_a and s_b is calculated as follows:

$$d(s_a, s_b) = \frac{\sum_{j=1}^{32} \sum_{i=1}^{42} d(s_{a,ij}, s_{b,ij})}{1344} \tag{5.16}$$

In this equation, $s_{a,ij}$ refers to the pixel in position (i, j) in s_a, i being the horizontal coordinate, and j the vertical coordinate. The total distance of the individual pixels is averaged over the total number of pixels in an image, $32 \times 42 = 1344$. The similarity of two states is consequently computed from the distance as follows:

$$sim(s_a, s_b) = e^{-\mu d(s_a, s_b)} \tag{5.17}$$

In Figure 5.4, sample images and their similarities are displayed. Using this similarity and a memory of exemplars containing data on which actions was undertaken in which situation, and the reinforcement received as a result and the final state, it is possible to calculate feedback values for individual actions. At the beginning of an experiment, a memory is created, with an exemplar for each behavior, and randomly generated images as start and final states for these exemplars. The aim of these random exemplars will be made clear below. The memory is a simple list of exemplars without further internal structure, and during the progress of the experiment, new exemplars are simply appended to it. The feedback for a behavior $a \in A$ given camera image s_c, denoted by $F(a, s_c)$, is calculated by summing the product of the reinforcement e_r of exemplars e in memory which have the action a, with the similarities $sim(s_c, e_s)$ of the start states of these exemplars e_s with the current situation s_c:

$$F(a, s_c) = \sum_{\{e \in M | e_a = a\}} sim(s_c, e_s) \cdot e_r \tag{5.18}$$

Once feedbacks for individual behaviors are calculated, probabilities for behaviors are computed using the Boltzmann distribution:

$$P(a \mid s_c) = \frac{e^{F(a,s_c)/\tau}}{\sum_{b \in A} e^{F(b,s_c)/\tau}} \tag{5.19}$$

Using this formula, a probability distribution is calculated, with a probability for each behavior, and the probabilities adding up to one. The reason for using

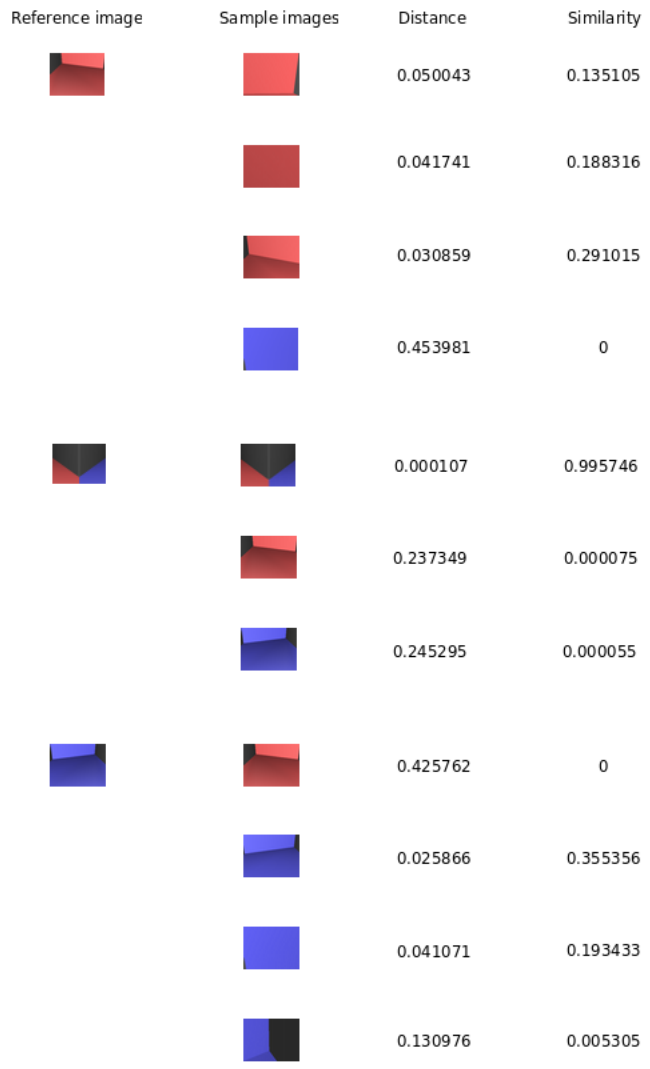

Figure 5.4: Sample distances and similarities, calculated using Equation 5.16 and Equation 5.17, with $\mu = 0.025$.

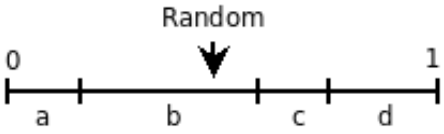

Figure 5.5: Using a random number to pick a behavior

the Boltzmann distribution is that, on the one hand, negative feedback values are mapped to positive probability values, albeit approaching zero with lower values, and on the other hand, the exponential function is increasing, so that with more positive feedback, the execution probability of a behavior increases, approaching unity. The τ parameter, traditionally called the heat parameter, can be used to control the speed of learning; for smaller τ values, an increase in feedback leads to a fast increase in probability, leading the learning agent to converge quickly on positive choices explored early in the interaction. The τ parameter can thus be used to balance exploration and exploitation. In the task for navigation, $\tau = 20$ was found out to provide a good balance of these two factors.

Through the described process, behaviors are mapped to probability values, which reflect the feedback for the behavior in the given state. In order to pick a behavior using this information, a random number between 0 and 1 is drawn. Afterwards, the probability values are added progressively, picking the first behavior with which the probability sum exceeds the previously drawn random number. See Algorithm 1 for an algorithmic explanation of this process. Another way of thinking about this action selection mechanism is that the actions are laid on a line segment from 0 to 1, the length of a segment for a behavior being proportional to the probability of the behavior. The behavior into whose segment the random number drawn falls is picked for execution, as depicted in Figure 5.5. This way of balancing exploration with exploitation by using the Boltzmann distribution to map activation values to probabilities adding up to one is called the softmax decision rule (Thrun, 1992). The method of selecting an option using probabilities mapped to an interval and a random number picked from that interval is called roulette wheel selection.

Algorithm 1 Action selection algorithm

```
X ← 0
R ← random value between 0 and 1
for each behavior a and its execution probability do
  Add execution probability to X
  if X > R then
    Return behavior a as the behavior to be executed
  end if
end for
```

One of the most important limitations of the approach presented here is that the behaviors are hand-coded. As explained above, the behaviors which are available for the Y-maze are one for turning right, one for turning left, and two for moving forward. The behaviors for turning right and left have the same motor speeds, only with different signs. The motor speeds for these two behaviors were set so that, when they were executed, the robot faced the red or the blue wall directly, so that the execution of the going forward behaviors would suffice to receive positive reinforcement.

Using the tools presented until here, associative learning, i.e. coupling of a certain behavior to similar perceptual situations, can be implemented. Learning is a result of the accumulation of exemplars; as the agent tries out different behaviors, the correct behaviors in various situations will receive more positive feedback, whereas the incorrect behaviors will incur negative feedback. Positive and negative feedback will lead to higher and lower probabilities, respectively, and the behavior with higher probability will get executed, leading to even more feedback, and higher probability. This positive feedback process causes the agent to rapidly home in on a probability close to 1 for the right behavior, for the situations in which there is such a behavior. The condition for this convergence is the visual distinguishability of the situations, i.e. that the similarity of states which require different behaviors is spread apart sufficiently. In the autoshaping stage, which will be explained later, the agent is placed facing the red or blue wall. It takes the simulated agents on average 30 trials until the execution probability of moving forward is virtually 1.

An important complication, however, appears once we consider how the agent is to make a decision in a situation which does not offer immediate positive feedback, but from which the agent can move into a position which can lead to positive reinforcement. The agent should, in such a situation, base its decision not solely on the reinforcement which would derive from the immediate application of a behavior, but additionally on the possibility that the behavior would bring the agent to a favorable position. In the opposite case, where the agent makes decisions only based on the feedback which can be derived from the current situation, it would be impossible for an agent in the intersection point of the Y-maze in Figure 5.2 to narrow in on an action for turning right or left, since these actions do not lead to direct positive reinforcement, but are the first step in a series of movements which would lead to positive reinforcement. The problem, to put it from the standpoint of learning a sequence of actions to be undertaken for achieving positive reinforcement, can be seen as the question of how to impute the reinforcement received to the various behavioral steps in the sequence. This problem is called the credit assignment problem in AI and machine learning (Minsky, 1961). In the particular setup used here, visual similarity of the consecutive states leading to positive reinforcement offers a relatively simple way of solving the credit assignment problem, which is implemented as follows. Among the exemplars in the memory, ones with zero reinforcement are picked out. The final states of these exemplars are aligned with the start states of the exemplars which have nonzero reinforcement. The feedback contribution of this combination is then calculated as the reverse ex-

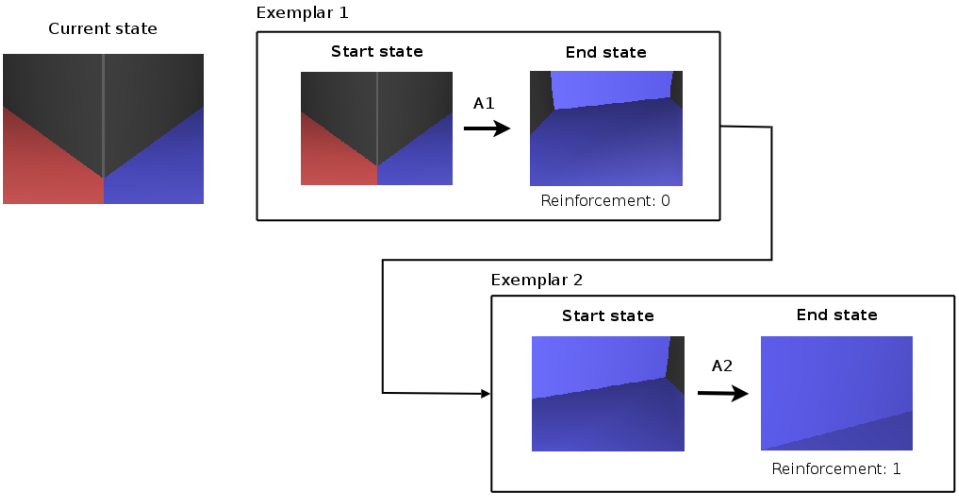

Figure 5.6: The chaining of exemplars

ponential of the sum of distances, multiplied by the reinforcement of the second exemplar. This feedback contribution is called secondary feedback, because its role is secondary, compared to the contribution of direct feedback for immediate action choices. The total secondary feedback $SF(a, s_c)$ for behavior a in situation s_c based on memory of exemplars M is thus calculated as follows:

$$SF(a, s_c) = \sum_{e \in M | e_r = 0, e_a = a} \; \sum_{g \in M | g_r \neq 0} g_r \cdot e^{-\mu(d(s_c, e_s) + d(e_f, g_s))} \tag{5.20}$$

As it can be seen from this explanation, the core idea in this process is to chain the exemplars and add their distance to arrive at an extended similarity, corresponding to the separation of the end state from the current state in the psychological space. This extended similarity is denoted in Equation 5.20 by $e^{-\mu(d(s_c, e_s) + d(e_f, g_s))}$. In this computation, one more behavioral step enters the behavioral trajectory to be followed by the agent. For computational reasons, chaining of exemplars was done only for two steps, because for a chain of length n with a memory of M exemplars, comparisons on the order of M^n are necessary. The chaining process is depicted in Figure 5.6. The state images in this figure, taken from actual experimental exemplars which followed each other in the movement of the agent, demonstrate the role of similarity in solving the credit assignment problem in this setup. When the robot turns right to face the blue end of the maze, the final state of the exemplar created for this movement has a high value of blue channel. Once the agent faces the blue wall, it will continue forward towards the end, and receive positive reinforcement after a few forward motions. The reason the agent moves forward when it faces the blue wall is primarily that when it is close enough to the wall, in which case

the blue channel is very high, moving forward causes positive reinforcement, and the accumulation of these exemplars with positive reinforcement causes the agent to move forward in cases where it has dominantly blue in the state image. The next time the agent faces the situation in which it has to turn right to face the blue wall, the final state of the exemplar formed in the former execution of the turning behavior will also have a high value for the blue channel, and this color will be matched with the start states of the exemplars which have been collected by driving against the blue wall. This way, the agent can engage in a series of actions which will end in positive reinforcement.

As the description of the experimental environment and the specific learning problem presented in this chapter make obvious, one of the standard learning algorithms used very often in reinforcement learning, Q-learning, can be considered as a viable choice for solving the learning problem presented here. The Q-learning algorithm (Watkins, 1989) principally involves the calculation of expected maximum reward for different actions in a reinforcement learning environment by updating the expected reward for a behavior in a state. The estimated value of action a in state s at time t, denoted as $Q(s_t, a_t)$ is updated as follows:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \tag{5.21}$$

This update involves the use of the reinforcement received upon execution of a action a at time $t + 1$, r_{t+1}, and the calculation of maximum possible further feedback $\max_a Q(s_{t+1}, a)$ that can be achieved from state s_{t+1} in which the agent will end if an action is carried out.

There are a number of difficulties which make the strict application of the Q-learning algrithm difficult in our scenario. The first of these is that the states here do not belong to clear-cut categories, making it difficult to attribute reward values to behaviors given a state based on a calculated maximum value for that state. Furthermore, the fact that a lazy learning method is implemented makes it difficult to buid in a mechanism whereby behavior values are updated at each step after the sensing of the final state. The fundamental idea of the Q-learning algorithm –calculating the value of an action in a given state based on the expected future reward– is nevertheless implemented in a different way through the inclusion of the chaining mechanism explained above. The maximum reward for a behavior, here called feedback for a behavior but in effect serving the same purpose, is updated when a decision has to be made, i.e. in a lazy manner. Also, each new sensory state is treated as a different state, determining feedback values for behaviors individually.

The calculation of feedback for a behavior is presented in algorithmic form in Algorithm 2. To summarize, learning is a result of the accumulation of exemplars which the agent then compares with the given situation, ordering behaviors according to their probability and using a random number to pick one. The feedback for behaviors is gathered from the reinforcement of the exemplars which are similar to the current situation and have nonzero reinforcement, multiplied with their similarity of the start states of these exemplars to the given

situation, and in addition from the chaining of exemplars, so that feedback is derived also from other end points in the memory.

Algorithm 2 Algorithm for calculating feedback for a behavior

Set total feedback for each behavior in A to zero
$s_c \leftarrow$ Current state from camera
for each behavior a in A **do**
  for each exemplar e in the memory with $e_r \neq 0$ and $e_a = a$ **do**
    Add $sim(s_c, e_s) \cdot e_r$ to the feedback for behavior a
  end for
  for each exemplar e in the memory with $e_r = 0$ and $e_a = a$ **do**
    for each exemplar g in the memory with $g_r \neq 0$ **do**
      Add $e^{-\mu(d(s_c,e_s)+d((e_f,g_s)))} \cdot g_r$ to the feedback for behavior a
    end for
  end for
end for
Calculate execution probabilities for each behavior according to the softmax rule.
Draw a behavior for execution

5.3 Language Games

The fundamental idea of language games as an experimental setup to study communication phenomena between artificial agents has been explained in Section 4.2.1.1. The setup used in the work presented here is very similar to that used by Steels, in that agents with categorization capabilities engage in episodes of activity in which they use labels to refer to objects. The main difference is that these categorization capabilities are not based on the splitting of sensory channels, but on behavioral categories, and the outcome of a round of interaction is signalled with the successful completion of a task. Partly due to these differences, the experiments consisted of four phases, which will be explained now (see also Türkmen and Zugic, 2008).

First step: Autoshaping In the first stage of the experiment, all the agents go through an autoshaping phase. The aim of this phase is to teach the agents the various possibilities in the environment. In the setup depicted in Figure 5.2, this consisted of the agents being placed in front of one of the walls with the red and blue colors and getting positive reinforcement for driving into those walls, and negative reinforcement for the black walls. This phase ended once the probability for driving into the designated wall reached 1 (i.e. the underlying software could not distinguish between 1 and the actual probability when the probability was displayed in text form). Autoshaping was carried out with both red and blue side, and took on average 30 trials for each color.

When an agent is created and placed in front of one of the target walls (that is, red or blue end walls), it starts with a memory of one random exemplar for each behavior, with randomly generated start and final images. Randomly created here refers to images of the same size as those generated by the simulated camera, but whose pixel RGB values are randomly determined. In the absence of the random exemplars, e.g. if simply a random behavior is drawn when the memory is empty, the feedback for the randomly drawn behavior is nonzero, whereas those for the other behaviors is zero. This leads to this behavior having a high probability, and getting executed continuously. To avoid the robot getting stuck in this state, and ensure initial exploration of the environment, each agent starts with the random memory, which ensures an initial positive probability for each behavior. Such exploration is necessary not only in the autoshaping phase, but also in the second step, where the agents have to find out that there are two possible choices at the intersection.

Second step: Learning to make choices In the second step, the agents are placed at the intersection of the red and blue paths. Positive reinforcement is given for bumping into the red or blue walls, negative reinforcement for every other wall. The criterion for finishing this step is that the execution probabilities of two behaviors are higher than the others and close to each other. The feedback for these behaviors is a result of the chaining of exemplars, with the exemplars with positive reinforcement from the first step providing connection to the blue and red target walls through similarity. When agents are trained in this setup, they learn to either go right or left, with the probabilities of the two behaviors close to each other and their sum close to 1. This is the state from which the agents are integrated into language games.

Algorithm 3 Behavior of an instructor in the training before a language game

```
Considered label ← Null
Labels already considered ← Empty set
while instructor is being trained do
  Behavior choice & applicable labels ← Evaluation of the situation
  if (Considered label = Null) then
    if There is a label L which is among the applicable labels and not in the
    set of labels already considered then
      Considered label ← L
      Add L to the set of labels already considered
    end if
  end if
  Carry out selected behavior and pick reinforcement
  if (reinforcement < 0) and (Considered label ≠ Null) then
    Considered label ← Null
  end if
end while
```

Third step: Training of an instructor In the third step, one agent that has gone through the first two steps is picked randomly as an instructor. This agent is then reinforced to make one of the choices in the task space. The procedure is the same as in the first step, except that only one of the choices is reinforced, and the robot starts from the intersection point, and not facing one of the walls. This step continues until the agent makes the right decision (that is, the path that is being reinforced in this training stage) a certain number of times in a row. This agent will serve as the instructor in the following language game. As it will be explained later, the number of consecutive runs in which the instructor had to make the same choice – the instructor success threshold – had a significant effect on the number of errors made and the emergence of a vocabulary.

One means available to the agents, and that is employed by the instructors and the students, is the accumulation of the exemplars which contribute to the different decisions for the detection and application of a label. The exemplars which are accumulated are the ones which have positive reinforcement, and the distance of their start state from the current state (in the case they are the second in a chain, the distance of their start state from the end state of the first exemplar in the chain, plus the distance of the start state of the first exemplar from the current state) is under a threshold. In the experiments, such a threshold of 0.12 was found to be functional. Because the decisions are made from a position which did not lead to any positive feedback with one single movement, these exemplars are mostly those that are second in a chain of two, with the first exemplar containing a turning movement. A label is added to the set of applicable labels in a situation when it is attached to more than half of the accumulated exemplars for a behavior. It should be mentioned that, due to the interplay of the memory and communication mechanisms, the labels were generally attached to all of the exemplars accumulated for a specific behavioral decision. The agents use this exemplar accumulation and label determination mechanism to find out the applicable labels for use in three different situations:

- For transmission to the interlocutor
- For interpretation of a received label
- For the coordination of the individual behavior

Another mechanism necessary for using a label is what is here called "considering" it, i.e. using it for coordinating behavior, be it for finishing the instructor training faster or for trying out a label in order to find out its meaning. When a label is being considered, during the evaluation of the current situation for the picking of a behavior, the feedback from the exemplars which include this label is multiplied with a positive constant so as to disproportionately increase the chances of the behavior coupled to this label. The considering of labels is illustrated in Figure 5.7.

If the agent picked as an instructor and undergoing the instructor training stage already has a label for the choice which is being reinforced, it uses this

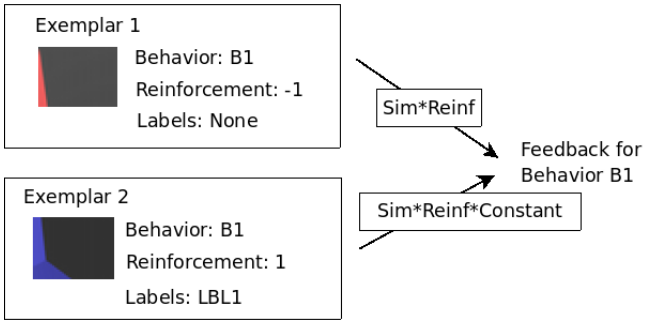

Figure 5.7: The use of labels when label LBL1 is being considered.

label for making the correct choice, thus finishing the instructor training step in a shorter time and without having to learn through the collecting of exemplars the correct choice. When the instructor makes a decision, it also determines whether there are any labels which apply in the given situation. Once these labels are determined and an action is carried out, the reinforcement received is used to find out which decision is being reinforced. A label corresponding to the reinforced choice is then picked as the label according to which the behavior is coordinated. Just as in the interpretation of a label by a student, this picked label is further used to make the correct choice, without having to go through the repetitive process of reinforcement learning. The pseudo-algorithm for the behavior of the instructor can be seen in Algorithm 3[16].

As this experimental step makes clear, the choice which has to be communicated among the agents in the language games is given to the agents not directly, but by having them engage in embodied interaction, and providing appropriate reinforcement. The instructor does not rely on an internal representation of the choice that was reinforced when it is asked for a label to be interpreted by the student in the next experimental step. Instead, it makes an evaluation to pick an action, and grounds the label it generates on this evaluation. Furthermore, in case the instructor has already engaged in a language game and has labels for actions, it can use these labels to coordinate its behavior. As it will be explained in the following paragraph, the instructor selects the exemplars which contribute to a decision. When a positive reinforcement is received, and the majority of the exemplars which contributed to that decision have a label attached to them, the instructor picks this label as the currently considered one. In the further trials in the third experimental step, the feedback from the exemplars which include this label is multiplied with a positive coefficient when picking a behavior, leading to the same decision being made consistently afterwards.

Fourth step: Communication In the fourth step, a student is randomly selected from the population of agents, excluding the current instructor. The

[16]The label exemplar mechanism which was added as a selection mechanism after the first set of experiments is also included in this algorithm.

Algorithm 4 Behavior of a student in a language game

```
Considered label ← Null
Meaning of label was guessed ← False
while language game is not over do
  Behavior choices & applicable signs ← Evaluation of the situation
  if there are two choices for actions then
    Ask instructor for label L specifying direction
    Considered label ← L
    if L is not among the applicable signs then
      Guess one of the choices (B) as the correct one
      Meaning of label was guessed ← True
      Attach L to the exemplars contributing to a decision for B
    end if
    Behavior choice ← Evaluation of the situation
  end if
  Carry out last selected behavior and pick reinforcement
  if (reinforcement < 0) and (Considered label ≠ Null) and (Label was
  guessed) then
    Remove label from the exemplars to which it was attached
    Meaning of label was guessed ← False
  end if
  if (reinforcement ≠ 0) and (Considered label ≠ Null) then
    Create exemplar with label and reinforcement
    Considered label ← Null
  end if
end while
```

student is placed at the intersection of the red and blue paths, in which position it has to make a choice, because the possibilities for going down one of these paths are comparable to each other. The student is also able to accumulate the exemplars which contribute to decisions, and find the applicable labels in a situation, as explained on p. 122 in the discussion of the instructor.

In order to decide on which path to take, the student "asks" the instructor, which has already been reinforced for one of the directions in the third step, for the direction to take. The response comes in the form of a three-letter word, which the instructor produces as follows. The instructor receives the image data which the student received, and evaluates it according to its own database, picking a behavior for execution.[17] The probability that the behavior which would lead the instructor in the direction which was reinforced in the third step is picked is high, but this is not guaranteed, which leads to the possibility of communication error. When the instructor makes a decision this way, it also selects the exemplars which have substantially contributed to this decision, just like the student. A label is then created or selected according to the following rules:

- If none of the selected exemplars has a label, create a word of three random letters.
- If the selected exemplars have only one label, return this label.
- If there are more than one labels attached to the selected exemplars, return the one which has been used or learned first.

These rules have turned out to be too simple, and a new way of picking labels has been devised, with a memory for label utterances. This mechanism will be explained later.

The student, after receiving the sign, checks whether the sign is among the applicable signs for this situation, applicability being determined through the exemplar accumulation and label detection mechanism explained above. If the sign was not among those in the exemplars similar to this situation, a random choice is made among the choices. The received sign is then attached to those exemplars which have been computed to contribute to the random choice made. Once this process of sign exchange is completed, the student reevaluates the current image. In this reevaluation, the received sign is considered, i.e. the feedback from the exemplars which include the label currently being tried out is multiplied with a positive coefficient, as explained above and shown in Figure 5.7. The probability for the random choice behavior is consequently much higher than the other ones. A behavior is picked, most probably the guessed choice, and carried out, after which the student arrives at one of the

[17] The availability of the current state to the instructor as it is given to the student is one more significant weakness of the model. The problem of how an agent can appropriate the point of view of another is a hotly disputed and significant topic in cognitive science, most recently discussed in the context of mirror neurons and theory of mind (see e.g. Gallagher, 2001). Simply passing on the same image as the current status is a simple solution to this problem.

end points. If the arrived endpoint is the correct one (the one picked at the beginning of the third step, and for which the instructor was reinforced), the agent gets a positive reinforcement; else, a negative reinforcement. In the case of positive reinforcement, the label which was attached to the exemplars is kept. Otherwise, the label is deleted, and the memory of the student is restored to its earlier state without the attached label. An algorithmic explanation of the behavior of the student can be seen in Algorithm 4[18].

At the end of a communication game, the instructor deletes the exemplars which were acquired during the instructor training. If these exemplars are not deleted, it takes the instructor a considerable amount of time until it learns to make a different choice in another episode in which it is an instructor, and the database becomes bigger, which lengthens the processing time for a decision. In a complete language game experiment, the third and fourth steps are repeated with different agents. In this process, the agents do not acquire any new exemplars, but the labels which are attached to these exemplars is changed.

5.4 Experiments and Results

In this section, various experiments carried out with agent populations of two different sizes (5 and 10 agents) consisting of agents trained to different degrees are explained and their results presented. In order to account for the various factors obstructing the emergence of a vocabulary which appeared in the progress of the experiments, different measures were implemented. The results of the experiments are presented in the form of graphs of the number of successful communication rounds in the last set of trials, with the size of this set depending on the size of the population, plotted against the number of trials in total. This enables a continuous visualization of the change of success in communication, at the same time giving sense of how long it takes to reach this degree of success. The experiments are carried on until the population reaches a pre-determined number of consecutive successful communication rounds; 10 rounds for 5 agents, and 20 rounds for 10 agent.

Using the above explained setup and computational mechanism, five agents were trained until the third phase. That is, these five agents, when placed in the intersection of the blue and red paths, had virtually two choices to make. Initially, the instructor success threshold for consecutive choices were set to 3 rounds. In this condition, the chances for error in communication stem from the instructor being trained for one choice, but failing to make that choice when asked for a label. At the end of the instructor training step, the probability for the instructor to make the reinforced choice when asked for a label is on average 0.6. Therefore, there is a chance of on average 0.4 that the instructor makes the wrong choice, giving to the agent a label corresponding not to the reinforced choice for that session, but to the incorrect one. If the student furthermore makes the correct choice, leading to positive feedback, there are now

[18]In Algorithm 4, the label-exemplar mechanism which will be explained below is also mentioned; in the first batch of experiments, this mechanism is not available.

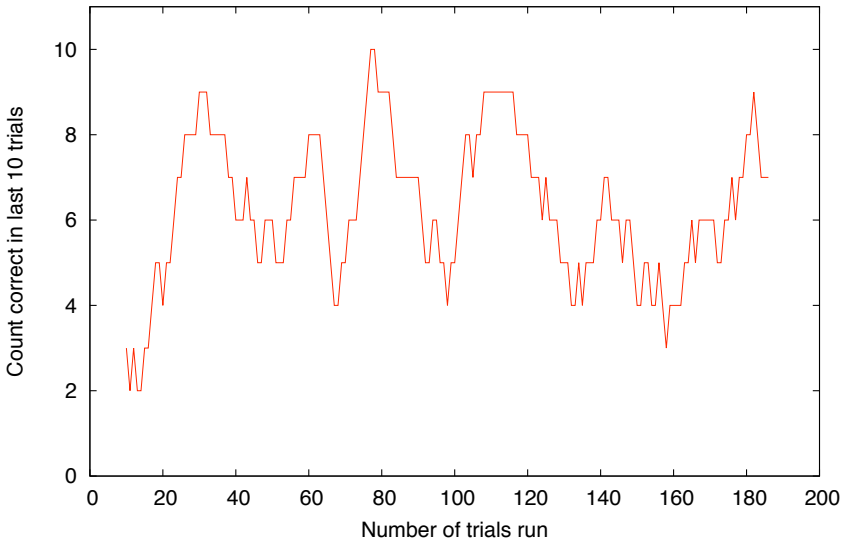

Figure 5.8: Lexicon formation with 5 agents and an instructor successful threshold of 3. Total number of correct responses in the last 10 trials is plotted against the number of trials run. Despite the considerable number of trials run, the success rate does not stabilize.

two agents with the same label referring to different paths. This becomes a source for continuous errors in the population, because there are no error correction mechanisms; the agents always pick the label which they either created themselves or used correctly when given by an instructor, as explained above. Furthermore, one such error in the early phases of the experiment leads to the arising of two groups of agents, each using the label with a different meaning. The reason for this is that the two agents which erroneously arrived at different meanings for the labels keep on using these labels in further language game episodes. The results of an experiment with an instructor success threshold of 3 are shown in Figure 5.8. In this figure, the number of correct communication attempts in the last 10 rounds is plotted against the number of trials which have taken place. As it can be seen in this figure, complete success is never reached, and the agents fail to converge on a vocabulary[19].

Increasing the instructor success threshold to 5 makes a big difference. With

[19]It should also be mentioned here that due to the nature of the experimental setup, involving a three-dimensional simulation, and the characteristics of the learning algorithm, with a comparisons of huge number of images at each decision step, the experiments take a considerable time to run. The above experiment which led to close to 200 trials to be performed took in excess of 6 hours to run.

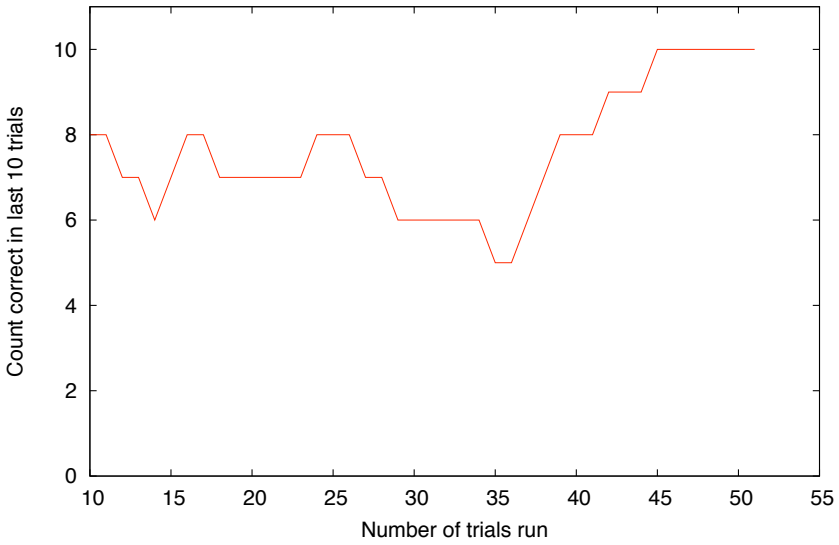

Figure 5.9: Lexicon formation with 5 agents with an instructor successful threshold of 5. Total number of correct responses in the last 10 trials is plotted against the number of trials run.

this higher threshold, the instructors have an average probability of 0.85 of making the correct choice when the student demands a label. This higher level of certainty leads to much lower chances for making an error. The results for an experiment in which five agents engaged in language games with an instructor success threshold of 5 rounds is shown in Figure 5.9. As it can be seen in this figure, the agents have managed to reach and maintain 100% correct for 7 trials in the last 10 trials at the end of a total of 57 trials. It must be mentioned that there is still a chance of an error being made, because the probability that the instructor makes the decision reinforced in that round is not unity. Such a perfect probability is the case only if there is already a label for the choice, as explained above. If an error does not happen in the early rounds of the experiment in which the labels for choices are established, the population of agents is able to agree on a vocabulary as shown in Figure 5.9.

In order to observe the effect of the size of the population on the formation of a vocabulary, the five agents mentioned above were each duplicated once, arriving at a population of ten agents. The exemplar basis, which provides the data on the environmental interaction on which the communication is based, is same for two sets of five agents in this community, but this does not have any effect on the dynamics of the label exchange, because the databases are copied (they simply point to the same images), and the labels are attached to

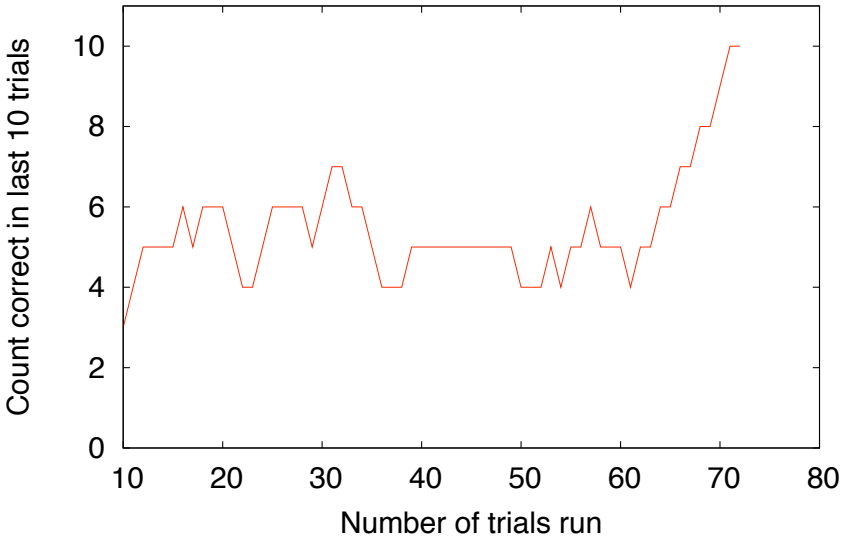

Figure 5.10: Lexicon formation with 10 agents, instructor successful threshold 5. Total number of correct responses in the last 10 trials is plotted against the number of trials run.

different exemplars. The results of the experiment with ten agents is displayed in Figure 5.10.

As it can be expected, the bigger population takes longer to arrive at a common lexicon than the smaller one. However, both populations are eventually able to communicate among each other on which path to take. Successful communication is a result of each of the agents learning all the labels that are used by the others. This fact can be seen in Figure 5.11, which depicts the number of different signs used in the last 10 rounds in the experiment of which the results are shown in Figure 5.10. There is no convergence in the labels used, and the number of labels used stays constant, meaning that the agents which are picked in later stages in the language game learn the labels created earlier in the game. The reason for this is that there is no mechanism which leads the agents to prefer labels which are used more often and more successfully. Such a mechanism would further lead to the avoidance of errors, because through the elimination of labels which are used differentially, the emergence of groups of agents with different label-choice couplings would be prevented. In the next section, the implementation and effects of such a learning mechanism will be discussed.

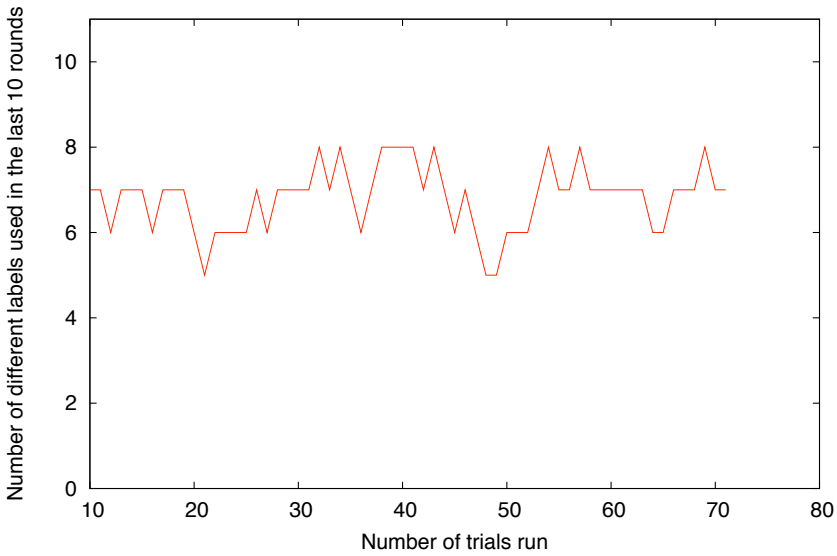

Figure 5.11: Number of different labels used in the last 10 rounds in a language game experiment with 10 agents, plotted against the number of trials run.

5.4.1 Probabilistic label selection

The simple label selection algorithm in which the agents use the labels which they learned or used the first time does not reliably lead to the formation of a vocabulary. This mechanism is also not conducive to error correction, in that the communication errors stemming from the use of the same label for different choices lead to repeated mistakes in communication, and the formation of groups which use the same label to refer to the same thing. In the overview of work on language games in Section 4.2.1.1, it was argued that a necessary process for the creation of a vocabulary is self-organization. The agents should use the more successful labels more often, in order to ensure coherence in the population. In the example of Steels and Kaplan (1999a), self-organization is achieved through the use of weights between words and meanings, increasing the value of such a weight between a word and a meaning when a word was used successfully in order to refer to a meaning. A similar effect can be achieved in the model presented here by using the action selection mechanism explained above to choose labels according to their successful use in communication. What is necessary for such a mechanism is a memory of utterances made and interpreted. Such a memory would correspond to the weighting mechanism of Steels and Kaplan (1999a), albeit without a single weight being connected to a meaning but with a set of exemplars.

The label memory G implemented to this end contains exemplars g which consist of a start state g_s, a label g_l, and the outcome of the language game g_r, i.e. $g = (g_s, g_l, g_r)$. Such an exemplar is stored by the speaker and the listener in each language game. The start state is the perceptual state in which the agents are when the language game takes place. Once the listener interprets the received label and receives a feedback, both agents store an exemplar g with the final outcome($g_r = +1$ for successful communication, $g_r = -1$ otherwise), the start state and the label. When an agent has to give a label to another agent, a procedure very similar to the selection of a behavior is applied. In case there is just one label for the behavior drawn by the speaker, this label is given to the listener. Otherwise, feedbacks are calculated for all labels in the database, and a label is drawn according to the softmax decision rule as in Equation 5.19, based on the feedbacks from the exemplars. The feedbacks are calculated as with the behaviors, according to Equation 5.18, with two differences. The first is that the behaviors in equation 5.18 are not matched; that is, which direction was signalled with the label is not included in the selection of the exemplars. It is rather the labels which are matched:

$$F(l, s_c) = \sum_{\{g \in G | g_l = l\}} sim(s_c, g_s) \cdot g_r \tag{5.22}$$

The second difference is the value of the τ temperature parameter in equation 5.19, through which probability values for the labels are obtained. The temperature parameter for behavior selection is relatively high, in order to lower the learning rate to enable discovery. In the language games, on the other hand, it is desirable that the agents learn at a faster rate, because the lexicon would

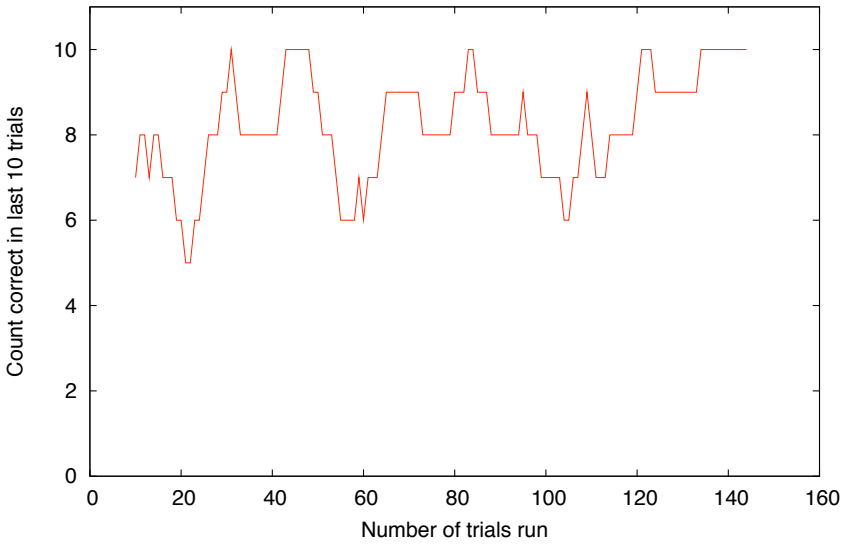

Figure 5.12: Lexicon formation with 5 agents and probabilistic label selection. Number of successful communication attempts in the last 10 trials is plotted against the number of trials run. The population reaches 100% correct in the last 10 trials and maintains it after 144 trials.

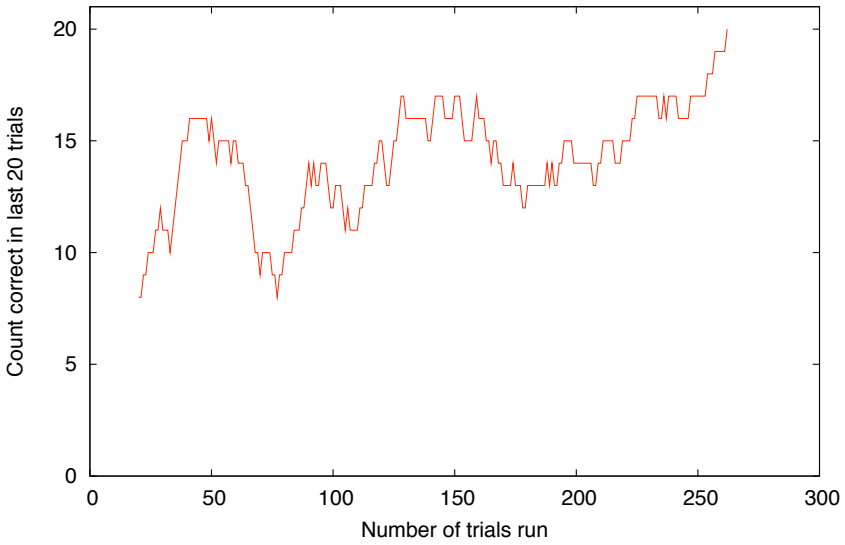

Figure 5.13: Lexicon formation with 10 agents and probabilistic label selection. Number of successful communication attempts in the last 20 trials is plotted against the number of trials run. The population reaches 100% correct in the last 20 trials after 262 trials.

otherwise stabilize in a much higher number of communication rounds. Therefore, the temperature value for the label selection process is one tenth of that for behavior selection, with $\tau = 2$.

The results of the experiments with probabilistic label selection are shown in Figure 5.12, where the population consisted of 5 agents, and Figure 5.13, where the population consists of 10 agents. As it can be seen from these figures, the agents manage to arrive at a lexicon also with the probabilistic label selection mechanism. The more interesting data, however, is the numbers of labels used and the co-occurrence frequency of the labels with the intended directions. The change in the number of labels used can be seen in Figure 5.14 for the experiment with 5 agents and Figure 5.15 for the experiment with 10 agents. In both experiments, the agents initially create a large number of labels. At the end of the first 10 trials, the population with 5 agents has 7 labels in use, and the population with 10 agents uses as much as 11 labels in a set of 20 trials. Each individual that acts as a speaker creates a new label if it does not already have one, leading to a proliferation of different labels. However, once all the agents have been involved in a communication round with each of the choices (blue and red), certain labels –especially ones which were used correctly the first time they come into play– get picked, and the more successful ones are selected more

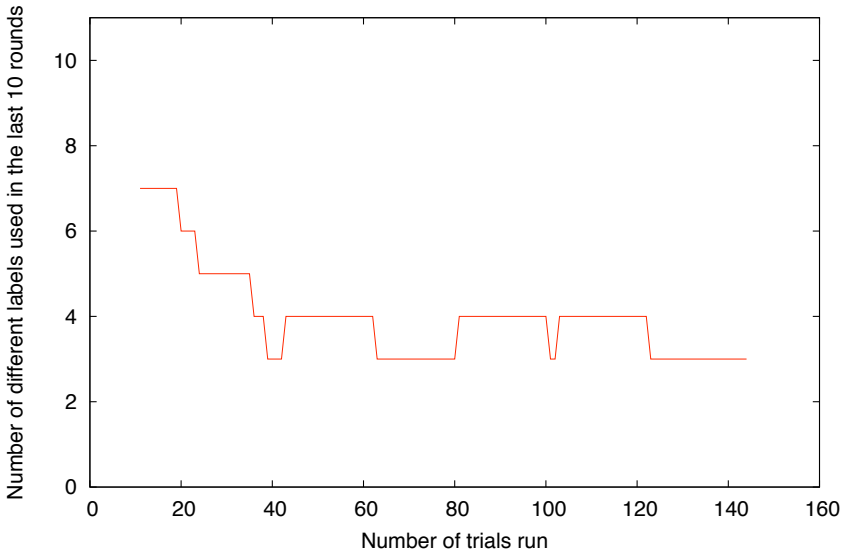

Figure 5.14: Number of different signs used in the last 20 rounds in an experiment with 5 agents, plotted against the number of trials run. As the population agrees on a limited number of labels for communication, the success rate in communication increases.

often. At the end, both populations converge on 3 labels, 2 for Blue and a single label for Red.

The convergence of the community on a few signs for the different choices can be seen more clearly in Figures 5.17 and 5.16. These depict the times each label was used in the last 10 or 20 trials on which Red was the given direction in experiments with 10 agents and 5 agents, respectively. As it can be seen from the figures, the communities arrive at single signs for the direction red in both experiments. In the community with 10 agents, a different label is used only once in the last 20 trials. The agreement of the community on the labels corresponds to the achievement of 100% correct communication, which shows that communicative success and the creation of a common vocabulary are parallel processes.

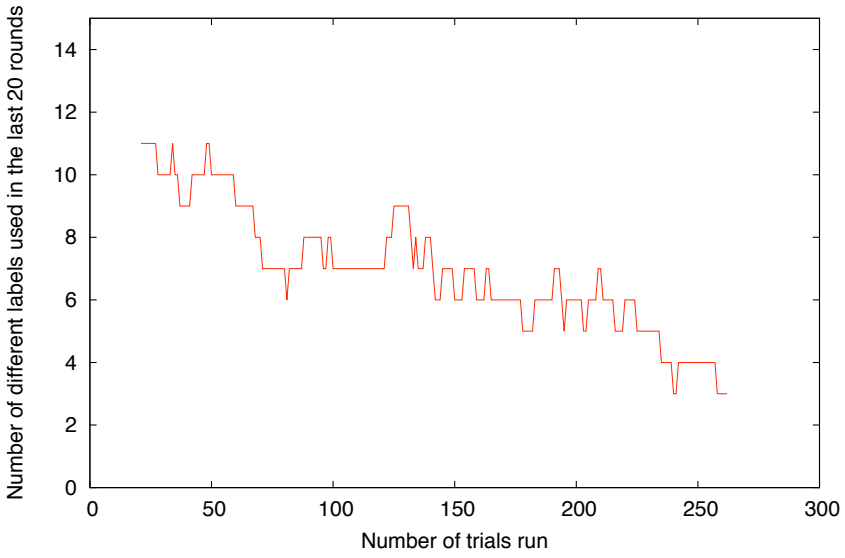

Figure 5.15: Number of different signs used in the last 20 rounds in an experiment with 10 agents, plotted against the number of trials run. The population starts with 11 labels, but through the progress of the experiment, arrives at 3 labels in total, 2 of these for Blue, and a single label for Red.

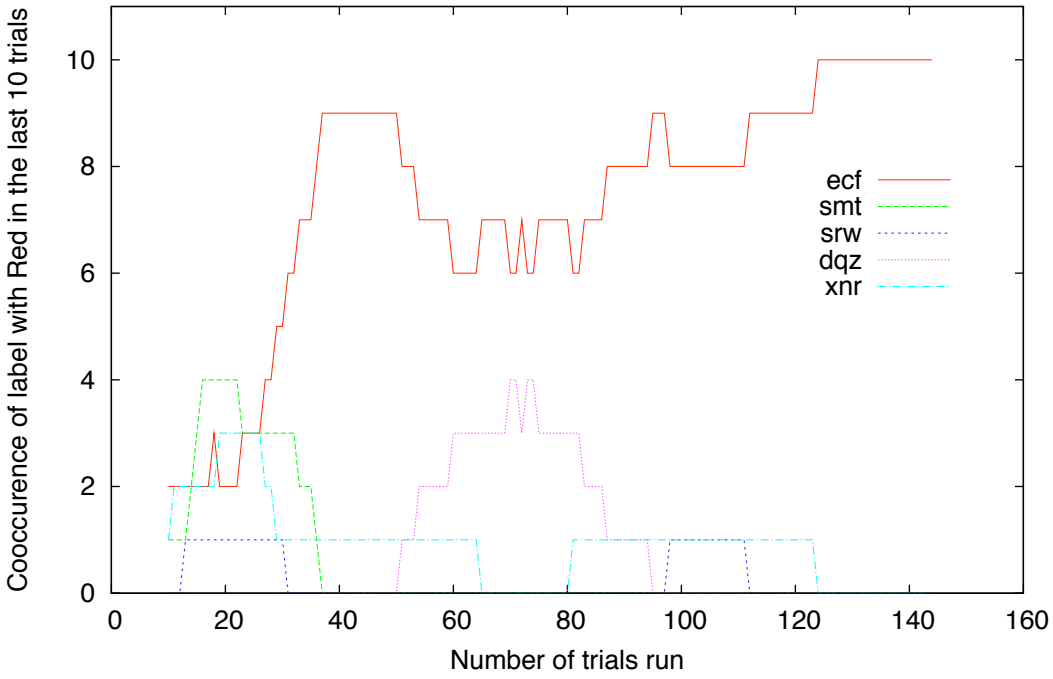

Figure 5.16: Co-occurrency of different labels with Red as the direction to be signalled in an experiment with 5 agents, plotted against the number of trials run. The agents agree on a single label for referring to Red, at the same time increasing their communicative performance.

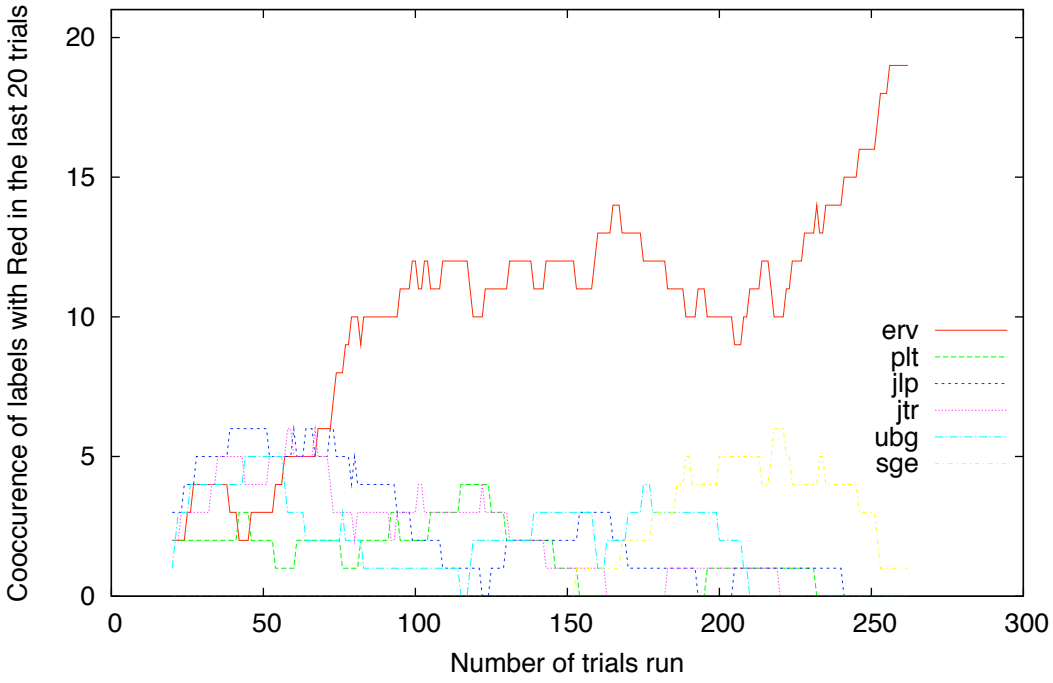

Figure 5.17: Co-occurrency of the labels with Red as direction in an experiment with 10 agents. A comparison with Figure 5.13 shows that the improvement of performance in communication games happens at the same time with the agreement on a single label. Labels which were used only a few times have not been included in this graphic in order to avoid clutter.

Chapter 6

Discussion

The preceding chapters have argued for understanding the symbolic representations used by humans in terms of situated representations. The distinguishing features of these kinds of representations were claimed to be their co-dependence with perceptual processing, and their public nature. The public nature of situated representations entails, as argued in Chapter 4, using multi-agent models for synthetic studies of situated representations. A dual role for symbolic representations was proposed, whereby they are created in communicative situations, but later serve agents in their individual behavior. These processes of creation and use were further studied in a computational simulation of a group of agents operating in a simple three-dimensional maze. This maze offered choices for the agents to take, and these choices served the agents to ground symbols. Situated representations were theorized to be a result of the intersection of the processes of communicative symbol use, self-organization of vocabularies and the utilization of such symbols in embodied behavior. One of the aims of the model was to understand situated representations in terms of the transactional perspective, according to which perceptual and motor data, concepts and names for things are coordinated in parallel.

The results of the computational model, presented in Chapter 5, show that it is possible to ground symbols in the embodied behavior of the agents. Communities of agents have been demonstrated to be able to arrive at a common vocabulary, referring to the choices in the environment with artificially created words and further using these words in their own embodied behavior in a situated manner. In this chapter, the significance of the results and the shortcomings of this particular approach and the model presented will be discussed. Some further extensions and improvements which can be made to the model will also be considered.

6.1 Language games and embodied grounding of symbols

In the synthetic studies of the dynamics of communication, the underlying idea of what purpose symbolic communication serves, and which fundamental processes enable agents to correctly use symbols in communication (what was taken to be a useful definition of meaning in Section 4.2) are evident in the organization of the experiments. For example, the evolutionary condition that successful communication should increase the survival chances of the individuals of a community are present in nearly all studies of the evolutionary emergence of communication. A tendency which was criticized in Chapter 2, the tendency to equate successful reference to objects or situations in the environment with task-independent internal representations, is one similar preconception that has a direct effect on the design of the experimental setup for language game experiments. The central contribution of the work presented here is the experimental setup used, and the conception of meaning which accompanies it, also affecting the design of the agents. Instead of understanding the process of meaning creation as located in a separate sphere to which individuals have private access, this process is viewed as a common way of organizing embodied behavior, which the individuals can then use for their private purposes. The final result is a community of agents which have differing individual experiences with the environment, represented in the form of a memory of exemplars. These agents command the same representational vocabulary for manipulating these differing memory stores and organizing their behavior conforming to the behavioral standards set by the reinforcement paradigm.

The reasons such a conception of meaning can be studied using the setup presented are twofold. First of all, reinforcement is given to the agents participating in a language game for the successful completion of an embodied task, in this case the navigation of a simple maze. The identification of choices which can be the subject of the episode of communication are not simply perceptual, but require action on the part of the agents. Successful communication about an object is not directly rewarded; what is rewarded instead is the successful use of communicated symbols in individual activity. This is coupled to the idea that symbols do not exist solely to correspond to things, but to enable the agents to organize their behavior in ways that would have been impossible or much more difficult without them, as argued in the context of situated representations in Section 3.3. The second reason is that the interaction between the agents takes the form of instruction giving. The special aspect to this kind of social interaction is that, as the work of Chapman (1991) discussed in Section 3.2 also demonstrated, it can build on existing common experience of the agents. The agents perceive certain possibilities in the environment, and their dialog is shaped according to these possibilities, taking them as given and as a part of the situation. The experience on which the communicative acts build is not of representational character, but instead it consists of low-level sensory data coupled with behavioral dynamics. This fact leads to the creation of symbols

as resources to use, instead of as parts of internal world models, to which the agents have privileged access.

The language game in which the agents engage is just one kind of such game, a game which can be called the instruction giving game. The interpretation of the symbol given by an agent thus corresponds to applying the given instruction, and trying out a certain kind of interaction in order to determine what the agent giving the instruction intended. One can envision other kinds of language games in order to model other kinds of interpretation, games in which the interpretation of the received symbol would involve a different kind of activity. An example for such an interaction is two agents cooperating at an activity, and using communication to coordinate with each other. This and other similar language games are a natural way to extend the setup presented. The used setup also makes clear that the language game in which the symbols were created and used is the basis on which a conception of the meaning of these symbols can be developed. In other language games, the symbols would acquire other roles, thereby extending the meaning of the symbols. An interesting question is how a symbol generated in the context of one language game could be used in another one; for example, the effects of using a symbol created for instructing another agent in information transfer, or the coordination of a group of agents.

One of the topics discussed was the role of training in correct communication. In the instructor training phase, the instructor goes through a number of runs in which the topic of the communication round is reinforced. If the number of these runs is low, the probability that the instructor picks the correct choice for attaching the label which it returns is not high. This leads to the probability that the agents attach a label to different choices, and further propagate these connections, leading to the creation of groups of agents that refer to different choices with the same label. When the number of these instructor training runs is increased, the probability of such errors decreases. One point of view that can be taken to put this phenomenon into context is understanding it as a reflection of embodied domain knowledge on communication. For a coherent communicative toolbox to arise, the agents should command a certain knowledge of their environment, both in the sense that they should be able to sense the choices that are available to them, and be able to reliably evaluate the correct choice of what they want to point out. If their capacities in both of these aspects is below a certain level, the population will not be able to arrive at a vocabulary. It was further shown that if the number of errors that are made is relatively low, a mechanism similar to the one used for embodied behavior can serve as an error-correction and feedback mechanism, leading to a simpler vocabulary and the avoidance of erroneous labels. This mechanism was further established using an utterance memory and a decision mechanism practically identical to the action decision mechanism, only with different constants which lead to faster learning.

6.1.1 Situatedness and the transactional view

Of course, it is perfectly possible to build an agent that functions in the discussed environment using the traditional methods, by creating a map of the area with perceptual filtering, creating internal representations for the choices which can be identified on this map, and then attaching labels to these representations to use in communication. The main target of the work presented was to model situated representations in a synthetic way, and the algorithmic approach to building the agents was a direct result of this concern, rather than reaching a high performance according to an independently measured standard. It was argued in Section 3.3 that what is referred to in a synthetic context as situated representations draws its inspiration from the human case, but it must initially stay as an approximation. One must therefore keep in mind what kind of situatedness the current setup and algorithmic design of the agents allow, and how far the sort of situatedness implemented here is from the natural human case.

The idea of situatedness which this model embodies is one of monitoring and reorganizing of behavior based on sensory information at each behavioral step. At each step in which a decision has to be made, the whole memory is reprocessed and the individual exemplars compared to the current situation. The relevance of past experiences is reevaluated with each new situation, whereby the individual episodes of experience are brought to bear on the current situation, and contribute to the current decision in accordance with their relevance. This processing leads to the determination of the actions which can be applied in that situation, and the symbols which can be used for behavioral coordination. In other words, symbols and actions are recontextualized with each step, and the execution of behaviors and utilization of symbols is fundamentally dependent on their being applicable in a given situation. Furthermore, once symbols and behaviors are recontextualized, they become part of the situation; agents can communicate on this basis, as argued above, making their interaction grounded in the low-level information from the sensors. The sense of situatedness here, therefore, is close to the role Clancey (1997) and Clark (2005) claim for situated representations in intelligence, in that symbols created for communication serve to augment the individual capacities of the agent. An important point here is that the extraction of such symbols from the memory for use in communication, for use in individual behavior, and the decision of which behavior to execute are carried out through the use of similar low-level processes operating on sensory data, without assuming ad hoc processes on top of the existing machinery for behavioral learning and embodied coordination.

Regarding situatedness, the principal limitation of the presented model is that the processing of sensory data is carried out in terms of the standard statement of the machine learning problem as discussed in Section 5.1.2. That is, for each pattern of sensory input, the memory is processed in an off-line manner. As a result, during this processing, the robot is not available for reacting to changes in the environment or to new contingencies. The output, a behavior, is then executed, and during this execution the robot is again off-line. These two

off-line phases point to the limits of the situatedness of this approach: Instead of being directed to the environment continuously, the agent is acting in executional steps, sampling the environment and processing data thus gathered to arrive at a decision. This fact is reflected also in the format in which the sensory data is gathered and stored, i.e. as static images which contain no information about the temporal contingencies of consecutive behaviors. The difference between an agent which is closely coupled to the environment and one that samples the environment is similar to two people, one following a moving pen with his eyes, and the other looking where the pen is, closing his eyes, and then opening them to find where the pen is once more. In the case of a continuously coupled agent, the internal dynamics are coupled to the external world at each moment. It is possible to extend the current methodology of exemplar-based learning to achieve this more genuine kind of situatedness, what can be called dynamic situatedness. A possible way of achieving dynamic situatedness will be discussed below.

The concept of situatedness was further discussed in terms of the transactional perspective in Section 3.2. The transactional perspective developed by Clancey (1997) argues that categorization, symbols and sensory-motor contingencies develop at the same time in behavior, constraining and determining each other in parallel. In the learning algorithm developed, the symbol to be used for behavioral coordination is determined based on the contents of the memory. Afterwards, how the exemplars in the memory affect the current decision process is determined by the symbol used to change behavior, which can be taken as a case of the co-constraining of lower-level sensory information and situated representations. However, as the above discussion of situated representations also demonstrated, there are a number of shortcomings in the realization of the transactional perspective. The first of these is the static and step-by-step nature of the processing mechanism; the sensory data and symbols are not shaped in parallel. The second important shortcoming is that the idea of the fundamental text, the data which cannot be contested, is still prevalent in the form of sensory data stored verbatim in memory. This data is not modified or shaped to fit a sensory category, as a result of constraints from other levels. This phenomenon is observed routinely in humans and is called categorical perception.

6.2 Shortcomings

The most significant shortcoming of the approach presented is the fact that the behaviors are hand-coded. Especially the turning behaviors are designed so that the agent faces one of the target walls when it executes these behaviors. There are only three kinds of behaviors: turning right or left and going forward. This is the minimal number necessary for carrying out the described language game experiments in this environment. This would enable one to claim that there is still internal representation of these choices, with these representations simply encoded as behaviors. The solution of this problem would be extending the machine learning mechanism to a regression learning system from simple

categorization. The machine learning task for the model was stated in terms of a categorization task, as discussed in Section 5.1. In categorization, input data is matched onto discrete pre-defined categories. In regression, however, input data has to be matched onto values from a continuous interval. Extending the current algorithm to enable a regression task would solve the problem of hand-coded behaviors.

At first blush, this appears to be a relatively simple extension. In its current form, the learning algorithm delivers feedback values for the various behaviors. Instead of calculating activation probabilities for individual behaviors based on these feedbacks values, one could multiply the speed values for the individual wheels with normalized feedbacks, thereby arriving at motor speeds which are in a continuous interval. The main problem with this simple solution was found out to be that the effect the accumulation of exemplars with positive reinforcement has in the categorization task – i.e. the increase in feedback and therefore in probability for the behavior responsible for the feedback – leads to an adverse effect in case the algorithm is minimally modified for regression. As the feedback for a certain range of values increases, the accumulation of exemplars leads to the motor speeds to overshoot, going over the optimal values. This excess feedback is initially accounted for by negative feedback from resulting exemplars which lead the robot to end in an undesirable position, but the learning process continuously oscillates around the optimal range, never settling in the desired range. Once this problem of oscillating around the optimal range of motor values is solved, the current learning algorithm can easily be further extended to accommodate regression learning.

One further important problem, alluded to on p. 125, is one faced by many multi-agent systems, and involves the shared perception of a scene from different perspectives. The instructor and student agents in the current setup both receive the same sensory data when the student has choices and asks for help to make a decision. When two people observe a scene, they have differing perspectives, but they have an understanding of what the other is seeing. Thanks to this mutual understanding of each other's perspectives, humans can refer to things, which might even be occluded from the perspective of someone else. This fundamental capability is a very difficult subject, and simply handing over the same perspectival data to both agents is a simple but unrealistic assumption to avoid the additional complexity that would be caused by the integration of a solution to this question into the current model; future work should take this issue more seriously and offer possible solutions.

6.2.1 Meaning, mattering and language games

One central question which the work presented has not addressed is one that is relatively vague, and is not usually asked either in AI or in studies of the dynamics of language. A glimpse into the nature of this question can be had when one considers how the experimental steps, i.e. the steps in the interaction of the student and instructor in a language game, are determined. These steps are programmed in by the designer, and their communicative purpose and success is a

derivative of our idea of what constitutes communication, or a prototypical case of communicative interaction. As explained in Section 2.4.1, Wittgenstein discussed such prototypical communicative contexts under the concept of language games. One important feature of language games is that there are numerous kinds of them. Humans create such games as they see suitable, and they are not scripted or pre-given: their form derives from the contingencies of the overall situation, and most importantly, from the aims of the agents, i.e. what they want to achieve in the given situation, based on their knowledge of the situation and each other.

Therefore, the main problem can be stated as "what is the motivation for agents to communicate?" In the setup presented, as mentioned above, the interaction is programmed in, and the agents are just going through these steps. If different language games are to be generated – either programmatically or by the agents – on the basis of what the agents want to achieve in a given situation, and how this can be done using their resources, the question of why communication should take place is of central importance. This can be seen when examples of the creation of language games are considered, for example in the case where one agent knows that the other one has certain information which it does not have, so it makes an utterance with which it intends to express its wish to acquire this information. Such a setting would constitute the language game of asking a question and answering. Another case, already mentioned above, would be one where the agents want to coordinate their actions, and they can achieve this by exchanging symbols which should specify to each other what each one of them is to do; this would constitute the language game of commanding.

In the field of language evolution, especially work in terms of artificial life and genetic algorithms research tries to tackle the problem of how and why communication can emerge; the work by Quinn (2001), or by Williams et al. (2008), discussed in Section 4.2.2 are good examples. The most important problem with these approaches is that communication is still taken as a given, and as an activity which qualifies the agents for being selected for reproduction; the fitness function, specifying the successful agents, is nearly a definition of what the agents have to do in the first place, giving them the standards for communicative behavior. The question of what forms the motivation to communicate takes, and how it leads to different language games, is an important one that has to be addressed by future studies.

6.3 Future directions

There are many directions in which the variation of language games experiments and communicative agents presented here can be extended. One possible extension to the current approach which has already been mentioned is the extension of the learning mechanism to regression learning. The simple version of using the feedback values as activations of various behavior nodes is unfortunately not practical, as explained above. If this problem can be solved, the existing model

can be extended to allow the determination of the optimal motor speed values autonomously, without the designer having to determine these values and code them in. Another simple addition is the inclusion of more complicated tasks by extending the capabilities of the robot, such as a gripping arm. A gripping arm is available for the real Khepera robot, allowing the picking up and carrying of small objects. Integrating a model of this physical extension in the simulation of the robot would allow the manipulation of such small objects in the simulation. When such an object is between the arms of the gripper, its top surface is in the range of the camera, making this surface a part of the visual scene; therefore, the simulated camera would still be a sufficient sensory mechanism for extending the setup with a gripper arm. The learning algorithm would have to be modified to accommodate the behaviors necessary for picking an object, but otherwise, the modifications which have to be made to the algorithm explained in Chapter 5 are minimal. The inclusion of the simple manipulation of objects (picking them up and carrying them to certain places) would enable the creation of more complex reinforcement schemes, which would lead to a better understanding of the phenomena of situated representations.

A more difficult but interesting extension concerns the representation of a situation. The representation of the situations in the exemplar memory is limited at the moment to image data from the simulated camera. The advantage of a single source for sensory data is that a uniform similarity measure can be easily defined; the Euclidean distance measure currently used can be applied to the different situations without concern for the weights of different parts, because the data as a whole belong to one channel. This becomes a limitation in case other sources of sensory data were to be added to the specification of a situation from the perspective of an agent. For example, if proximity sensor data from eight sensors on the original Khepera robot were to be integrated simply by appending the sensor readings at the end of the image data, the information from these sensors would drown in the mass of the image data due to the differences between the amounts of data from these different sources – $3 \times 32 \times 42$ bytes versus 8 bytes. What would be necessary is a weighting of the different sensory channels according to their significance and the amount of data they provide. Such a weighting mechanism is a problem that has to be solved or circumvented by any attempt to include further sensory data into the representation of a state.

As a more complicated and speculative extension, the idea of a situation can be broadened not only by including sensory data from different channels, but also by qualitatively extending a situation both temporally, and structurally. Temporally, the situations, as they are defined and processed at the moment, are completely static and contain no information about the sequence in which sensory data is acquired. The way the perceptual situation is transformed as a result of the behavioral dynamics is a significant resource from which an agent can learn, and these transformations reveal more about the way the environment is structured than static snapshots. The image data is currently saved without any reference to what has come before it and what comes after, except for the second image in an exemplar depicting the perceptual state after the

execution of the behavior. The dynamics of the behaviors, as mentioned above, are completely ignored. If what was referred to as dynamic situatedness above is to be achieved, the transformational properties of the environment have to be integrated into the learning mechanism. A rough outline of a possible method to achieve this can be drawn as follows. Instead of defining exemplars as single images and individual behaviors, a record of the sensory transformation (i.e. how the sensory state has changed over the course of the execution of the behavior) and the unfolding of the behavioral coordination can be stored together as an exemplar. What is then needed are computational mechanisms to compare sensory transformations with each other and, on the basis of these comparisons, create a new specification for motor transformations. These mechanisms are akin in spirit to those proposed by Tomasello (2000) as the necessary mechanisms for language acquisition, i.e. computational resources for analogical reasoning and structure mapping.

A final possibility for building on the presented model concerns the way the images are saved and compared to each other. The perceptual data is saved in raw pixel form, and the comparison of the images is done pixel by pixel. The overlapping of the pixels, i.e. the fact that what is at the top right of an image will be compared with the top right part of all the other images, is the only sort of structure which is effective in the processing of perceptual data. In case there are relatively small details in an otherwise uniform image, these should be paid more attention to. Learning here is achieved by an accumulation of exemplars, but in case the feature is small, the learning of this feature (e.g. a small black mark at the top right of the image) takes too long, and is sometimes even impossible. The solution can be inspired by natural perceptual systems. Processing in the human visual system starts in the retina, in that simple features such as edges are formed through processes such as lateral inhibition. The important thing is that these processes do not deliver a description of a scene, but modify the perceptual data to make certain things more salient. A similar effect can be achieved by modifying the images through simple filters which change the pixel values to the degree that there is an edge in that area, creating a color gradient around the edges, or any other features which can be brought to prominence using such a procedure. This way, the similarity of images with similar features in close areas can be increased, at the same time decreasing the similarity of images with this feature and those that do not, without having to create structured images.

Chapter 7

Conclusion

The aim of the work presented here has been to contribute to the debate on the use of representational entities by living beings, and their explanatory role in synthetic models of intelligence. This has been done on the basis of a discussion of how this subject has traditionally been treated in cognitive science and the philosophical background which has provided the historical seeds for this treatment. In the spirit of cognitive science, a computational model of the grounding of a common vocabulary by a community of agents was presented. "Grounding" here refers to understanding the use of symbols by agents on the basis of their physical capacities and the situation as it is perceived by them and organized according to their bodily capacities. Taking the recent advances in situated and embodied cognitive science as a background, it has been argued that symbolic representations as they are used by humans are fundamentally situation-dependent and arise out of social interactions. The situated view of representations was further brought to work in the multi-agent model of their production and use by embodied agents. In this concluding chapter, a summary of the discussion will be given, and the main conclusions will be stated.

One of the central discussion topics in cognitive science and AI in recent years has been the status of symbolic representations as the central means of intelligence, and as explanatory constructs to explain the mind. A closely related topic is to what extent language and communicative use of symbols contribute to the capacities of intelligent beings. The default position on these questions has been one that took over the individualistic and mentalist tradition of rationalism, postulating an inner sphere in which the external world is modelled. The application of this methodology to AI has been the assumption that intelligence is a result of the manipulation of symbols. Symbolic representations were therefore used extensively in traditional approaches to AI, what was called GOFAI by Haugeland (1985). According to this paradigm, also called cognitivism, the human mind, and possibly any intelligent agent, is a symbol-processing system. Symbols in such a system stand in representational relations to entities in the world. Through the manipulation of these symbols, agents can reason about the world, and act in a goal-directed and structured manner. The posi-

tion on communication concurrent with the cognitivist position on the role of symbols in intelligence is one that attributes to language the role of a vessel, in that language serves to carry contents of a mental language from one head into the other. This position has been further bolstered by the dominant linguistic paradigm, the Chomskian school, which claimed an innate, encapsulated language organ, independent of the workings of the rest of the mind, and solely responsible for the transformation of thoughts into sentences.

The cognitivist position, especially as it was practiced in AI, has run into a number of technical problems. One example for these problems is the frame problem, which appears when a symbol-based system has to reason about a changing world. Another example is the common sense problem, in which the whole of common sense knowledge has to be programmed in to build a system capable of reasoning about the most simple human tasks. A more general problem, which encompasses these two to a certain extent, is the relevance problem. An agent functioning rationally, i.e. computing an action based on given facts, faces combinatorial explosion when the number of facts which it has to take into account grows. Partitioning such a system into modules does not solve the problem, because information relevant to a situation can come from any one of the various modules. It was argued in Chapter 2 that these problems were a result of the individualist and mentalist tendencies of cognitivism; cognition is seen as going inside the head of individual agents, facing an unknown world which has to be made sense of, and the community a group of agents which have to be deciphered.

The enduring technical obstacles in AI and criticisms stemming from similar concerns in cognitive psychology, coupled with philosophical criticisms of the underlying rationalist framework, led to the development of a new paradigm which rejected the representation-centric view of traditional cognitive science, and instead claimed the primacy of agent-environment coupling and grounding of representational tools in this coupling. This paradigm, called situated and embodied cognitive science, aimed to understand the high-level capacities which have been the primary subject matter of traditional AI and cognitive science as grounded in and fundamentally dependent upon the so-called low-level capacities, such as embodied behavior, and social capacities. The most significant claims of the situated-embodied paradigm have centered around the role of internal models and symbolic representations in AI. Against the detachment of artificial agents from their world through the use of internal models, the tight coupling of agents with the environment with sensory-motor loops has been championed, under the motto "the world is its best own model". Instead of seeing mental representations as private resources to which an agent has privileged access, and through which it can reason about the world without engaging with the world, representations were studied as resources which have to be interpreted, in a similar way to maps or plans for activity. This point also opens the way to an alternative understanding of the role of language in human behavior, namely as a resource which can also be used in individual activity. Relying on this alternative understanding of symbol use, various theories of the use of linguistic means by the individual have been proposed. To summarize,

the situated and embodied approaches have concentrated on how behavioral capabilities are adapted to their environment, how they can appear structured and planned by being coupled to the regularities in the environment, and how symbolic capabilities derive their benefit for the agent from their situation- and context-dependency, instead of their being abstract context-independent entities.

The insights from embodied-situated cognitive science into the role of communities in the creation of meaning, coupled with the multi-agent modelling methodologies developed especially in artificial life, have led to the development of the field of evolution of language. This field studies the emergence of a vocabulary in a community of agents which interact locally, communicating about meanings – either as concepts in the head or about objects in the environment – and updating their behavioral strategies as a result of the success of these communicative episodes. These studies are methodologically either based on genetic methods, or use agents with learning algorithms. They can further either be situated, in that the agents function in an environment and the symbols they use in communication refer to the objects in this environment, or alternatively they communicate about pre-given meanings. Various experiments with different kinds of agents in different environments have shown that a community consisting of agents with certain mechanisms can achieve a common vocabulary, i.e. a common means of referring to shared meanings. Among the useful mechanisms for converging on symbols are self-organization, common means of categorization, reinforcement learning on the results of communication, and selection of the successful symbols. An interesting possibility offered by the multi-agent modelling of the emergence of symbolic communication is that symbols created in communicative contexts can further be used by the individual agents, offering a way of modelling symbol creation and use without falling into the trap of an internalized language.

If this possibility is to be realized, a number of lessons learned from situated-embodied AI have to be taken more seriously. The models of the dynamics of communication have a number of shortcomings, stemming from deep-seated convictions on the role of language, and how communicative symbols relate to meanings. These shortcomings exhibit themselves primarily in the organization of the experiments, where the aim of the agents is to communicate correctly, and they receive reinforcement for doing so. Because the agents have to simply transmit labels referring to the correct objects or meanings, the degree of their situatedness is also limited. In studies on the evolution of language, this bias takes the form of choosing the agents for reproduction which are the most successful in communicative tasks.

As a proposal on how the studies in the dynamics of language can be improved based on the discussions on the role of representations in intelligence, and the overview of work done in this field, an experimental setup, and a model which functioned in this setup, were presented. The main aim of the presented setup is to embed communication and the use of symbols in a concrete task context, and avoid delegating to the symbols solely the role of labels for internal representations. This is achieved by giving reinforcement not for successfully re-

ferring to an entity, but instead for the successful completion of a task. In order to achieve situatedness, this task involves an embodied coordination problem. The embodied coordination problem requires a physical environment, which was in the case of the setup presented a simulated Y-maze. This Y-maze presents the agents with two choices to make. These choices, going left or right, are color-coded as red or blue, in that the end locations and the paths have these colors. The agents are simulated Khepera robots with two wheels as actuators, and speeds for these wheels constitute different behaviors which enable the agents to navigate the maze.

An episode of communication between an instructor and a student consists of the instructor being trained to make a certain choice (left or right) in this maze, and afterwards giving the student a symbol corresponding to this choice. The student then has to make a guess as to which direction this symbol corresponds to. Based on the reinforcement which it receives when this guess is executed, the student either keeps the symbol or discards it as incorrect. The learning algorithm for the agents is similarity-based, and uses exemplars for the representation of episodes of action. The contents of the exemplars consists of the start and end states in the form of image data, action undertaken, and the reinforcement received for this action. The images are compared to each other using the Euclidean distance and similarity is calculated from this distance with the reverse exponential. At each behavioral step, feedbacks for the various behaviors are calculated based on the exemplars in the memory. These feedbacks are then converted to probabilities. On the basis of these probabilities, a behavior is drawn for execution. In order to avoid the credit assignment problem, similarity is used to chain exemplars in the calculation of feedbacks.

Experiments carried out with this setup and agent communities of different sizes have shown that in such a setup the self-organization of a vocabulary can be observed. The most important condition for the emergence of a vocabulary is that the probability of communication error, stemming from the instructor making a different choice from the topic, is not too high. Without a selection mechanism, the emergence of the vocabulary takes the form of each agent learning each and every label through repeated games. In order to better model the emergence of a vocabulary, an additional memory of utterance exemplars was created. This memory includes exemplars which represent an episode of communicative interaction, with the label used and the outcome of the interpretation. Through the use of this memory, more successful symbols are used more often, leading to the emergence of a simpler vocabulary.

The model presented here is limited in many respects. The experimental environment is too limited, in that only two choices are presented to the agents. In parallel with the limitations of the environment, the motor and perceptual apparatus of the agents is also relatively limited, with only two motors as actuators and a simulated camera for perceptual data. Also, due to the specific nature of the learning algorithm, the individual behaviors have to be hand-coded. In order to improve on these aspects, a number of possible extensions have been proposed in Section 6.3. Some examples for these extensions are including objects that can be manipulated and a robotic arm which can move these around,

or including data from other perceptual sources. The action selection mechanism can be extended to carry out regression on the image data, and not just simple categorization. Other more promising extensions concern the representation of the start and end situations. Instead of saving the images as they are received from the simulated camera on the robot, certain salient features on the image (such as edges, lines etc.) can be brought to bear on the similarity without creating a structured representation of the scenes. This can be done by modifying the image to create gradients around these features. A speculative but very interesting extension concerns the coupling of the behavioral dynamics and the perceptual transformation in order to arrive at an alternative way of representing exemplars. The crucial component of such an extension would be an algorithm for generating various behavioral transformations based on how the perceptual transformations have been affected by the different behavioral dynamics, as such information is available in the memory.

The role of language in intelligent behavior and the status of symbolic representations in models and natural beings are crucial topics for cognitive science. New insights on these subjects will deliver the key to many other problems, and also new ways of thinking which might help us better realize alternative visions. The work presented is intended to be a step in the direction of such a vision, and it is hoped that it may be of use for further progress.

Bibliography

Agre, P. E. (1985). Routines. AI Memo 828, MIT Artificial Intelligence Laboratory.

Agre, P. E. (1995). Computational research on interaction and agency. *Artificial Intelligence*, 72:1–52.

Agre, P. E. (1997). *Computation and Human Experience*. Cambridge University Press, Cambridge, MA.

Agre, P. E. and Chapman, D. (1987). Pengi: An implementation of a theory of activity. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, pages 196–201, Seattle.

Agre, P. E. and Chapman, D. (1990). What are plans for? In Maes, P., editor, *Designing Autonomous Agents: Theory and Practice from Biology to Engineering and Back*, pages 17–34. MIT Press, Cambridge, MA.

Aha, D. W. and Goldstone, R. L. (1992). Concept learning and flexible weighting. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, pages 534–539, Hillsdale, NJ. Erlbaum.

Allen, S. W. and Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General*, 120(1):3–19.

Anderson, M. and Perlis, D. (2002). Symbol systems. In Nadel, L., Chalmers, D., Culicover, P., French, B., and Goldstone, R., editors, *Encyclopedia of Cognitive Science*. Macmillan Publishers, London.

Anderson, M. L. (2003). Embodied cognition: A field guide. *Communications of the ACM*, 149(1):91–130.

Arkin, R. C. (1998). *Behavior-Based Robotics*. The MIT Press, Cambridge, MA.

Baker, G. P. and Hacker, P. M. S. (1984). *Wittgenstein: Meaning and Understanding*. Basil Blackwell, Oxford.

Barsalou, L. (1998). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4):577–660.

Barsalou, L. (2008). Grounded cognition. *Annual Review of Psychology*, 59:617–645.

Barsalou, L. W. (1999). Language comprehension: Archival memory or preparation for situated action? *Discourse Processes*, 28(1):61–81.

Bechtel, W. (1988). *Philosophy of Mind: An Overview for Cognitive Science*. Lawrence Erlbaum, Hillsdale, NJ.

Beer, R. (1995). A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence*, 72:173–215.

Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Science*, 4(3):91–99.

Beer, R. D. (2009). Dynamical systems and embedded cognition. In Frankish, K. and Ramsey, W., editors, *The Cambridge Handbook of Artificial Intelligence*. Cambridge Uniersity Press.

Berk, L. and Garvin, R. (1984). Development of private speech among low-income appalachian children. *Developmental Psychology*, 20(2):271–286.

Bickerton, D. (1990). *Language and Species*. University of Chicago Press, Chicago.

Bickerton, D. (2003). Symbol and structure: a comprehensive framework for language evolution. In Christiansen, M. and Kirby, S., editors, *Language Evolution: The States of the Art*. Oxford University Press.

Blume, T. and Demmerling, C. (1998). *Grundprobleme der analytischen Sprachphilosophie*. UTB, Stuttgart.

Boden, M. A., editor (1990). *The Philosophy of Artificial Intelligence*. Oxford University Press, New York.

Bontempi, G., Birattari, M., and Bersini, H. (1999). Lazy learning for modeling and control design. *International Journal of Control*, 72(7/8):643–658.

Bourgine, P. and Stewart, J. (2004). Autopoiesis and cognition. *Artificial Life*, 10:327–345.

Boysen, S. T., Bernston, G., Hannan, M., and Cacioppo, J. (1996). Quantity-based inference and symbolic representation in chimpanzees (pan troglodytes). *Journal of Experimental Psychology: Animal Behavior Processes*, 22:76–86.

Brooks, R. A. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, 2(1):14–23.

Brooks, R. A. (1990). Elephants don't play chess. *Robotics and Autonomous Systems*, 6(1&2):3–15.

Brooks, R. A. (1991). Intelligence without reason. In Myopoulos, J. and Reiter, R., editors, *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI-91)*, pages 569–595, Sydney, Australia. Morgan Kaufmann publishers Inc.: San Mateo, CA, USA.

Call, J., Hare, B., Carpenter, M., and Tomasello, M. (2004). 'unwilling' versus 'unable': Chimpanzees' understanding of human intentional action. *Developmental Science*, 7:488–498.

Cangelosi, A. (2001). Evolution of communication and language using signals, symbols and words. *IEEE Transactions on Evolutionary Computation*, 5:93–101.

Cangelosi, A. (2004). The sensorimotor bases of linguistic structure: Experiments with grounded adaptive agents. In Schaal, S., Ijspeert, A. J., Billard, A., Vijayakumar, S., Hallam, J., and Meyer, J.-A., editors, *Proceedings of the Eighth International Conference on the Simulation of Adaptive Behaviour: From Animals to Animats 8*, pages 487–496, Cambridge, MA. MIT Press.

Cangelosi, A. and Harnad, S. (2001). The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language in perceptual categories. *Evolution of Communication*, 4(1):117–142.

Cangelosi, A. and Parisi, D. (1998). The emergence of a "language" in an evolving population of neural networks. *Connection Science*, 10(2):83–97.

Cangelosi, A. and Parisi, D. (2001). How nouns and verbs differentially affect the behavior of artificial organisms. In Moore, J. D. and Stenning, K., editors, *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*, pages 170–175. Lawrence Erlbaum Associates, London.

Carpenter, M., Akhtar, N., and Tomasello, M. (1998a). 14- through 18-month-old infants differentially imitate intentional and accidental actions. *Infant Behavior and Development*, 21:315–330.

Carpenter, M., Nagell, K., and Tomasello, M. (1998b). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, 63.

Carruthers, P. (1998). Thinking in language?: Evolution and a modularist possibility. In Carruthers, P. and Boucher, J., editors, *Language and Thought*, pages 94–119. Cambridge University Press.

Carruthers, P. (2008). Language in cognition. In Margolis, E., Samuels, R., and Stich, S., editors, *The Oxford Handbook of Philosophy of Cognitive Science*. Oxford University Press.

Carruthers, P. and Boucher, J. (1998). Introduction: Opening up options. In Carruthers, P. and Boucher, J., editors, *Language and Thought: Interdisciplinary Themes*, pages 1–18. Cambridge University Press.

Chalmers, D. J., French, R. M., and Hofstadter, D. R. (1992). High-level perception, representation, and analogy: a critique of artificial intelligence methodology. *J. Exp. Theor. Artif. Intell.*, 4(3):185–211.

Chapman, D. (1985). Planning for conjunctive goals. *Artificial Intelligence*, 32(3):333–377.

Chapman, D. (1991). *Vision, Instruction, and Action*. MIT Press, Cambridge, MA.

Cheng, K. (1986). A purely geometric module in the rat's spatial representation. *Cognition*, 23:149–178.

Chomsky, N. (1988). *Language and problems of knowledge : the Managua lectures*. MIT Press, Cambridge, MA, USA.

Christiansen, M. H. and Kirby, S. (2003). Language evolution: consensus and controversies. *Trends in Cognitive Sciences*, 7(7):300–307.

Clancey, W. J. (1997). *Situated cognition : on human knowledge and computer representations*. Cambridge University Press, Cambridge, U.K. ; New York, NY, USA.

Clark, A. (1989). *Microcognition: philosophy, cognitive science, and parallel distributed processing*. MIT Press, Cambridge, MA, USA.

Clark, A. (1997). *Being There: Putting Brain, Body and World Together Again*. MIT Press, Cambridge, MA.

Clark, A. (1998). Magic words: How language augments human computation. In Carruthers, P. and Boucher, J., editors, *Language and Thought: Interdisciplinary Themes*, pages 168–183. Cambridge University Press.

Clark, A. (1999). An embodied cognitive science? *Trends in cognitive science*, 3(9):345–351.

Clark, A. (2003). Artificial intelligence and the many faces of reason. In *The Blackwell Guide to Philosophy of Mind*, pages 309–321. Blackwell.

Clark, A. (2005). Beyond the flesh: Some lessons from a mole cricket. *Artificial Life*, 11(1):233–244.

Clark, A. (2006). Language, embodiment and the cognitive niche. *Trends in Cognitive Sciences*, 10(8).

Cohen-Cole, J. (2005). The reflexivity of cognitive science: The scientist as model of human nature. *History of the human sciences*, 18(4):107–139.

Costall, A. (1991). 'graceful degradation': Cognitivism and the metaphors of the computer. In Still, A. and Costall, A., editors, *Against Cognitivism: Alternative Foundations for Cognitive Psychology*, pages 151–169. Harvester Wheatsheaf, New York, NY.

Davis, M. (2000). *The Universal Computer: The Road from Leibniz to Turing*. W. W. Norton & Co., Inc., New York, NY.

de Boer, B. (2000). Emergence of vowel systems through self-organisation. *AI Communications*, 13(1):27–39.

Dean, J. (1998). Animats and what they can tell us. *Trends in cognitive science*, 2(2):60–67.

Dennett, D. (1978). *Brainstorms: Philosophical Essays on Mind and Psychology*. Bradford Books, Montgomery, VT.

Di Paolo, E. A. (1998). An investigation into the evolution of communication. *Adaptive Behavior*, 6(2):285–324.

Dreyfus, H. (1979). *What Computers Can't Do*. The MIT Press, Cambridge, MA.

Dreyfus, H. L. (2007). Why heideggerian ai failed and how fixing it would require making it more heideggerian. *Artificial Intelligence*, 171(18):1137–1160.

Dreyfus, H. L. and Dreyfus, S. (1988). Making a mind versus modeling the brain: Artificial intelligence back at a branchpoint. *Daedalus*, 117(1):15–44.

Dummett, M. (1994). *Origins of Analytical Philosophy*. Harvard University Press, Cambridge, MA.

Durkheim, E. (1938). *The rules of sociological method*. The Free Press, New York, NY.

Fitch, W. T. (2000). The evolution of speech: a comparative review. *Trends in cognitive sciences*, 4(7):258–267.

Floreano, D. and Mondada, F. (1996). Evolution of Plastic Neurocontrollers for Situated Agents. In Maes, P., Mataric, M., Meyer, J.-A., Pollack, J., and Wilson, S., editors, *4th International Conference on Simulation of Adaptive Behavior (SAB'1996)*, Cambridge, MA. MIT Press.

Fodor, J. (1981). *Representations*. MIT Press, Cambridge, MA. cited in Costall (1991).

Fodor, J. (1994). Fodor's guide to mental representation: The intelligent auntie's vade-mecum. In Stich, S. S. and Warfield, T. A., editors, *Mental Representation: A Reader*, pages 9–33. Blackwell Pulishers, Cambridge, Massachusetts.

Fodor, J. (2000). *The mind doesn't work that way: the scope and limits of computational psychology*. The MIT Press, Cambridge, MA.

Fodor, J. A. (1980). Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioural and Brain Sciences*, 3:63–109.

Fodor, J. A. and Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28:3–71.

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science*, 29:737–767.

Gallagher, S. (2001). The practice of mind: Theory, simulation, or primary interaction. *Journal of Consciousness Studies*, 8:83–108.

Gallagher, S. (2008). Philosophical antecedents to situated cognition. In Aydede, M. and Robbins, P., editors, *Cambridge Handbook of Situated Cognition*, pages 35–53. Cambridge University Press.

Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. MIT Press, Cambridge, MA.

Gardner, H. (1985). *The Mind's New Science: A History of the Cognitive Revolution*. Basic Books, New York, NY.

Garner, W. R. (1974). *The Processing of Information and Structure*. Lawrence Erlbaum, Potomac, MD.

Gasser, M. (1993). The structure-grounding problem. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, pages 149–152, NJ: Erlbaum.

Gergely, G., Bekkering, H., and Kiraly, I. (2002). Rational imitation in preverbal infants. *Nature*, 415(755).

Gibbs, R. W. (2006). *Embodiment and Cognitive Science*. Cambridge University Press.

Goldfarb, W. D. (1983). I want you to bring me a slab: Remarks on the opening sections of the *Philosophical Investigations*. *Synthese*, 56:265–282.

Goldstone, R. L. (1994). The role of similarity in categorization: providing a framework. *Cognition*, 52:125–157.

Goldstone, R. L. (1999). Similarity. In Wilson, R. A. and Keil, F. C., editors, *MIT Encyclopedia of the Cognitive Sciences*, pages 763–765. MIT Press, Cambridge, MA.

Goldstone, R. L. and Kersten, A. (2003). Concepts and categorization. In Healy, A. F. and Proctor, R. W., editors, *Comprehensive Handbook pf Psychology, Vol. 4: Experimental Psychology*, pages 599–621. Wiley, Hoboken, NJ.

Goodman, N. (1972). Seven strictures on similarity. In Goodman, N., editor, *Problems and Projects*, pages 437–447. Bobbs-Merrill, New York.

Greenwood, J. D. (1999). Understanding the cognitive revolution in psychology. *Journal of the History of the Behavioral Sciences*, 35(1):1–22.

Grim, P., Denis, P. S., and Kokalis, T. (2002). Learning to communicate: The emergence of signaling in spatialized arrays of neural nets. *Adaptive Behavior*, 10(1):45–70.

Grim, P., Kokalis, T., Alai-Tafti, A., Kilb, N., and Denis, P. S. (2004). Making meaning happen. *Journal of Experimental & Theoretical Artificial Intelligence*, 16(4):209–243.

Grim, P., Kokalis, T., Tafti, A., and Kilb, N. (2001). Evolution of communication with a spatialized genetic algorithm. *Evolution of Communication*, 3:105–134.

Guignon, C. B. (1983). *Heidegger and the Problem of Knowledge*. Hackett Publishing Company, Indianapolis, Indiana.

Hahn, U. and Chater, N. (1998). Similarity and rules: distinct? exhaustive? empirically distinguishable? *Cognition*, 65:197–230.

Hammer, B. and Hitzler, P., editors (2007). *Perspectives of Neural-Symbolic Integration*. Springer Verlag.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, 42:335–346.

Harvey, I., Di Paolo, E., Wood, R., and Quinn, M. (2005). Evolutionary robotics: A new scientific tool for studying cognition. *Artificial Life*, 11(1-2):79–98.

Harvey, I., Husbands, P., Cliff, D., Thompson, A., and Jakobi, N. (1997). Evolutionary robotics: the sussex approach. *Robotics and Autonomous Systems*, 20:205–224.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Verlag.

Haugeland, J. (1978). The nature and plausibility of cognitivism. *Behavioral and Brain Sciences*, 1:215–260.

Haugeland, J., editor (1981a). *Mind Design: Philosophy, Psychology, Artificial Intelligence*. MIT Press, Cambridge, MA.

Haugeland, J. (1981b). Semantic engines: An introduction to mind design. In Haugeland (1981a), pages 1–34.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. The MIT Press, Cambridge, MA.

Haugeland, J. (1990). The intentionality all-stars. *Philosophical Perspectives*, 4:383–427. great stuff. cognitivists: neo-cartesians. norms and language.

Hauser, M. D. (1996). *The evolution of communication*. MIT Press, Cambridge, MA.

Hauser, M. D., Chomsky, N., and Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science*, 298(5598):1569–1579.

Hayes, P. J. (1979). The naive physics manifesto. In Michie, D., editor, *Expert Systems in the Micro-Electronic Age*, pages 242–270. Edinburgh University Press.

Hayes, P. J., Ford, K. M., and Agnew, N. M. (1996). Goldilocks and the frame problem. In Ford, K. M. and Pylyshyn, Z. W., editors, *The robot's dilemma revisited: the frame problem in artificial intelligence*, pages 135 – 137. Ablex Publishing Corp., Norwood, NJ, USA.

Haykin, S. (1999). *Neural Networks, A Comprehensive Foundation*. Prentice Hall, Upper Saddle River, New Jersey.

Heidegger, M. (2001). *Sein und Zeit*. Max Niemeyer Verlag, Thübingen.

Hermer, L. and Spelke, E. (1996). Modularity and development: the case of spatial reorientation. *Cognition*, 61:195–232.

Hermer-Vazquez, L., Spelke, E., and Katsnelson, A. (1999). Sources of flexibity in human cognition: Dual-task studies of space and language. *Cognitive Psychology*, 39:3–36.

Hinton, G. E., McClelland, J. L., and Rumelhart, D. E. (1986). Distributed representations. In Rumelhart, D. E. and McClelland, J. L., editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, pages 77–109. MIT Press.

Hofstadter, D. (1982). Waking up from the boolean dream, or, subcognition as computation. *Scientic American*. Reprinted in Metamagical Themas: Questing for the Essence of Mind and Pattern (1985).

Holland, J. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor.

Hurford, J. (1989). Biological evolution of the saussurean sign as a component of the language acquisition device. *Lingua*, 77(2):187–222.

Husserl, E. (1982). *Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie*. Felix Meiner Verlag, Hamburg.

Hutchins, E. (1996). *Cognition in the Wild*. The MIT Press, Cambridge, MA.

Hutchins, E. and Hazlehurst, B. (1995). How to invent a lexicon: the development of shared symbols in interaction. In Gilbert, G. N. and Conte, R., editors, *Artificial Societies: The computer simulation of social life*, pages 157–189. UCL Press, London.

Izquierdo, E., Harvey, I., and Beer, R. D. (2008). Associative learning on a continuum in evolved dynamical neural networks. *Adaptive Behavior*, 16(6):361–384.

Jackendoff, R. (1999). Possible stages in the evolution of the language capacity. *Trends in Cognitive Sciences*, 3(7):272–279.

Jäkel, F., Schölkopf, B., and Wichmann, F. A. (2007). A tutorial on kernel methods for categorization. *Journal of Mathematical Psychology*, 51(6):343–358.

Jäkel, F., Schölkopf, B., and Wichmann, F. A. (2008). Generalization and similarity in exemplar models of categorization: Insights from machine learning. *Psychonomic Bulletin and Review*, 15(2):256–271.

John, R. S., Türkmen, U., and Zugic, R. (2006). Comparative cognitive robotics: Autonomous robots as empirical models of animal learning. In Opwis, K. and Penner, I.-K., editors, *Proceedings of Cognitio 2006: Beyond the brain: embodied, situated & distributed cognition*.

Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax. In Briscoe, T., editor, *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge University Press.

Kirby, S. and Hurford, J. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In Cangelosi, A. and Parisi, D., editors, *Simulating the Evolution of Language*, pages 121–148. Springer Verlag, London.

Kosslyn, S. M., Pinker, S., Smith, G. E., and Shwartz, S. P. (1979). On the demystification of mental imagery. *Behavioral and Brain Sciences*, 2:535–548.

Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological Review*, 99:22–44.

Kuhn, T. S. (1996). *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, IL. 3rd edition.

Kwisthout, J., Vogt, P., Haselager, P., and Dikstra, T. (208). Joint attention and language evolution. Technical Report 2007-039, Department of Information and Computing Sciences, Utrecht University.

Lachman, R., Lachman, J. L., and Butterfield, E. C. (1979). *Cognitive Psychology and Information Processing: An Introduction*. Lawrence Erlbaum, Hillsdale, NJ.

Lakoff, G. (1990). *Women, Fire, and Dangerous Things*. University Of Chicago Press, Chicago, USA.

Lave, J., editor (1988). *Cognition in Practice: Mind, mathematics, and culture in everyday life*. Cambridge University Press, Cambridge , UK.

Lenat, D. B. and Feigenbaum, E. A. (1991). On the thresholds of knowledge. *Artificial Intelligence*, 47:185–250.

MacLennan, B. and Burghardt, G. M. (1993). Synthetic ethology and the evolution of cooperative communication. *Adaptive Behavior*, 2(2):161–187.

Maes, P., editor (1990). *Designing Autonomous Agents: Theory and Practice from Biology to Engineering and Back*. The MIT Press, Cambridge, MA.

Maes, P. (1993). Behavior-based artificial intelligence. In *From Animals to Animats 2: Proceedings of the Second International Conference on Simulation of Adaptive Behavior*, pages 2–10, Cambridge, MA. MIT Press/Bradford Books.

Margolis, E. and Laurence, S. (1999a). Concepts and cognitive science. In Margolis and Laurence (1999b), pages 3–81.

Margolis, E. and Laurence, S., editors (1999b). *Concepts: core readings*. MIT Press, Cambridge, MA.

Markman, A. B. and Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, 25:431–467.

Marocco, D., Cangelosi, A., and Nolfi, S. (2002). The role of social and cognitive abilities in the emergence of communication: Experiments in evolutionary robotics. In *EPSRC/BBSRC International Workshop Biologically-Inspired Robotics Bristol*, pages 174–181.

Maturana, H. and Varela, F. (1980). *Autopoiesis and cognition: The realization of the living*. D. Reidel, Boston.

McCarthy, J. and Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence*, 4:463–502.

McDermott, D. (1981). Artificial intelligence meets natural stupidity. In Haugeland (1981a), pages 143–160.

McDermott, D. (1987a). We've been framed: Or, why ai is innocent of the frame problem. In Pylyshyn (1987), pages 3–81.

McDermott, D. V. (1987b). A critique of pure reason. *Computational Intelligence*, 3:151–237.

McGinn, M. (1997). *Wittgenstein and the Philosophical Investigations*. Routledge.

McReynolds, P. (1980). The clock metaphor in the history of psychology. In Nickles, T., editor, *Scientific Discovery: Case Studies*, pages 97–112. Reidel, Dordrecht.

Medin, D. L., Goldstone, R. L., and Gentner, D. (1993). Respects for similarity. *Psychological Review*, 100(2):254–278.

Medin, D. L. and Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85(3):207–238.

Meltzoff, A. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, 31:838–850.

Minsky, M. L. (1961). Steps toward artificial intelligence. *Proceedings of the Institute of Radio Engineers*, 49:8–30.

Monahan, J. S. and Lockhead, G. R. (1977). Identification of integral stimuli. *Journal of Experimental Psychology: General*, 106(1):94–110.

Moody, J. and Darken, C. J. (1989). Fast learning in networks of locally tuned processing units. *Neural Computation*, 1:281–294.

Neisser, U. (1967). *Cognitive Psychology*. Appleton-Century- Crofts., New York, NY.

Newell, A. (1982). The knowledge level. *Artificial Intelligence*, 18:87–127.

Newell, A. and Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 19(3):113–126. Reprinted in Boden (1990).

Nosofsky, R. M. (1985). Luce's choice model and thurstone's categorical judgment model compared: Kornbrot's data revisited. *Perception & Psychophysics*, 37(1):89–91.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115(1):39–57.

Nosofsky, R. M. (1990). Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology*, 34:393–418.

Nosofsky, R. M., Kruschke, J. K., and McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(2):211–233.

Osbeck, L. M., Malone, K. R., and Nersessian, N. J. (2007). Dissenters in the sanctuary: Evolving frameworks in 'mainstream' cognitive science. *Theory & Psychology*, 17(2):197–230.

Oudeyer, P.-Y. (2005). The self-organization of speech sounds. *Journal of Theoretical Biology*, 233(3):435–449.

Parisi, D. (1997). An artificial life approach to language. *Brain and Language*, 59(1):121–146.

Parisi, D., Cecconi, F., and Cerini, A. (1995). Kin-directed altruism and attachment behaviour in an evolving population of neural networks. In Gilbert, N. and Conte, R., editors, *Artificial societies: The computer simulation of social life*, pages 238–251. UCL Press, London.

Parsons, L. M. (1987). Imagined spatial transformations of one's hands and feet. *Cognitive Psychology*, 19:178–241.

Pfeifer, R. and Scheier, C. (2001). *Understanding Intelligence*. MIT Press, Cambridge, MA, second edition.

Phattanasri, P., Chiel, H. J., and Beer, R. D. (2007). The dynamics of associative learning in evolved model circuits. *Adaptive Behavior*, 15(4):377–396.

Piattelli-Palmarini, M. (1989). Evolution, selection and cognition: From "learning" to parameter setting in biology and in the study of language. *Cognition*, 31(1):1–44.

Pinker, S. (1984). *Language learnability and language development*. Harvard University Press, Cambrdige, MA.

Pinker, S. (1994). *The language instinct: How the mind creates language*. William Morrow, New York.

Pinker, S. and Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences*, 13(4):707–784.

Pinker, S. and Jackendoff, R. (2005). The faculty of language: What's special about it? *Cognition*, 95(2):201–236.

Port, R. F. and van Gelder, T. (1995). *Mind as Motion: Explorations in the Dynamics of Cognition*. The MIT Press.

Pothos, E. M. (2005). The rules versus similarity distinction. *Behavioral and Brain Sciences*, 28:1–49.

Premack, D. (1986). *Gavagai! Or the future history of the animal language controversy*. Bradford, MIT Press, Cambridge, MA.

Preston, B. (1993). Heidegger and artificial intelligence. *Philosophy and Phenomenological Research*, 53(1):43–69.

Pylyshyn, Z. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin*, 80:1–24.

Pylyshyn, Z. W., editor (1987). *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*. Ablex Publishing, Norwood, NJ.

Quinn, M. (2001). Evolving communication without dedicated communication channels. In Kelemen, J. and Sosik, P., editors, *Proceedinga of the 6th European Conference on Artificial Life, ECAL 2001*, pages 357–366, Berlin. Springer Verlag.

Rips, L. J. (1989). Similarity, typicality, and categorization. In Vosniadou, S. and Ortony, A., editors, *Similarity and analogical reasoning*, pages 21–59. Cambridge University Press, New York, NY, USA.

Rosch, E. and Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7:573–605.

Ross, H. S. and Lollis, S. P. (1987). Communication within infant social games. *Developmental Psychology*, 23:241–248.

Schmitz, J., Dean, J., Kindermann, T., Schumm, M., and Cruse, H. (2001). A biologically inspired controller for hexapod walking: Simple solutions by exploiting physical properties. *Biological Bulletin*, 200:195–200.

Schölkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA.

Schön, D. (1978). Generative metaphor: A perspective on problem setting in social policy. In Ortony, A., editor, *Metaphor and Thought*, pages 254–283. Cambridge University Press, Cambridge.

Shanahan, M. (2008). The frame problem. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy (Fall 2008 Edition)*.

Shepard, R. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, 22(4):325–345.

Shepard, R. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, 1:54–87.

Shepard, R. and Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, 171:701–703.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237:1317–1323.

Shusterman, A. and Spelke, E. (2005). Language and the development of spatial reasoning. In P. Carruthers, S. L. and Stich, S., editors, *The Innate Mind : Structure and Contents*. Oxford University Press.

Simon, H. A. (1969). *The Sciences of the Artificial*. MIT Press, Cambridge, Massachusetts, first edition.

Smith, B. C. (1991). The owl and the electric encyclopedia. *Artificial Intelligence*, 47:251–288.

Smith, B. C. (1996). *On the Origin of Objects*. MIT Press, Cambridge, MA.

Smith, E. E. and Medin, D. L. (1981). *Categories and concepts*. Harvard University Press, Cambridge, MA.

Smith, E. E. and Sloman, S. A. (1994). Similarity- versus rule-based categorization. *Memory & Cognition*, 22(4):377–386.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11(1):1–74.

Steels, L. (1994). The artificial life roots of artificial intelligence. *Artificial Life Journal*, 1(1).

Steels, L. (1996a). Emergent adaptive lexicons. Cambridge, MA. MIT Press/Bradford Books.

Steels, L. (1996b). The origins of intelligence. In *Proceedings of the Carlo Erba Foundation Meeting on Artificial Life*, Milano. Foundazione Carlo Erba.

Steels, L. (1996c). Perceptually grounded meaning creation. In Tokoro, M., editor, *Proceedings of the International Conference on Multi-Agent Systems*, Cambridge, MA. MIT Press.

Steels, L. (1997a). Constructing and sharing perceptual distinctions. In van Someren, M. and Widmer, G., editors, *Proceedings of the European Conference on Machine Learning (ECMLA 97)*, pages 4–13, Berlin. Springer.

Steels, L. (1997b). The synthetic modeling of language origins. *Evolution of Communication*, 1(1):1–34.

Steels, L. (1998). Synthesizing the origins of language and meaning using coevolution, self-organisation and level formation. In Hurford, J., Knight, C., and Studdert-Kennedy, M., editors, *Approaches to the Evolution of Language: Social and Cognitive Bases*, pages 384–404. Edinburgh University Press.

Steels, L. (1999). The spontaneous self-organization of an adaptive language. In Furukawa, K., Michie, D., and Muggleton, S., editors, *Machine Intelligence 15*, pages 205–224, St. Catherine's College, Oxford. Oxford University Press. Machine Intelligence Workshop: July 1995.

Steels, L. (2000). Language as a complex adaptive system. In Schoenauer, M., editor, *Proceedings of PPSN VI*, Lecture Notes in Computer Science, Berlin, Germany. Springer Verlag.

Steels, L. (2002). Simulating the evolution of a grammar for case. In *EvoLang 200: Fourth International Conference on the Evolution of Language*.

Steels, L. (2003). The evolution of communication systems by adaptive agents. In Alonso, E., Kudenko, D., and Kazakov, D., editors, *Adaptive Agents and Multi-Agent Systems: Adaptation and Multi-Agent Learning. LNAI 2636*, pages 125–140. Springer Verlag, Berlin.

Steels, L. (2004). Constructivist development of grounded construction grammars. In *Proceedings Annual Meeting Association for Computational Linguistics Conference*, Barcelona.

Steels, L. (2006). How to do experiments in artificial language evolution and why. In *Proceedings of the 6th International Conference on the Evolution of Language*, pages 323–332.

Steels, L. (2008). The symbol grounding problem has been solved. so what's next? In de Vega, M., editor, *Symbols and Embodiment: Debates on Meaning and Cognition*, chapter 12. Oxford University Press, Oxford.

Steels, L. and Baillie, J.-C. (2003). Shared grounding of event descriptions by autonomous robots. *Robotics and Autonomous Systems*, 43(2-3):163–173.

Steels, L. and Kaplan, F. (1998). Stochasticity as a source of innovation in language games. In Adami, C., Belew, R., Kitano, H., and Taylor, C., editors, *Artificial Life VI*, Los Angeles. MIT Press.

Steels, L. and Kaplan, F. (1999a). Collective learning and semiotic dynamics. In Floreano, D., Nicoud, J.-D., and Mondada, F., editors, *ECAL99*, pages 679–688. Springer-Verlag.

Steels, L. and Kaplan, F. (1999b). Situated grounded word semantics. In Dean, T., editor, *IJCAI 99*, pages 862–867. Morgan Kaufmann Publishers.

Steels, L. and Kaplan, F. (2001). AIBO's first words : The social learning of language and meaning. *Evolution of Communication*, 4(1).

Steels, L., Kaplan, F., McIntyre, A., and Looveren, J. V. (2002). Crucial factors in the origins of word-meaning. In Wray, A., editor, *The Transition to Language*, pages 252–271. Oxford University Press, Oxford, UK.

Steels, L. and Vogt, P. (1997). Grounding adaptive language games in robotic agents. In Husbands, C. and Harvey, C., editors, *Proceedings of the Fourth European Conference on Artificial Life (ECAL' 97)*, London. MIT Press.

Suchman, L. A. (1987). *Plans and Situated Actions: the problem of human machine communication.* Cambridge University Press, Cambridge, MA.

Sutton, R. S. (1991). Integrated modeling and control based on reinforcement learning and dynamic programming. In Lippmann, R., Moody, J., and Touretzky, D. S., editors, *Advances in Neural Information Processing Systems 3*, pages 471–478. Morgan-Kauffman.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction.* MIT Press.

Symes, E., Ellis, R., and Tucker, M. (2007). Visual object affordances: Object orientation. *Acta Psychologica*, 124:238–255.

Taylor, C. (1986). Overcoming epistemology. In Baynes, K., Bohman, J., and McCarthy, T., editors, *After Philosophy: End or Transformation?*, pages 464–485. The MIT Press, Cambridge, MA.

Thelen, E., Schoner, G., Scheier, C., and Smith, L. B. (2001). The dynamics of embodiment: A field theory of infant perseverative reaching. *Behavioral and Brain Sciences*, 24:1–86.

Thompson, A. (1998). *Hardware Evolution: Automatic design of electronic circuits in reconfigurable hardware by artificial evolution*. Springer Verlag.

Thompson, E. and Zahavi, D. (2007). Philosophical issues: Phenomenology. In Zelazo, P. D., Moscovitch, M., and Thompson, E., editors, *The Cambridge Handbook of Consciousness*, pages 67–87. Cambridge University Press, Cambridge.

Thompson, R. K. R., Oden, D. L., and Boysen, S. T. (1997). Language-naive chimpanzees (pan troglodytes) judge relations between relations in a conceptual matching-to-sample task. *Journal of Experimental Psychology: Animal Behavior Processes*, 23(1):31–43.

Thrun, S. (1992). The role of exploration in learning control. In White, D. and Sofge, D., editors, *Handbook for Intelligent Control: Neural, Fuzzy and Adaptive Approaches*. Van Nostrand Reinhold, Florence, Kentucky.

Tomasello, M. (1995). Language is not an instinct. *Cognitive Development*, 10:131–156.

Tomasello, M. (1998). *The new psychology of language: Cognitive and functional approaches*. Erlbaum, Mahwah, NJ.

Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Harvard University Press.

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition*, 74:209–253.

Tomasello, M. (2003). The key is social cognition. In Gentner, D. and Goldin-Meadow, S., editors, *Language in mind: Advances in the study of language and thought*, pages 47–57. MIT Press, Cambridge, MA.

Tomasello, M., Akhtar, N., Dodson, K., and Rekau, L. (1997). Differential productivity in young children's use of nouns and verbs. *Journal of Child Language*, 24:373–387.

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28:675–735.

Tucker, M. and Ellis, R. (1998). On the relations between seen objects and components of potential actions. *Journal of Experimental Psychology: Human Perception and Performance*, 24:830–846.

Tugendhat, E. (1970). *Der Wahrheitsbegriff bei Husserl und Heidegger*. Walter de Gruyter & Co., Berlin.

Türkmen, U. (2007). Situated representations and the modelling of the evolution of language. In Frings, C., Mecklinger, A., Opitz, B., Pospeschill, M., Wentura, D., and Zimmer, H. D., editors, *Kognitionsforschung 2007: Beiträge zur 8. Jahrestagung der Gesellschaft für Kognitionswissenschaft*, Aachen. Shaker Verlag.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84:327–352.

Türkmen, U. and Zugic, R. (2008). Modelling the social coordiantion of behavior with public symbols. In Bullock, S., Noble, J., Watson, R., and Bedau, M. A., editors, *Artificial Life XI: Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems*, page 812. MIT Press, Cambridge, MA.

Vera, A. H., Herbert, and Simon, A. (1993). Situated action: a symbolic interpretation. *Cognitive Science*, 17:7–48.

Vogt, P. (2005a). The emergence of compositional structures in perceptually grounded language games. *Artificial Intelligence*, 167(1-2):206–242.

Vogt, P. (2005b). On the acquisition and evolution of compositional languages: Sparse input and the productive creativity of children. *Adaptive Behavior*, 13(4):325–346.

von Foerster, H. (1980). Thoughts and notes on cognition. In Gavin, P., editor, *Cognition: A Multiple View*, pages 25–48. Spartan Books, New York, NY.

Wagner, K., Reggia, J. A., Uriagereka, J., and Wilkinson, G. S. (2003). Progress in the simulation of emergent communication and language. *Adaptive Behavior*, 11(1):37–69.

Watkins, C. (1989). *Situated representation: Solving the handcoding problem with emergent Structured representation*. PhD thesis, Binghamton University.

Webb, B. (1994). Robotic experiments in cricket phonotaxis. In Cliff, D., Husbands, P., Meyer;, J.-A., and Wilson, S. W., editors, *From Animals to Animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behavior*, pages 45–54, Cambridge, MA. MIT Press.

Werner, C. W. and Rehkämper, G. (1999). Discrimination of multidimensional geometrical figures by chickens: categorization and pattern-learning. *Animal Cognition*, 2:27–40.

Werner, C. W. and Rehkämper, G. (2001). Categorization of multidimensional geometrical figures by chickens (Gallus gallus f. domestica): fit of basic assumptions from exemplar, feature and prototype theory. *Animal Cognition*, 4:37–48.

Werner, G. and Dyer, M. (1992). Evolution of communication in artificial organisms. In Langton, C., Taylor, C., Farmer, D., and Rasmussen, S., editors, *Artificial Life II*, pages 659–687, Redwood City, CA. Addison-Wesley Pub.

Wheeler, M. (1995). Escaping from the cartesian mind-set: Heidegger and artificial life. In *Proceedings of the Third European Conference on Advances in Artificial Life*, pages 65–76, London, UK. Springer Verlag.

Williams, P., Beer, R., and Gasser, M. (2008). Evolving referential communication in embodied dynamical agents. In Bullock, S., Noble, J., Watson, R., and Bedau, M. A., editors, *Artificial Life XI: Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems*, pages 702–709. MIT Press, Cambridge, MA.

Wilson, S. W. (1991). The animat path to ai. In Meyer, J.-A. and Wilson, S. W., editors, *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior*, pages 15–21. MIT Press, Cambridge, MA.

Wittgenstein, L. (1984). *Tractatus Logico-Philosophicus*. Suhrkamp Verlag, Frankfurt am Main.

Wittgenstein, L. (1991). *Das Blaue Buch. Eine Philosophische Betrachtung. (Das Braune Buch). Werkausgabe Band 5*. Suhrkamp Verlag, Frankfurt am Main.

Wittgenstein, L. (2001). *Philosophische Untersuchungen*. Suhrkamp Verlag, Frankfurt am Main.

Zwaan, R. and Madden, C. J. (2005). Embodied sentence comprehension. In Pecher, D. and Zwaan, R., editors, *Grounding cognition: The role of perception and action in memory, language, and thinking*, pages 224–245. Cambridge University Press, Cambridge, UK.

Zwaan, R. A., Stanfield, R. A., and Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. *Psychological Science*, 13:168–171.